

Domenico Iorio-Fili



GATTO

versione 3.3

Manuale d'uso

versione GATTO: 3.3

ultima modifica

28 maggio 2012

INDICE

INDICE	3
INTRODUZIONE	6
CAP. 1 - PANORAMICA SU GATTO	6
CAP. 2 - INTRODUZIONE ALL'USO DI GATTO	14
<i>Lezione 1. Esempio di uso di GATTO: lettura di statistiche e impostazioni del corpus Demo; ricerca delle forme uomo e uomini</i>	21
PARTE I - PREPARAZIONE DEI DATI	24
CAP. 3 - PREPARAZIONE DI UN FILE TESTO	24
<i>Lezione 2. Caratteri validi nei file testo all'interno delle parole. Forme polirematiche. Ordinamento alfabetico delle forme. Digitazione delle lettere accentate. Segni di interpunzione</i>	26
<i>Lezione 3. Come preparare un file testo usando i soli codici base: testo, riferimenti organici, numeri di pagina, di volume e di colonna, numeri di riga</i>	30
<i>Lezione 4. Come preparare un file testo: parti in versi e parti in prosa, numeri di verso, conteggio dei versi per distici, terzine, ecc..., codifica della rima al mezzo, riferimenti organici a più livelli</i>	36
<i>Lezione 5. Come preparare un file testo: campi formula, scioglimenti, espunzioni, corsivi, citazioni, interpunzioni, parlante. Lemmi e iperlemmi. Inizio del file testo</i>	41
CAP. 4 - USO DI NOTE E DI EDIZIONI DIVERSE O TRADUZIONI	45
<i>Lezione 6. Come utilizzare delle note in un file testo.</i>	45
<i>Lezione 7. Come preparare il testo associato contenente le note</i>	48
<i>Lezione 8. Come preparare un file testo combinandolo con una diversa edizione o una traduzione in altra lingua</i>	49
<i>Lezione 9. Come gestire la presenza contemporanea di note e di una diversa edizione o traduzione in altra lingua</i>	52
CAP. 5 - RACCOLTA DELLE INFORMAZIONI BIBLIOGRAFICHE	54
<i>Lezione 10. Datazione dei testi. Concetto di data codificata, passaggio dalla data descrittiva alla data codificata, costruzione della data codificata</i>	56
CAP. 6 - DEFINIZIONE DEI PARAMETRI DA ASSEGNARE AL CORPUS	60
PARTE II - COSTRUZIONE DI UN CORPUS E INSERIMENTO DI TESTI	63
CAP. 7 - COSTRUZIONE DI UN CORPUS E DEL RELATIVO ARCHIVIO BIBLIOGRAFICO	63
<i>Lezione 11. Come costruire un corpus completo di archivio bibliografico</i>	64
<i>Lezione 12. Come inserire e correggere i dati in un archivio bibliografico</i>	65
CAP. 8 - INSERIMENTO DI UN TESTO IN UN CORPUS	67
<i>Lezione 13. Come inserire un testo in un corpus.</i>	71
<i>Lezione 14. Come inserire nel corpus i testi associati, contenenti note, diverse edizioni o traduzioni.</i>	72
PARTE III - LEMMATIZZAZIONE E INSERIMENTO DI IPERLEMMI	74
CAP. 9 - CONCETTO E MECCANISMI DI LEMMATIZZAZIONE. FUSIONE DI PIÙ LEMMATIZZAZIONI	74
CAP. 10 - LEMMATIZZAZIONE SUL CORPUS	79
<i>Lezione 15. Come effettuare la lemmatizzazione</i>	81
<i>Lezione 16. Uso della finestra a contesto singolo</i>	87
<i>Lezione 17. Selezione del tipo di occorrenze e comandi di spostamento tra contesti</i>	88
<i>Lezione 18. Funzioni accessorie, ulteriori informazioni sulle ricerche di partenza e rassegna dei comandi per visualizzare informazioni</i>	88
CAP. 11 - LEMMATIZZAZIONE PER SINGOLO TESTO	91
<i>Lezione 19. Come lemmatizzare un testo</i>	93
<i>Lezione 20. Come lemmatizzare con la finestra a contesto singolo o in modalità sequenziale</i>	96
<i>Lezione 21. Opzioni di lemmatizzazione (tipo di forme, tipo di occorrenze) e comandi di spostamento tra occorrenze della stessa forma e tra forme diverse</i>	98
<i>Lezione 22. Funzioni accessorie e rassegna dei comandi per visualizzare informazioni</i>	100
CAP. 12- INSERIMENTO DI IPERLEMMI	102
<i>Lezione 23. Come inserire iperlemmi e collegarli ai lemmi.</i>	104
<i>Lezione 24. Come associare direttamente gli iperlemmi alle singole occorrenze.</i>	108

CAP. 13- LEMMI E COPPIE FORMA-LEMMA MUTI. ESPORTAZIONE E IMPORTAZIONE DI LEMMARIO E IPERLEMMARIO	109
<i>Lezione 25. Come trasferire il lemmario (ed eventualmente l'iperlemmario) di un corpus in un altro corpus</i>	110
<hr/>	
PARTE IV - ESTRAZIONE DI INFORMAZIONI LESSICOGRAFICHE	112
CAP. 14 - ESTRAZIONE DI INFORMAZIONI DA UN CORPUS GATTO	112
CAP. 15 - RICERCHE PER FORME	115
<i>Lezione 26. Come effettuare ricerche per forme</i>	118
<i>Lezione 27. Altre opzioni disponibili. Rimario. Ordinamenti dei risultati</i>	123
<i>Lezione 28. Uso delle varianti. Salvataggio e recupero dell'accumulatore</i>	125
<i>Lezione 29. Visualizzazione dei contesti ottenuti</i>	127
<i>Lezione 30. Raffinamento</i>	137
<i>Lezione 31. Visualizzazione dei contesti a partire dalla lista dei testi che li contengono</i>	139
<i>Lezione 32. Come esportare i contesti trovati</i>	140
<i>Lezione 33. Ricerca delle forme incluse in un brano di testo. Ricerca delle forme mai lemmatizzate</i>	143
<i>Lezione 34. Come visualizzare ed esportare le note, una diversa edizione o una traduzione</i>	145
CAP. 16 - RICERCHE BASATE SUI LEMMI	147
<i>Lezione 35. Come effettuare ricerche per lemmi</i>	148
<i>Lezione 36. Come effettuare le altre ricerche basate sui lemmi: per categorie grammaticali, per disambiguatori, concordanze lemmatizzate. Come effettuare ricerche combinate.</i>	151
CAP. 17 - GENERAZIONE DI INDEX LOCORUM	152
<i>Lezione 37. Come generare Index Locorum per forme e lemmatizzati</i>	153
CAP. 18 - RICERCHE PER IPERLEMMI	154
<i>Lezione 38. Come effettuare ricerche per iperlemmi</i>	156
CAP. 19 - RICERCHE DI COCCORRENZE	157
<i>Lezione 39. Come impostare ed effettuare ricerche di cooccorrenze</i>	160
CAP. 20 - RICERCHE LIMITATE A SOTTOCORPUS. DOMINIO DI RICERCA	163
<i>Lezione 40. Come definire dei sottocorpus</i>	163
<i>Lezione 41. Come effettuare ricerche limitate a uno o più sottocorpus</i>	166
CAP. 21 - ANALISI DELLE FREQUENZE	167
<i>Lezione 42. Come effettuare analisi delle frequenze</i>	167
CAP. 22 - LETTURA DEI TESTI. PRODUZIONE DI FORMARI E LEMMARI CON RELATIVI CONTESTI.	
GENERAZIONE DI INCIPITARI	170
<i>Lezione 43. Come visualizzare brani di testo sulla base dei riferimenti organici</i>	170
<i>Lezione 44. Come ottenere formario, formario esclusivo, lemmario, lemmario esclusivo o incipitario di un singolo testo</i>	170
<i>Lezione 45. Come ottenere formario, formario esclusivo, lemmario, lemmario esclusivo o incipitario di uno o più sottocorpus</i>	172
CAP. 23 - ACCESSO ALLE INFORMAZIONI BIBLIOGRAFICHE E STATISTICHE	173
<i>Lezione 46. Come prendere visione dei dati bibliografici relativi a testi inclusi in corpus o sottocorpus</i>	173
<hr/>	
PARTE V - OPERAZIONI DI GESTIONE DEI DATI	175
CAP. 24 - GESTIONE DEI CORPUS	175
<i>Lezione 47. Come modificare, compattare, copiare, cancellare un corpus. Come attivare l'ordinamento dei contesti per sintagmi.</i>	176
CAP. 25 - GESTIONE DEI TESTI	177
<i>Lezione 48. Come cancellare un testo da un corpus</i>	182
<i>Lezione 49. Come estrarre un testo lemmatizzato dal corpus per modifiche o copia in altro corpus</i>	183
<i>Lezione 50. Come modificare con un editore un testo lemmatizzato</i>	184
<i>Lezione 51. Come (re)inserire un testo modificato, lemmatizzato o meno</i>	185
<i>Lezione 52. Come modificare e reinserire un testo associato</i>	186
<i>Lezione 53. Operazioni di natura accessoria eseguibili sui testi</i>	186
<i>Lezione 54. Come effettuare una stessa operazione su più testi</i>	188
CAP. 26 - GESTIONE DEI LEMMI	189
<i>Lezione 55. Come modificare i dati di un lemma</i>	190
<i>Lezione 56. Come fondere due lemmi introdotti erroneamente come diversi</i>	191
<i>Lezione 57. Come sostituire con altro lemma un certo lemma associato a più occorrenze di una stessa forma</i>	192
<i>Lezione 58. Come rimuovere coppie forma-lemma mute e lemmi muti</i>	193
CAP. 27 - GESTIONE DEGLI IPERLEMMI	194

<i>Lezione 59. Come modificare o cancellare un iperlemma</i>	195
<i>Lezione 60. Come fondere due iperlemmi introdotti erroneamente come diversi</i>	195
<i>Lezione 61. Come navigare all'interno della struttura dell'iperlemmario</i>	196
CAP. 28 - GESTIONE DEGLI ARCHIVI BIBLIOGRAFICI	196
<i>Lezione 62. Come effettuare le operazioni di gestione di un archivio bibliografico</i>	197
PARTE VI - COMANDI E AZIONI COMUNI A PIU' AMBIENTI	199
CAP. 29 - COMANDI DI USO GENERALE	199
<i>Lezione 63. Come stampare il contenuto dell'accumulatore</i>	200
APPENDICE A - LIMITI DI GATTO	203
APPENDICE B - IMPOSTAZIONE DELLE OPZIONI GENERALI	204
APPENDICE C - CODICI UTILIZZABILI NEI FILE TESTO, NOTE, TRAD	207
APPENDICE D - AVVIO DI GATTO CON PARAMETRI	214
APPENDICE E - RISOLUZIONE DI PROBLEMI PARTICOLARI	216
APPENDICE F - GLOSSARIO	218
APPENDICE G - CARATTERI ANSI	225

INTRODUZIONE

CAP. 1 - PANORAMICA SU GATTO

FINALITÀ DEL PROGRAMMA

Dal 1985, col nome Centro Studi Opera del Vocabolario Italiano, l'attuale Istituto Opera del Vocabolario Italiano (OVI) del CNR ha ereditato dall'Accademia della Crusca il compito di realizzare il Vocabolario Storico della Lingua Italiana.

Il progetto iniziale era quello di un vocabolario che andasse dalle origini ai giorni nostri. In seguito apparve evidente che, data la limitatezza delle risorse disponibili, era opportuno procedere concentrandosi inizialmente sulla sezione cronologica antica, ovvero il Tesoro della Lingua Italiana delle Origini, limitando per il momento le risorse destinate alle fasi successive.

Sulla base di queste premesse, il lavoro svolto all'OVI prende le mosse da ciò che è edito in modo ragionevolmente affidabile e che interessi la lingua dei primi secoli, valutabile in circa 22 milioni di occorrenze. Si tratta di tradurre in formato elettronico questa vasta documentazione, previo un lavoro filologico sui testi che va dalla verifica dei criteri editoriali (necessaria per alcune indispensabili regolarizzazioni, ad esempio delle parentesi) fino talvolta ad una vera e propria verifica testuale. Nel contempo i testi vengono corredati delle necessarie codifiche. A quel punto è possibile inserirli nel corpus testuale dell'OVI.

La successiva operazione da effettuare è quella della lemmatizzazione. In assenza della possibilità di interrogare il corpus per lemmi il redattore è costretto a reperire il materiale cercandolo tra tutte le varianti grafiche che una forma può assumere; questa strada comporta elevate perdite di tempo e incertezza sulla completezza del risultato. L'inserimento della lemmatizzazione è in grado per contro di imprimere una drastica accelerazione al lavoro. A questo riguardo va notato che l'estrema varietà linguistica presente nei testi presi in considerazione rende del tutto inapplicabile il ricorso a lemmatizzatori automatici, risultando di gran lunga più efficiente una lemmatizzazione manuale, sia pure assistita dal software.

Il corpus, una volta corredato della lemmatizzazione, diviene la fonte di informazioni utilizzata dai redattori per redigere le voci del vocabolario storico.

L'esigenza di disporre di uno strumento in grado di gestire in maniera integrata le diverse fasi del lavoro ha portato alla nascita di GATTO, il cui nome è per l'appunto un acronimo che sta per **G**estione degli **A**rchivi **T**estuali del **T**esoro delle **O**rigini.

GATTO è nato quindi come strumento finalizzato alla costruzione, gestione ed interrogazione del corpus di testi che è alla base del Vocabolario Storico della Lingua Italiana in corso di realizzazione presso l'OVI.

Più in generale GATTO permette di effettuare ricerche lessicografiche su un qualunque archivio testuale a condizione che i testi costituenti, opportunamente codificati, siano disponibili su file unitamente ai relativi dati bibliografici.

Poiché GATTO gestisce corpus e non singoli testi (si possono però costruire corpus formati anche da un unico testo o limitare le operazioni a sottocorpus, ovvero a parti di un corpus), per utilizzarlo è necessario disporre di un corpus già esistente oppure costruirne uno ed introdurre i testi precedentemente preparati, cioè scritti su file e completi di codifiche. Un corpus può in seguito essere modificato per aggiunta o eliminazione di testi. È possibile disporre di più corpus da utilizzare alternativamente, però tutte le funzioni di GATTO si applicano di volta in volta ad un solo corpus; è possibile includere copie dello stesso testo in più corpus.

Ogni corpus al momento della creazione è solo un contenitore vuoto, destinato ad includere uno o più testi; ognuno di essi è tratto da un file contenente il testo vero e proprio dell'opera, il *file testo*, corredato internamente da opportune codifiche che indichino i numeri di pagina, le ripartizioni del testo, i brani in versi, le parti da escludere dalle ricerche e via dicendo. Per realizzare un file di questo tipo è sufficiente disporre di un editore o di un programma di videoscrittura in grado di produrre un file ANSI.

Per ogni testo che si intende introdurre in un corpus è poi necessario inserire i corrispondenti dati bibliografici all'interno di un archivio apposito gestito da GATTO.

GATTO consente di effettuare ricerche non solo per forme ma anche per lemmi: a tale scopo è previsto l'utilizzo di testi lemmatizzati. La lemmatizzazione viene effettuata mediante il modulo di lemmatizzazione incluso in GATTO; i testi lemmatizzati possono essere trasferiti in altri corpus completi della lemmatizzazione.

Una volta che un determinato corpus è stato costruito e dotato dei testi desiderati, eventualmente lemmatizzati, è possibile svolgere al suo interno ricerche lessicografiche ottenendo in risposta una serie di forme e/o lemmi e successivamente di contesti estratti dal corpus. Tali contesti possono essere visualizzati, stampati su carta, memorizzati su file.

GATTO è stato specificamente progettato e realizzato da **Domenico Iorio-Fili**, ricercatore presso l'OVI. La versione attuale del programma (3.3), scritto in Visual Basic 6.0, è la 10^a della serie, è costituita da circa 34.000 linee di codice (poco meno di 50.000 includendo la specificazione dell'interfaccia grafica, a sua volta basata su 30 finestre) ed ha una struttura fortemente modulare, necessaria al fine di rendere possibile il mantenimento dell'allineamento tra le diverse parti del programma, date le sue dimensioni.

L'esatta definizione delle problematiche da affrontare e un insostituibile contributo nel definire l'impostazione ottimale delle funzioni e dell'interfaccia in rapporto alle esigenze degli utenti, colleghi e non, sono stati forniti da **Valentina Pollidori**.

Allo sviluppo del programma ha dato un ampio contributo **Francesco Leoncino**.

Dell'organizzazione ed effettuazione di test e verifiche sperimentali e della ideazione di una vasta gamma di suggerimenti migliorativi essenziali si è fatto pressoché totalmente carico **Paolo Squillacioti**.

Inoltre lo sviluppo del programma ha potuto avvalersi delle prove sul campo e dei suggerimenti offerti da numerosi colleghi.

AMBIENTI

GATTO è internamente articolato in quattro parti logicamente separate, dette *ambienti*, corrispondenti ai gruppi di funzioni che il programma è chiamato a svolgere. Questi ambienti sono denominati **Ricerche**, **Lemmatizzazione sul corpus**, **Lemmatizzazione per singolo testo** e **Gestione della base di dati**; ad essi si accede partendo dalla finestra di avvio di GATTO. È possibile passare da un ambiente all'altro durante la stessa sessione di lavoro (per esempio, non è necessario uscire da GATTO per correggere un errore di lemmatizzazione scoperto durante un'interrogazione). Verrà ora fornita una breve panoramica di ciascun ambiente.

Ambiente Ricerche

Scopo di GATTO è quello di consentire ricerche lessicografiche all'interno di uno o più testi raccolti in corpus, o in sottoinsiemi di questi ultimi, *sottocorpus*, definiti dinamicamente.

La maggior parte delle ricerche ha come obiettivo finale la localizzazione, all'interno dei testi indicati, di occorrenze delle forme specificate. Altre operazioni consentono di estrarre informazioni di diversa natura, come liste di forme o di lemmi, incipitari, index locorum, statistiche di frequenze.

L'ambiente **Ricerche** è quello nel quale si effettuano tutte le ricerche di informazioni su un corpus precedentemente costruito ed eventualmente lemmatizzato in tutto o in parte. Le ricerche possono riguardare forme, lemmi, categorie grammaticali, iperlemmi. In questo ambiente si definiscono dinamicamente eventuali sottocorpus e si possono estrarre informazioni di altra natura, come formari o lemmari relativi a singoli testi, a singoli sottocorpus o a combinazioni di più sottocorpus, visualizzazioni di parti di testo, frequenze comparate di forme, uscite grafiche come distribuzione integrale delle occorrenze in un testo e distribuzione temporale dei contesti localizzati.

I risultati ottenuti possono essere stampati o copiati su file per un utilizzo successivo.

I testi di un corpus possono essere interrogati in blocco, singolarmente, o per gruppi definiti sulla base dei dati bibliografici. Si possono quindi effettuare ricerche alternativamente sul corpus o su una o più di queste selezioni, indicate col termine *sottocorpus*. La composizione di ogni sottocorpus, può venire memorizzata su disco per essere richiamata nelle sessioni successive. Si possono definire fino a 6 sottocorpus simultaneamente.

Per cercare un elemento (forma, lemma,...) sono disponibili dei riquadri (*selettori*) in ciascuno dei quali si può inserire una parola da cercare. È possibile fare uso di *caratteri jolly* per estendere la ricerca a più parole, così come è disponibile una serie di opzioni per modificare le modalità di ricerca.

Nel caso di ricerca per forme esiste la possibilità di prelevare un brano testuale da un documento qualunque con tecniche di *copia e incolla*, trasferendolo poi in una apposita finestra di ingresso di GATTO: il testo verrà analizzato e verranno ricercate nel corpus le occorrenze delle parole presenti al suo interno.

Ogni ricerca è normalmente realizzata in 4 tempi (in genere è anche possibile saltare i punti 2 e 3, passando direttamente dall'impostazione della ricerca ai contesti estratti):

1. impostazione della ricerca con indicazione degli oggetti cercati;

2. estrazione dagli archivi e visualizzazione in una griglia di tutte le forme o i lemmi che soddisfano le impostazioni iniziali;
3. copia o combinazione in una seconda finestra, ACCUMULATORE, degli elementi di effettivo interesse; l'operazione può essere ripetuta più volte, anche sulla base di ricerche di natura diversa (da cui il nome *Accumulatore*); la *copia* consente di combinare gli esiti di più ricerche in modalità OR, accettando cioè gli elementi che soddisfano almeno una delle ricerche impostate; la *combinazione* permette di mantenere nell'ACCUMULATORE solo gli elementi che vi comparivano già ma che, nel contempo, figurano anche tra gli elementi ottenuti con la nuova ricerca (modalità AND);
4. selezione dall'ACCUMULATORE di uno o più elementi ed estrazione dal corpus dei relativi contesti.

È disponibile una serie di opzioni per dimensionare e ritagliare i contesti che si intendono estrarre dal corpus e stampare su carta o copiare su file.

Il risultato finale della ricerca è una serie di contesti. Ogni contesto così ottenuto comprende:

- occorrenza localizzata,
- contesto,
- riferimento topografico,
- riferimento organico.

Ad ogni contesto sono associati i dati bibliografici del testo che lo contiene.

Se l'occorrenza è lemmatizzata sono accessibili le informazioni relative al lemma (ed eventualmente all'iperlemma) utilizzato.

La dimensione di ogni singolo contesto può essere ridefinita separatamente in base al numero delle parole o dei periodi costituenti o per selezione diretta delle parole di inizio e fine contesto; è anche possibile l'uso di contesti in modalità *kwic* (**KeyWord In Context**).

I contesti che verranno estratti al termine delle operazioni saranno presentati ordinandoli a scelta:

- in ordine cronologico del testo di appartenenza e, a parità di testo, in ordine alfabetico di forme localizzate;
- in ordine alfabetico di forme localizzate e, a parità di forma, in ordine cronologico del testo di appartenenza;
- in ordine cronologico del testo di appartenenza e, a parità di testo, in ordine di successione al suo interno.

L'ordinamento dei testi, *per default* effettuato su base cronologica, è modificabile ed è possibile utilizzare criteri diversi (ad esempio ordine alfabetico di autore o titolo).

Per alcuni tipi di ricerche sono previsti ordinamenti dei risultati un po' diversi.

In generale gli ordinamenti alfabetici relativi alle forme e/o ai lemmi possono basarsi sui loro caratteri iniziali, come ovvio, oppure sulle loro terminazioni, permettendo in questo caso di realizzare ordinamenti *inversi*, sia crescenti che decrescenti.

Il materiale così ricavato può essere estratto, tutto o in parte, e stampato su carta o copiato su file scritti in formato RTF (**Rich Text Format**).

I tipi di ricerche disponibili in GATTO sono:

- per forme,
- per forme mai lemmatizzate,
- per lemmi,
- per categorie grammaticali,
- per disambiguatori (particolari proprietà dei lemmi),
- concordanze lemmatizzate,
- index locorum per forme,
- index locorum lemmatizzati,
- per iperlemmi,
- cooccorrenze,
- estrazione di formari di singoli testi o di sottocorpus,
- estrazione di formari esclusivi di singoli testi o di sottocorpus,
- estrazione di lemmari di singoli testi o di sottocorpus,
- estrazione di lemmari esclusivi di singoli testi o di sottocorpus,
- stampa di incipitari di singoli testi o di sottocorpus,
- visualizzazione di parti di singoli testi,
- generazioni di curve forme-frequenze relative a singoli testi o sottocorpus,

- istogrammi della distribuzione temporale delle occorrenze di specifiche forme o gruppi di forme.

Ambiente Lemmatizzazione sul corpus

All'interno di GATTO è possibile procedere alla lemmatizzazione dei testi inclusi in un corpus. In GATTO *lemmatizzare* significa associare un certo lemma ad una specifica occorrenza di una specifica forma in uno specifico testo. Se il lemma utilizzato non era presente precedentemente nel lemmario del corpus vi viene aggiunto automaticamente.

Una volta portata a termine una lemmatizzazione, questa ha una duplice conseguenza: realizza una connessione tra il lemma e l'occorrenza cui è stato associato (e questo era ovvio) e crea inoltre una connessione generale tra il lemma e la forma di cui l'occorrenza era espressione.

Per avviare la lemmatizzazione si effettua una ricerca di contesti in maniera analoga a quanto si fa nell'ambiente **Ricerche**; la ricerca può essere impostata per forme, per lemmi, per categorie grammaticali o per disambiguatori e può essere estesa all'intero corpus o limitata a dei sottocorpus. I contesti ottenuti al termine della ricerca sono disponibili per essere lemmatizzati e su di essi si possono effettuare tutte le operazioni di lemmatizzazione desiderate; ogni singola operazione permette di associare un lemma a uno o più contesti, eventualmente anche a tutti quelli ottenuti con la ricerca, con un unico comando; le lemmatizzazioni possono essere successivamente corrette o rimosse.

Nel corso della lemmatizzazione è possibile, volendolo, associare a un'occorrenza oltre al lemma anche un iperlemma.

Ogni volta che nel corso di una lemmatizzazione vengono utilizzati un lemma o un iperlemma non presenti fino a quel momento nel corpus, questi vi vengono aggiunti automaticamente; in particolare il lemma entra a far parte del dizionario di macchina della forma lemmatizzata e come tale verrà proposto all'utente, come possibile scelta, nel corso di successive lemmatizzazioni riferite a quella stessa forma, anche se in testi diversi dello stesso corpus; un discorso analogo vale per l'iperlemma.

La lemmatizzazione è agevolata dall'uso del dizionario di macchina e da una *tabella degli omografi* operanti a livello di corpus, a loro volta modificati automaticamente in tempo reale dalle operazioni di lemmatizzazione.

Sempre in questo ambiente è possibile modificare o rimuovere le lemmatizzazioni precedentemente assegnate.

Ambiente Lemmatizzazione per singolo testo

Questa seconda modalità di lemmatizzazione è più specificamente mirata alla lemmatizzazione molto ampia o esaustiva di singoli testi.

La lemmatizzazione può essere effettuata in maniera *standard* o *sequenziale*: in tutti i casi opera su un solo testo per volta.

La lemmatizzazione *standard* consente di scegliere la forma ed eventualmente la specifica occorrenza che si intende lemmatizzare e di passare quindi alle altre occorrenze della stessa forma o alle forme alfabeticamente adiacenti.

La lemmatizzazione *sequenziale* passa in rassegna, consentendone la lemmatizzazione, le parole che compongono il testo nell'ordine in cui vi compaiono. Questo è probabilmente lo strumento più rilevante offerto da questa seconda modalità di lemmatizzazione, specie se usato in combinazione con la maniera *standard* (tenere presente, a questo riguardo, la Lezione 20).

Nel corso della lemmatizzazione è possibile, volendolo, associare all'occorrenza oltre al lemma anche un iperlemma.

Valgono tutte le altre considerazioni fatte per la **Lemmatizzazione sul corpus**; in particolare, entrambe le modalità di lemmatizzazione sono agevolate dall'uso del dizionario di macchina e della *tabella degli omografi*.

Ambiente Gestione della base di dati

Questo è l'ambiente nel quale è possibile costruire, modificare, rimuovere la maggior parte degli elementi costituenti la base di dati su cui GATTO opera, ovvero corpus, testi, lemmi, iperlemmi, dati bibliografici. I file testo, una volta preparati e testati, possono essere usati per introdurre in un corpus i testi che lo costituiranno, eventualmente corredati di apparati e versioni in altra lingua. I corpus possono essere modificati tramite successivi inserimenti o rimozioni di testi. Inoltre è possibile accedere al lemmario e all'iperlemmario del corpus per correggere errori o per inserire o rimuovere iperlemmi di qualunque livello. La lista che segue riassume, schematicamente, le operazioni possibili.

Operazioni sui corpus:

- creazione,
- modifica di alcune caratteristiche,
- duplicazione,
- compattazione,
- cancellazione.
- attivazione dell'ordinamento per sintagmi

Operazioni sui testi:

- controllo di correttezza, inserimento o sostituzione in un corpus di *testi base* ed eventualmente *testi associati di tipo Note o Trad* contenenti apparati, traduzioni o diverse edizioni,
- cancellazione dal corpus di testi,
- estrazione dal corpus di file contenenti i testi e le eventuali lemmatizzazioni.

Operazioni sui lemmi:

- correzione dei dati dei lemmi,
- fusione di due lemmi in uno,
- sostituzione di un lemma in una coppia forma-lemma,
- rimozione lemmi e coppie forma-lemma muti.

Operazioni sugli iperlemmi:

- inserimento di un iperlemma di livello qualunque,
- correzione dei dati di un iperlemma,
- modifica delle connessioni di un iperlemma con gli iperlemmi di livello superiore o con gli iperlemmi di livello inferiore o i lemmi,
- cancellazione di un iperlemma.

Operazioni sugli archivi bibliografici:

- creazione,
- inserimento di un nuovo testo,
- modifica dei dati bibliografici di un testo,
- rimozione di un testo,
- cancellazione.

STRUTTURE DATI

Forme

Per *forma* si intende ogni singola parola, distinta dalle altre esclusivamente in base all'assetto grafico, che può occorrere un qualsiasi numero di volte in un corpus o in un testo.

Il concetto di *forma* non deve essere confuso con quello di *occorrenza*, che rappresenta un evento, ovvero una singola comparsa di una forma in un testo.

Perciò, ad esempio, il brano *visualizzazione di numeri di pagina* è costituito complessivamente da 5 occorrenze di 4 forme in tutto:

- due occorrenze della forma *di*
- una occorrenza della forma *numeri*
- una occorrenza della forma *pagina*
- una occorrenza della forma *visualizzazione*

L'elenco delle forme diverse presenti in un corpus è racchiuso in un archivio detto *formario*. Non viene fatta differenza tra lettere maiuscole e minuscole, per cui ad esempio le stringhe *Paolo*, *paolo* e *PAOLO* sono riferite alla stessa forma.

Le forme possono essere *monorematiche* o *polirematiche*. Normalmente tutte le parole prelevate dai testi sono considerate forme monorematiche; appositi codici permettono di raccogliere più parole di testo in forme polirematiche, trattate e lemmatizzabili come un tutt'uno. Ad esempio la stringa "Nicola di

Bonaccorso" genera normalmente nel formario le 3 forme "Nicola", "di" e "Bonaccorso", lemmatizzabili e ricercabili separatamente; se la stringa viene invece racchiusa tra i codici che definiscono una forma polirematica, GATTO memorizzerà nel formario l'unica forma "Nicola di Bonaccorso".

Una *forma* può essere costituita da una qualunque sequenza di lettere, con e senza diacritici.

In fase di creazione di un nuovo corpus è possibile specificare se si vuole che le forme possano o meno contenere anche le cifre arabe; in questo caso tutti i numeri presenti nei testi del corpus, così come le eventuali sequenze miste di lettere e cifre, verranno indicizzati.

Lemmi

Il *lemma* è l'elemento cui viene riportato un insieme di forme che si distinguono fra loro soltanto per l'assetto grafico (varianti grafiche, con o senza valore di varianti fonetiche) e/o perché sono forme della flessione dello stesso verbo o sostantivo o aggettivo. Corrisponde di regola, ma non necessariamente, ad un'entrata del vocabolario.

Anche nel lemmario non si fa distinzione fra lettere maiuscole e minuscole, mentre è possibile tenere distinti lemmi omografi precisandone le categorie grammaticali; ad ogni lemma è poi associabile una stringa disambiguatrice, detta appunto *disambiguatore*, da usare in presenza di lemmi omografi aventi anche uguale categoria grammaticale; infine, ad ogni lemma è associabile un breve commento, il quale è un attributo il cui contenuto è regolarmente visualizzato, ma che non contribuisce a definire il lemma o a distinguerlo dagli altri. In sostanza, affinché in GATTO due lemmi siano riconosciuti come distinti essi devono differire per l'entrata lessicale e/o per la categoria grammaticale: nel caso di due lemmi diversi ma caratterizzati da uguaglianza di entrambi questi elementi è possibile introdurre nel campo disambiguatore di uno di essi o di entrambi stringhe arbitrarie finalizzate a distinguerli. Mentre nel definire un nuovo lemma entrata lessicale e categoria grammaticale vanno obbligatoriamente assegnati, disambiguatore e commento sono attributi facoltativi.

Tutti questi dati possono essere modificati dall'utente, mentre l'inserimento e la cancellazione di lemmi dal lemmario sono operazioni normalmente controllate in modo automatico da GATTO.

Quando negli ambienti di lemmatizzazione si associa un lemma ad una specifica occorrenza di una forma, viene creata un'associazione lemma-forma valida in tutto il corpus, che permetterà di risalire alla forma partendo da una ricerca per lemmi, e viceversa. Poiché allo stesso lemma verranno probabilmente collegate occorrenze di più forme, effettuando una ricerca (per lemmi) del lemma considerato si potranno localizzare simultaneamente le occorrenze di tutte (o alcune delle) forme collegate con esso; sarà inoltre possibile, volendolo, includere o escludere da tale ricerca le occorrenze specificamente lemmatizzate col lemma in questione, quelle lemmatizzate con altri lemmi, quelle non lemmatizzate.

Così come la prima associazione di un lemma ad un'occorrenza inserisce automaticamente il nuovo lemma nel lemmario del corpus, la rimozione dell'ultima lemmatizzazione che ne fa uso, per modifica della lemmatizzazione o cancellazione del testo dal corpus, produce in genere l'automatica cancellazione del lemma dal lemmario.

Fanno eccezione a questo meccanismo i cosiddetti *lemmi muti*, trattati al termine di questo paragrafo.

Un altro meccanismo di inserimento di lemmi nel lemmario è legato all'acquisizione al corpus di testi già lemmatizzati in precedenza, magari estratti da un altro corpus: in questo caso ogni testo porta con sé i propri lemmi e questi vengono automaticamente inglobati nel lemmario, qualora non già presenti.

L'insieme dei lemmi associati ad una forma nell'ambito di un corpus ne costituisce il *dizionario di macchina*. Questo dizionario, utilizzato in lemmatizzazione, è gestito in maniera quasi del tutto automatica: ogni voce del dizionario, ovvero ogni associazione forma-lemma, nasce al momento del suo primo abbinamento ad un'occorrenza in un testo del corpus e viene normalmente rimossa al momento in cui l'ultimo utilizzo di tale associazione risultasse rimosso dal corpus.

Un *lemma* può essere costituito da una qualunque sequenza di lettere, con e senza diacritici.

In fase di creazione di un nuovo corpus è possibile specificare se si vuole che i lemmi possano o meno essere costituiti anche da cifre arabe.

Lemmi muti (e coppie forma-lemma mute)

Il meccanismo di gestione dinamica del lemmario e del dizionario di macchina adottato in GATTO, ottimale nella maggior parte dei casi, ne rende però difficoltosa l'applicazione ad un corpus diverso da quello sulla base del quale tali archivi sono stati costruiti.

Per ovviare a questa difficoltà è stata introdotta una nuova categoria di lemmi e collegamenti forma-lemma, detti *muti* perché sono presenti negli archivi senza però essere riferiti ad alcuna occorrenza del

corpus; in altre parole, data una coppia forma-lemma muta, non esiste nel corpus nessuna occorrenza della forma lemmatizzata col lemma; analogamente non esiste nessuna occorrenza del corpus lemmatizzata con un lemma *muta*.

Gli elementi *muti* sono strumenti che un corpus può ereditare da un altro già lemmatizzato e sfruttarli ai fini della propria lemmatizzazione.

Iperlemmi

Si immagini che in un corpus sia stato definito un certo numero di lemmi relativi ad animali (*cane, gatto, lepre, bue, aquila, falco, fringuello, piccione,...*). Una ricerca relativa ad occorrenze di forme associate ad un lemma indicante un mammifero potrebbe essere impostata, ad esempio, come ricerca per lemmi di cane, gatto, ecc., con qualche rischio di incompletezza nel caso non si citino tutti i lemmi presenti. Esiste la possibilità di associare una volta per tutte a ciascuno dei primi 4 lemmi dell'esempio un *iperlemma*, diciamo l'iperlemma *mammifero*, con un meccanismo analogo a quello con cui si associano i lemmi alle forme.

A questo punto, una ricerca per iperlemmi, a partire dall'iperlemma *mammifero*, porterebbe agli stessi risultati di una ricerca per lemmi in cui sia stato specificato di cercare tutte le occorrenze relative a *cane, gatto, lepre e bue*.

Sempre con riferimento all'esempio iniziale, in modo analogo si potrebbe associare l'iperlemma *uccello* a *aquila, falco, fringuello, piccione*, con conseguenze analoghe.

Estendendo il concetto, possiamo pensare di introdurre l'iperlemma *animale*, associandolo non tanto a dei lemmi quanto agli iperlemmi *mammifero* e *uccello*. Una ricerca per iperlemmi dell'iperlemma *animale* porterebbe a localizzare tutte le occorrenze delle forme lemmatizzate con i lemmi a loro volta associati, risalendo questo schema di connessioni (tramite *mammifero* e *uccello*), all'iperlemma *animale*.

Nella terminologia di GATTO *mammifero* e *uccello* sono detti *iperlemmi di livello 1*, in quanto connessi direttamente con dei lemmi; *animale* è invece un *iperlemma di livello 2*, in quanto *punta* a iperlemmi di livello 1. È possibile definire liberamente iperlemmi di livelli superiori, purché esista una catena di connessioni che partendo da essi *scenda* fino ad individuare uno o più lemmi presenti nel corpus.

Un secondo uso degli iperlemmi consiste nell'associarli direttamente a specifiche occorrenze. Gli iperlemmi utilizzati o creati in questa circostanza sono sempre di livello 1. Corrispondentemente è possibile impostare ricerche per iperlemmi specificando che le occorrenze che si intende localizzare non sono tutte quelle lemmatizzate con i lemmi associati agli iperlemmi specificati ma solo quelle espressamente associate agli iperlemmi indicati. Come conseguenza, questo secondo tipo di ricerca per iperlemmi è limitato alle occorrenze associate a specifiche terne forma-lemma-iperlemma.

In conclusione, un iperlemma può avere contemporaneamente connessioni con iperlemmi di livello superiore e con iperlemmi di livello inferiore (o lemmi se è un iperlemma di livello 1). Nell'esempio, *mammifero* (iperlemma di livello 1) è connesso *verso l'alto* (cioè al livello 2) con *animale* e *verso il basso* con *cane, gatto, lepre, bue*.

GATTO identifica univocamente un iperlemma mediante 3 attributi: assetto grafico (obbligatorio), disambiguatore (facoltativo), livello (obbligatorio).

Testi

Un *testo* per GATTO consiste nel contenuto di un file testo. All'interno di GATTO un testo è individuato univocamente da una sigla formata da uno a tre caratteri alfanumerici.

Normalmente un testo coincide con un documento letterario completo. Niente impedisce però di ripartire in più file, e quindi in più *testi* per GATTO, un'unica opera: si potrebbe ad esempio avere interesse a ripartire in tre file la *Commedia*, assegnando ad ognuna delle tre cantiche una diversa sigla. Al contrario, si potrebbe inserire in uno stesso file, trattandola quindi come un unico documento, una raccolta di novelle di uno stesso autore.

Va comunque tenuto presente che ad ogni file testo è associato un record di bibliografia, quindi un unico autore, un unico titolo e via dicendo.

Materialmente un testo è contenuto in un file testo (detto anche *testo base*) posto nella *directory dei testi*. Tale file, scritto in ANSI, va preparato usando un editore o un programma di videoscrittura adatto. Eventuali note andranno inserite in un altro file ANSI detto *testo associato di tipo Note*. Una eventuale seconda edizione o traduzione in altra lingua può essere inserita in un terzo file ANSI detto *testo associato di tipo Trad*.

Corpus e sottocorpus

Un corpus è l'ambito testuale entro il quale si svolgono le operazioni effettuate con GATTO, siano esse lemmatizzazioni, ricerche, modifiche di dati, ecc. Ogni corpus è individuato da un nome formato da 1 a 12 caratteri alfanumerici. Fisicamente un corpus viene collocato in una directory il cui nome deriva da quello del corpus (es. il corpus Dante, inserito nel disco C, verrà collocato automaticamente nella directory C:\dante.gat). Si possono avere corpus con nomi diversi nello stesso disco o anche corpus aventi lo stesso nome, posti in tal caso su dischi diversi.

Gli elementi di un corpus sono i testi via via introdotti al suo interno. È possibile raggruppare i testi di un corpus in sottocorpus definiti dinamicamente sulla base dei loro dati bibliografici. Si possono definire contemporaneamente fino a 6 sottocorpus. Uno stesso testo può appartenere a più sottocorpus.

La struttura di un sottocorpus, intesa come lista dei testi costituenti, può essere salvata su file e richiamata in una successiva sessione per riutilizzare lo stesso sottocorpus.

Uscendo da un corpus senza aver salvato gli eventuali sottocorpus su file questi ultimi andranno perduti, nel senso che in una successiva sessione andranno ridefiniti, se ancora necessari.

Corpus e sottocorpus possono essere formati anche da un solo testo.

Le ricerche possono essere effettuate all'interno del corpus, di un sottocorpus o di una combinazione di più sottocorpus; eventuali duplicati nei risultati verranno rimossi automaticamente.

Un corpus può contenere oltre 16.000 testi e 2 miliardi di occorrenze.

Archivi bibliografici

Ad ogni testo incluso in un corpus è associato un pacchetto di informazioni bibliografiche (titolo, autore, edizione, sigla, ecc.) che va inserito in un record di una base di dati apposita, indicata coi termini *bibliografia* o *archivio bibliografico*.

Di conseguenza tra le funzionalità di GATTO vi sono la creazione e la cancellazione di archivi bibliografici, l'inserimento, la modifica e la cancellazione di record all'interno di un archivio bibliografico.

All'atto della creazione di un nuovo corpus all'utente viene chiesto di indicare il file contenente l'archivio bibliografico da associare a quel corpus: se tale file al momento non esiste viene creato.

Ad un corpus è associato un unico archivio bibliografico, che alla fine dovrà quindi contenere tutti i record relativi ai testi del corpus; nell'archivio possono essere presenti ulteriori record, relativi a testi non inclusi nel corpus e magari appartenenti a corpus diversi. A seconda degli ambienti GATTO consente di esplorare l'intero archivio bibliografico o di vederne solo la parte costituita dai record associati ai testi del corpus di volta in volta considerato.

È sempre possibile sostituire l'archivio bibliografico associato ad un corpus: è compito dell'utente verificare che anche il nuovo archivio contenga tutti i record necessari; in caso contrario le operazioni che richiederanno accesso a testi privi di dati bibliografici verranno fermate e il problema segnalato.

Come detto in precedenza, uno stesso archivio bibliografico può essere utilizzato da corpus diversi.

Fisicamente un archivio bibliografico è un file, di nome e collocazione scelti dall'utente.

AVVIO DI GATTO CON PARAMETRI PER USO LIMITATO ALLE RICERCHE SU DI UN UNICO CORPUS PREDEFINITO

Esiste la possibilità di utilizzare GATTO in modalità di sola interrogazione limitata ad un corpus predefinito.

Avviando il programma con questa modalità si libera l'utente dalla necessità di selezionare il corpus da utilizzare; in tal caso sono però predefiniti anche il tipo di ricerca con il quale si comincia e gli elementi da cercare.

Mediante questa tecnica è possibile avviare GATTO da un altro programma o dalla linea comando DOS, passandogli direttamente come parametri le indicazioni relative alla prima ricerca da effettuare. Una volta 'entrati' nel programma è possibile continuare ad utilizzarlo effettuando tipi diversi di ricerche, ma sempre e soltanto sul corpus preimpostato all'avvio.

Quasi tutti i tipi di ricerca possono essere avviati mediante questa modalità, che permette peraltro di fare uso di tutte le opzioni disponibili con l'accesso standard.

I parametri da fornire al programma sono:

- 1 lettera identificativa del disco nel quale si trova il corpus da interrogare (es. "C"),
- 2 nome del corpus da interrogare (es. "demo"),
- 3 codice che denota il tipo di ricerca iniziale da effettuare (es. "f" per ricerca per forme),
- 4 elementi da cercare (es. "casa" per cercare tale forma).

I dettagli dell'avvio di GATTO con parametri, la sintassi da seguire e alcuni esempi di chiamata sono riportati nell'Appendice D - *Avvio di GATTO con parametri*.

CAP. 2 - INTRODUZIONE ALL'USO DI GATTO

CONSIDERAZIONI SULL'USO DEL PROGRAMMA

GATTO opera con tutte le versioni di Windows a 32 bit, a partire da Windows 95.

I requisiti di memoria centrale dipendono essenzialmente dal sistema operativo in uso. Nel caso di costruzione di corpus testuali di dimensioni rilevanti, i requisiti di memoria considerati minimi per le diverse versioni di Windows potrebbero rivelarsi non del tutto adeguati.

Il programma crea ed utilizza directory aventi nomi scelti dall'utente ed estensione **gat**. È consigliabile, se possibile, evitare la presenza nei dischi del PC e in quelli delle macchine ad esso collegate in rete locale di altre directory aventi un nome terminante con l'estensione **gat**.

GATTO si appoggia alle *estensioni* presenti nei nomi di file (ovvero ai caratteri finali, successivi al punto, come exe, txt, bat, doc, ..) per indicare i testi cui i file si riferiscono; per questo motivo è spesso necessario per l'utente conoscere tali estensioni. *Gestione risorse* di Windows in partenza è configurato in modo da non mostrarle: si suggerisce di modificare senz'altro questa opzione di *Gestione risorse* tramite i suoi menu [VISUALIZZA + OPZIONI CARTELLA.. + VISUALIZZA + NASCONDI LE ESTENSIONI DEI FILE PER I TIPI DI FILE CONOSCIUTI]: questa ultima opzione va **disattivata** (a seconda delle diverse versioni di Windows percorsi e diciture per questa operazione potrebbero differire).

Uscire sempre dal programma usando gli appositi comandi. Le interruzioni realizzate diversamente possono avere effetti dannosi sull'integrità della base di dati. È comunque fortemente consigliato di disporre di una copia di salvataggio aggiornata della base di dati stessa: a questo riguardo tenere presente che le operazioni effettuate negli ambienti **Lemmatizzazione sul corpus**, **Lemmatizzazione per singolo testo** e **Gestione base dati** possono modificare i dati preesistenti, mentre le ricerche si limitano a leggerli senza modificarli.

Per ogni corpus costruito, la base di dati della quale è opportuno avere una copia di salvataggio è costituita da:

1. cartella (il cui nome ha estensione **gat**) in cui viene inserito il corpus, insieme alle sue sottocartelle,
2. cartella contenente l'archivio bibliografico,
3. cartella contenente i testi (*directory dei testi*).

GATTO è un programma registrato ed è protetto dalle vigenti leggi italiane ed internazionali sulla tutela del software.

GATTO può essere liberamente usato e riprodotto per scopi di studio e ricerca senza fini di lucro, con l'obbligo della citazione in tutte le pubblicazioni che ne derivino e in tutte le applicazioni nelle quali venga utilizzato.

Qualunque altro uso deve essere oggetto di un accordo scritto preventivo con l'Istituto Opera del Vocabolario Italiano.

(GATTO - Copyright 8-2-99 - Registrazione n. 001172.)

PROTEZIONE DEI TESTI

Vi possono essere casi nei quali l'utente che opera con GATTO ha interesse a costruire e fornire ad altri un corpus testuale gestito dal programma consentendo loro di utilizzarlo per estrarne informazioni lessicografiche, senza tuttavia permettere che i destinatari possano disporre in forma integrale dei testi inclusi.

GATTO rende possibile questo tipo di rapporto tra costruttore ed utenti di un corpus combinando due meccanismi:

- **cifratura dei testi.** Tutti i riferimenti interni ai testi inclusi in un corpus GATTO sono cifrati, per cui non è possibile risalire ai testi in chiaro partendo dai dati distribuiti nei diversi file e dal database del corpus;

- **vincoli all'estrazione dei testi dal corpus.** Se nel corpus è inserito un testo, è normalmente possibile generare una copia in chiaro del testo stesso (vedere ad esempio la Lezione 49); tuttavia questa operazione è tecnicamente possibile solo se nella *directory dei testi* del corpus il file testo relativo è già presente: in altre parole, può ottenere una copia del testo solo chi lo possiede già in partenza. Notare che questa condizione non rende priva di senso l'estrazione di un testo dal corpus, in quanto la copia estratta, a differenza di quella di partenza, risulterà arricchita dell'eventuale lemmatizzazione operata successivamente all'inserimento del testo nel corpus.

LIMITAZIONI NEGLI ACCESSI SIMULTANEI ALLO STESSO CORPUS

Si possono mettere in esecuzione contemporaneamente più copie di GATTO, aprendo corpus diversi, senza problemi di alcun genere.

Se invece le copie di GATTO accedono allo stesso corpus, o se quest'ultimo è collocato su un server di rete e più utenti intendono accedervi allo stesso tempo, vi sono alcune limitazioni di accesso da tenere presente, in dipendenza dagli ambienti dai quali si apre il corpus:

- **accesso da ambiente Ricerche:** l'accesso in multiutenza allo stesso corpus è possibile senza difficoltà,
- **accesso da ambiente Lemmatizzazione per singolo testo:** l'accesso in multiutenza è consentito, ma solo un utente per volta può lemmatizzare un certo testo;
- **accesso da ambienti Lemmatizzazione sul corpus e Gestione base dati:** l'accesso è in modalità esclusiva. Ciò significa che perché un utente possa aprire un corpus bisogna che nessun altro utente vi sia collegato; una volta aperto il corpus da questi ambienti nessun altro utente (o altra copia di GATTO dallo stesso utente) potrà accedervi finché il corpus non verrà chiuso o fino a che quell'utente non passerà ad utilizzarlo a partire da un ambiente diverso.

LIMITAZIONI NEGLI ACCESSI SIMULTANEI ALLA STESSA BIBLIOGRAFIA

Quanto detto nel paragrafo precedente in relazione agli accessi multipli ad uno stesso corpus vale, in termini molto simili, in caso di uso dello stesso file di bibliografia, da parte di più corpora. Più precisamente:

- **uso da ambiente Ricerche:** l'uso dello stesso file di bibliografia da parte di più corpora è possibile senza difficoltà,
- **uso da ambienti Lemmatizzazione e Gestione base dati:** l'uso è in modalità esclusiva. Ciò significa che perché un utente possa utilizzare un corpus bisogna che nessun altro corpus sia collegato alla stessa bibliografia; una volta aperto il corpus da questi ambienti nessun altro utente (o altra copia di GATTO dallo stesso utente) potrà accedere a corpora facenti uso dello stesso file di bibliografia.

Volendo aprire contemporaneamente due o più corpora che fanno uso della stessa bibliografia, con almeno una apertura da ambiente Lemmatizzazione o Gestione base dati, fare preventivamente una copia della bibliografia (file con estensione mdb) rinominarla o collocarla altrove e reindirizzare verso quest'ultima, ad esempio, i corpora aperti in modalità Ricerche.

DATI E SOFTWARE ACCESSORI NECESSARI PER USARE GATTO

Come detto in precedenza per utilizzare GATTO è necessario disporre in una forma qualunque dei testi sui quali si vogliono operare ricerche lessicografiche: GATTO infatti non contiene al suo interno testi, ma è solo un gestore di testi forniti dall'utente. Se si vuole fare riferimento ad una specifica edizione, sarà necessario disporre di quest'ultima al fine di introdurre i corrispondenti caratteri di fine riga e i numeri di pagina. Nella misura in cui se ne vuole fare uso è poi necessario disporre dei relativi dati bibliografici.

È importante sapere che non tutti i programmi di videoscrittura disponibili sul mercato sono adatti per scrivere i file testo utilizzati da GATTO. In particolare uno dei più diffusi, ovvero Microsoft Word, a partire dalla versione 97 è divenuto inadeguato a scrivere ed editare tali file. Il motivo è legato ai convertitori associati a Word:

- il formato **doc** (formato base di Word) inserisce una serie di caratteri di controllo invisibili per l'utente ma non per l'interprete di Gatto che non è in grado di gestirli;
- il formato **txt** (solo testo o solo testo con layout) al momento del salvataggio del testo sostituisce automaticamente, senza che lo si possa impedire, una serie di caratteri: ad esempio, « e » vengono

modificati in "; © viene modificato in (c), ecc. (non e' stato possibile trovare nella documentazione di Word una lista, oltre a una motivazione, per queste sostituzioni). Il rischio e' quindi che codici e caratteri la cui presenza all'interno del testo e' necessaria siano modificati senza possibilità di ripristino e senza che l'utente se ne accorga;

- il formato **txt** (solo testo con interruzione di linea), qualora la lunghezza di una riga di testo superi quella del righello di Word, inserisce senza possibilità di impedirlo una interruzione di linea, che altera la struttura della pagina per quanto riguarda il numero di riga rispetto all'edizione.

L'unico formato utilizzabile sarebbe il formato txt (solo testo) del vecchio Word 6.0, il quale peraltro non e' in grado di gestire i *nomi lunghi* dei file, del tipo di filgat.xxx.txt. Da queste considerazioni deriva il suggerimento di utilizzare programmi come Notepad, o meglio Wordpad, per scrivere e editare i file testo o, se si vuole uno strumento sofisticato, editori più potenti come ad esempio KEDIT del Mansfield Software Group.

USO E CONSERVAZIONE DEI *FILE TESTO*¹

Con questo termine si indicano i file, già citati ad esempio nel paragrafo precedente, che contengono materialmente i testi da inserire in un corpus. Come detto nel paragrafo *Testi* del capitolo precedente i file testo vanno collocati in un'apposita cartella detta *directory dei testi*.

A dispetto della loro importanza, i file testo vengono utilizzati da GATTO solo in due circostanze:

1. inserimento di un testo, lemmatizzato o meno, nel corpus;
2. estrazione di un testo dal corpus (operazione che ha un senso solo per testi lemmatizzati successivamente al loro inserimento nel corpus).

In tutti gli altri casi, ad esempio durante le operazioni di lemmatizzazione o di interrogazione degli archivi, i file testo non vengono usati e potrebbero anche non essere presenti all'interno della directory dei testi.

Caso di assenza di modifiche alla lemmatizzazione

Da quanto detto segue che se un testo non è stato lemmatizzato o, pur essendolo in partenza, non si ha interesse a modificarne la lemmatizzazione, l'uso del materiale testuale si articola in due fasi successive:

- I. il testo viene inserito nel corpus: in questa fase il file testo è usato per estrarne le informazioni testuali e deve essere presente nella directory dei testi;
- II. il corpus è interrogato, senza fare uso del file testo, che può non essere presente nella directory dei testi.

Se si presentasse la necessità di modificare il testo sarebbe sufficiente, ad esempio, rimuovere il testo dal corpus, modificare il file testo e ripetere l'inserimento nel corpus sulla base del file testo modificato.

Caso di effettuazione di modifiche alla lemmatizzazione

Diversa e leggermente più delicata è la situazione nel caso in cui, successivamente alla sua introduzione nel corpus, un testo venga lemmatizzato o più in generale ne venga modificato lo stato di lemmatizzazione rispetto al contenuto iniziale del file testo (mediante aggiunte, correzioni o rimozioni di lemmatizzazioni o iperlemmatizzazioni).

In questo caso infatti la procedura per apportare eventuali modifiche al testo dopo che è stato inserito nel corpus è la seguente:

- I. il testo viene estratto dal corpus. In questa fase il file testo originale deve essere presente nella directory dei testi e il suo contenuto deve risultare rigorosamente inalterato rispetto a quando è stato usato per inserire il testo nel corpus: infatti il file testo usato in quella circostanza costituisce la "*chiave*" di riconoscimento del testo stesso, che GATTO utilizza per autorizzare l'estrazione del testo dal corpus;
- II. il file testo così estratto, contenente la lemmatizzazione nella versione modificata, viene sostituito al file testo originale nella directory dei testi;
- III. a questo punto si possono apportare tutte le modifiche volute nel nuovo file testo;

¹ Per comodità in questo paragrafo col termine *file testo* si intende in realtà una coppia di file, ovvero **filgat.xxx.txt** e **codlemmi.xxx**, dove xxx è la sigla del testo relativo.

IV. la nuova versione del testo viene inserita nel corpus in luogo della precedente (per *sostituzione* oppure per *cancellazione* seguita da *inserimento*): durante questa fase il nuovo file testo deve essere presente nella directory dei testi e diviene la nuova "**chiave**" di riconoscimento del testo.

(Data la sua importanza, la descrizione di questa procedura compare anche in altre parti del manuale).

Ciò che deve essere assolutamente chiaro è che **in presenza di alterazioni del file testo rispetto alla sua versione originale, il testo correntemente inserito in un corpus non potrà più esserne estratto, rendendo impossibile sia apportare correzioni al testo di partenza che inserire in un altro corpus il testo, completo della lemmatizzazione apportatavi dopo il suo inserimento nel corpus attuale.**

Quindi si suggerisce di conservare copie di sicurezza dei file testo (intendendo con questo i file filgat.xxx.txt e codlemmi.xxx, citati in altre parti del manuale, dove xxx indica la sigla del testo), collocate possibilmente in posti diversi dal disco rigido in uso (floppy disc, CD, dischi rigidi di altri PC,...).

AIUTO IN LINEA

Al programma è associata una guida in linea che può essere richiamata in ogni momento. Esistono due modi per attivarla:

1. tramite i menu [**Guida**] dove presenti: la Guida in linea viene avviata a partire dalla pagina iniziale ("Il programma GATTO");
2. mediante il tasto <F1>: viene visualizzata la specifica pagina della Guida relativa alla finestra aperta al momento e all'operazione in corso di esecuzione.

La Guida in linea contiene un Sommario che funge da tutorial per quanto riguarda gli argomenti principali. Anche se il Sommario non è in grado di sostituire il manuale cartaceo, costituisce comunque un utile punto di riferimento per quanto riguarda la descrizione di operazioni complete.

La Guida contiene anche un piccolo dizionario di termini usati al suo interno. Tra questi ve ne sono alcuni particolarmente tecnici: ad ogni prima citazione di uno di essi in una pagina della Guida è sempre associato un collegamento che permette di visualizzarne la definizione senza abbandonare la finestra aperta al momento.

La maggior parte delle pagine della Guida in linea si riferisce a corrispondenti finestre di GATTO: in tali casi compaiono i paragrafi *Come arrivare qui* e *Come proseguire* che agevolano la comprensione della struttura del programma e possono guidare l'utente nel corso delle operazioni.

STRUTTURA DI QUESTO MANUALE

All'interno della più generale finalità di gestire e utilizzare una base di dati testuale, GATTO consente di espletare una varietà di funzioni, talvolta interconnesse, modulabili tramite una serie di opzioni. Il programma si presenta di conseguenza caratterizzato da una notevole articolazione interna, solo in parte semplificata dalla strutturazione in ambienti logicamente separati. Come detto, questo spettro di operazioni eseguibili tramite GATTO deriva dalla sua natura di strumento di lavoro destinato ad essere applicato alle diverse fasi dell'attività dell'OVI. Il singolo utente non avrà probabilmente le stesse esigenze e si limiterà ad usare un sottoinsieme delle funzioni disponibili in GATTO. In ogni caso l'uso del programma non è semplice e richiede una fase di apprendimento, sia pure breve e limitata alle cose di interesse.

In conseguenza delle considerazioni precedenti il manuale di GATTO è stato pensato come *tutorial* del programma. Oltre a questa introduzione il manuale è diviso in 6 parti più alcune appendici, in funzione della logica di uso. Ogni parte si articola in *capitoli* a loro volta contenenti delle *Lezioni*. Le parti iniziali dei capitoli, di impostazione più teorica, contengono spiegazioni di massima sul funzionamento di GATTO ed introducono i concetti che verranno poi ripresi nelle *Lezioni*; queste ultime, più tecniche, completano le informazioni fornite in precedenza dando un quadro completo dei comandi e delle operazioni previste; nella maggior parte dei casi le *Lezioni* contengono una o più esercitazioni pratiche relative ai concetti illustrati: a questo scopo le *Lezioni* contengono in genere uno o più esempi di uso del programma che l'utente può semplicemente leggere oppure provare ad eseguire; poiché la maggior parte degli esercizi è riferita al corpus testuale **Demo** fornito insieme al programma l'utente ha la possibilità di verificare la correttezza dei risultati che ottiene comparandoli con quelli riportati nel manuale.

Le *Lezioni* relative alla gestione della base di dati utilizzeranno invece di regola il corpus **Prova** che l'utente stesso verrà guidato a costruire e a popolare con i tre testi **es1**, **es2** e **es3** ottenibili, via Web o CD, unitamente a questo manuale.

Di seguito vengono riassunti i contenuti delle parti in cui il manuale è diviso.

Parte I - Preparazione dei dati

Per usare GATTO è necessario costruire (almeno) un corpus, inserendovi uno o più testi con relative informazioni bibliografiche. Per effettuare queste operazioni bisogna per prima cosa disporre dei dati necessari, la cui preparazione è l'oggetto della prima parte del manuale.

L'argomento centrale è la scrittura del file testo, ovvero della sua versione digitale, con la descrizione dei codici diacritici utilizzati per inserirvi le informazioni necessarie (ad esempio numeri di pagina): l'inserimento dei codici diacritici è quella che viene chiamata tecnicamente *marcatura* del testo.

Un testo può anche essere corredato da note, così come talvolta capita di disporre di una traduzione in altra lingua o di una seconda edizione che si vuole inserire per metterla a confronto con quella base: in questi casi GATTO prevede la costruzione di file aggiuntivi, la cui preparazione è trattata insieme ai relativi codici diacritici.

Completano la prima parte del manuale la descrizione dei dati bibliografici richiesti e dei parametri che andranno assegnati al corpus al momento della costruzione.

Parte II - Costruzione di un corpus e inserimento di testi

Una volta predisposti i dati necessari è possibile passare alla effettiva costruzione di un nuovo corpus, inteso inizialmente come contenitore vuoto destinato ad includere futuri testi, e di un archivio bibliografico. Subito dopo si può passare ad inserire nel corpus i testi precedentemente preparati, eventualmente corredati da testi accessori (detti *testi associati*) contenenti eventuali note e/o diverse edizioni o traduzioni.

L'utente viene guidato a costruire un corpus e un archivio bibliografico, ad inserire testi e testi associati nel corpus, fronteggiando situazioni di errore appositamente inserite a titolo di esercizio.

Parte III - Lemmatizzazione e inserimento di iperlemmi

Questa parte riguarda solo coloro che sono interessati a lemmatizzare i testi introdotti nel corpus.

Vengono descritti natura e uso di lemmi e iperlemmi e successivamente vengono spiegate in dettaglio le tecniche disponibili per associare questi elementi alle forme presenti nei testi inseriti nel corpus.

L'uso di lemmi e iperlemmi come oggetti di ricerca e la loro estrazione in connessione con i testi, ad esempio per un eventuale trasferimento in un altro corpus o per consentire modifiche o correzioni ai testi lemmatizzati, sono invece trattati in parti successive di questo manuale.

Parte IV - Estrazione di informazioni lessicografiche

Si tratta evidentemente della parte centrale, in quanto tutte le altre operazioni eseguibili con GATTO hanno significato in funzione delle ricerche lessicografiche da eseguire all'interno di un corpus testuale.

Vengono inizialmente passate in rassegna le modalità di ricerca previste in GATTO, illustrandone le fasi da quella iniziale di scelta della modalità desiderata fino alla estrazione dei risultati mediante visualizzazione, stampa su carta o scrittura su file. Nella maggior parte dei casi descritti i risultati consistono in contesti estratti dai testi del corpus.

GATTO consente di definire dinamicamente fino a 6 sottocorpus contemporaneamente. Pressoché tutti i tipi di ricerche possono essere limitati ai testi appartenenti a uno o più di tali insiemi. Vengono mostrate le modalità per definire i sottocorpus e per utilizzarli in fase di ricerca; viene anche mostrato come salvarne la struttura per riutilizzarli nelle sessioni successive senza la necessità di definirli nuovamente.

Un'altra categoria di informazioni ottenibili è quella relativa alle frequenze delle forme. Viene spiegato come ottenere e mettere a confronto le frequenze in testi o gruppi di testi diversi (incluso il corpus completo).

Si passa poi a descrivere la ricerca di informazioni testuali di altra natura, ovvero non definibili in termini di contesti estratti, quali formari, lemmari, incipitari, index locorum; tali informazioni possono essere riferite a singoli testi, singoli sottocorpus, insiemi di più sottocorpus, intero corpus.

Infine sono descritte le azioni necessarie per risalire alle informazioni bibliografiche e statistiche relative ai singoli testi del corpus, oltre ai dati statistici e alle impostazioni di quest'ultimo.

Parte V - Operazioni di gestione dei dati

La base di dati costruita mediante la creazione di un corpus e l'inserimento di uno o più testi al suo interno non costituisce una struttura rigida e può essere modificata mediante una serie di operazioni di gestione dei dati, descritte in questa parte del manuale.

Gestione dei corpus: un corpus può essere rimosso, con tutto il suo contenuto; alcune delle impostazioni inizialmente assegnate ad un corpus possono essere modificate.

Gestione dei testi: un testo (lemmatizzato o meno) inserito in un corpus può essere estratto (senza rimuoverlo dal corpus), modificato e reinserito al suo interno sostituendone la versione corrente, oppure può essere rimosso definitivamente; stesse possibilità valgono per eventuali testi associati. Nuovi testi possono essere aggiunti.

Gestione dei lemmi: l'inserimento di lemmi può dar luogo, nel tempo, alla presenza nel lemmario di lemmi errati, sostanzialmente duplicati o associati, magari ripetutamente, a forme ed occorrenze sbagliate. Apposite funzioni consentono di correggere queste situazioni senza perdere il lavoro di lemmatizzazione già fatto.

Gestione degli iperlemmi: questi ultimi possono essere inseriti, modificati, rimossi dall'iperlemmario del corpus.

Gestione degli archivi bibliografici: vengono descritte le fasi di creazione, modifica e cancellazione di archivi bibliografici e di inserimento, modifica e cancellazione di record al loro interno.

Parte VI - Comandi e azioni comuni a tutti gli ambienti

Esistono alcune operazioni ricorrenti, comuni a più ambienti di GATTO, come la selezione di uno o più elementi da una tabella visualizzata, la stampa, parziale o completa, del suo contenuto, la chiusura delle finestre.

Per non costringere il lettore a sfogliare il manuale alla ricerca del punto in cui queste operazioni sono presentate, è stato deciso di riunirle e trattarle in un'apposita parte del manuale.

Appendici

Appendice A: riporta i principali limiti operativi di GATTO, come numero massimo di testi inseribili in un corpus.

Appendice B: impostazione delle opzioni generali. GATTO dispone di un file di inizializzazione in cui sono memorizzati alcuni valori di partenza utilizzati per default dal programma, come la directory in cui stampare i file contenenti i contesti, oppure il numero di parole costituenti i contesti da visualizzare. Questi valori possono essere modificati dall'utente di volta in volta ma ad ogni riavvio GATTO riparte dal contenuto del file di inizializzazione. Volendo modificare in maniera permanente questi default (ferma restando la possibilità di modificare comunque in seguito i valori durante la sessione), si può fare ricorso alla finestra OPZIONI DI PERSONALIZZAZIONE: l'Appendice B elenca i dati personalizzabili e spiega come modificarli.

Appendice C: elenca in forma compatta e descrive brevemente le codifiche utilizzate nei file testo e nei testi associati. L'argomento è descritto compiutamente nei capitoli 3 e 4.

Appendice D: descrive la sintassi da usare per l'avvio di GATTO con parametri. Questa modalità di avvio è utile quando si vuole chiamare GATTO da un altro programma, preimpostando sia il nome del corpus da interrogare che il tipo di ricerca e gli elementi da cercare inizialmente.

Appendice E: tratta di alcuni problemi di natura piuttosto particolare che, sia pure molto raramente, possono presentarsi agli utenti di GATTO e che non dipendono dal codice. L'appendice spiega come riconoscerli e come risolverli.

Appendice F: contiene un piccolo glossario dei termini più tecnici usati nel manuale.

Appendice G: elenca i codici ANSI delle lettere accentate e spiega come inserire nei testi tutte quelle non presenti direttamente sulla tastiera del computer.

CONVENZIONI USATE NEL MANUALE

I nomi delle finestre sono indicati in maiuscoletto. Es. APERTURA DEI CORPUS.

I menu sono indicati in grassetto maiuscoletto ridotto, tra parentesi quadre. Es. [BIBLIOGRAFIA].

Le scelte comportanti il passaggio attraverso menu e sottomenu sono indicate usando "+" come separatore tra di essi. Es: [RICERCHE.. + OPERAZIONI SU SINGOLO TESTO + FORMARIO].

I pulsanti sono indicati in grassetto tra parentesi tonde. Es: (Annulla).

I tasti della tastiera sono indicati in grassetto tra parentesi angolari. Es: <Invio>.

I termini più tecnici sono riportati in corsivo: normalmente è possibile trovarne una breve descrizione nel Glossario (Appendice F).

I termini *cartella* e *directory* sono usati come sinonimi.

Dove necessario si supporrà che il corpus **Demo** sia stato installato nel disco C, l'archivio bibliografico nella cartella C:\demobiblio e la *directory dei testi* sia C:\demotesti.

Spesso, particolarmente nelle parti che descrivono i meccanismi di lemmatizzazione, i termini *occorrenza* e *contesto* verranno usati come sinonimi.

STRUMENTI DI CONTROLLO PRESENTI NELL'INTERFACCIA GRAFICA

Clic e doppio clic del mouse

Le finestre di GATTO non supportano il doppio clic del mouse. Tutti gli oggetti presenti nell'interfaccia grafica (menu, griglie, pulsanti, ecc.) possono essere pilotati con un singolo clic sul tasto sinistro del mouse, talvolta su quello destro. Il doppio clic, qualora eseguito, verrebbe interpretato dal codice come lo stesso comando dato 2 volte in rapida successione: il codice stesso include una serie di protezioni volte ad evitare possibili malfunzionamenti derivanti da questa azione; tuttavia, dato l'elevato numero di *oggetti grafici* presenti nell'interfaccia di GATTO e potenzialmente in grado di rispondere in maniera non prevedibile al doppio clic, si raccomanda di non eseguire questa manovra.

Menu

Comando posto nella parte più alta di una finestra. Lo si attiva cliccandolo col tasto sinistro del mouse o digitando <Alt> seguito dal tasto corrispondente al carattere sottolineato presente nella dicitura. Ad esempio, per attivare [SOTTOCORPUS] digitare <Alt> e poi <s>; per attivare [RICERCHE.. + PER FORME] digitare <Alt>, poi <r> e infine <f>. Alcuni menu riportano, accanto alla dicitura, una combinazione di uno o più tasti di scelta rapida la cui digitazione permette di attivarli senza fare uso del mouse.

Pulsante (o 'bottone')



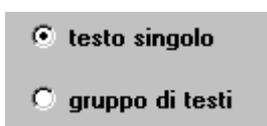
Elemento grafico presente in una finestra e caratterizzato da una scritta descrittiva dell'azione che verrà svolta dal programma cliccando sul bottone stesso col tasto sinistro del mouse. Per attivarlo si può anche digitare <Alt> seguito dal tasto corrispondente al carattere sottolineato nella dicitura, dove presente.

Check box



Riquadro che permette di attivare o disattivare una certa opzione del programma effettuando un singolo clic al suo interno. Un *check box* è attivato quando evidenzia al suo interno la spunta √. Lo stato di attivazione di ogni *check box* è indipendente da quello di altri *check box* eventualmente presenti nella stessa finestra. Per attivarlo si può anche digitare <Alt> seguito dal tasto corrispondente al carattere sottolineato nella dicitura, dove presente.

Option box



Finestrina circolare che permette di selezionare un'opzione del programma, scegliendo tra più alternative, ognuna controllata da uno specifico *option box*; gli *option box* tra i quali va operata una scelta mutuamente esclusiva sono racchiusi, ed eventualmente separati da altri *option box* presenti nella stessa finestra, da appositi pannelli o riquadri. Per attivare l'opzione associata ad un *option box*, disattivando automaticamente le opzioni alternative, fare clic su di esso. Un *option box* è attivato quando evidenzia al suo interno un dischetto nero.

Box a tendina



Elemento grafico normalmente composto da un riquadro e una freccia, diretta in basso, posta alla sua destra (immagine più a sinistra); cliccando sulla freccia apparirà una lista di dati alternativi tra i quali scegliere (immagine più a destra); un clic su uno di questi lo inserirà nel riquadro in alto come *dato selezionato*. Una volta effettuata la selezione, la lista dei dati del box a tendina si richiude e resta visibile il solo dato selezionato.

LEZIONE 1. ESEMPIO DI USO DI GATTO: LETTURA DI STATISTICHE E IMPOSTAZIONI DEL CORPUS DEMO; RICERCA DELLE FORME *UOMO* E *UOMINI*

In questa Lezione, così come nelle successive, si suppone che sia stato installato il corpus **Demo** fornito in associazione con GATTO.

Scopo di questa Lezione è aprire il corpus **Demo**, vederne statistiche e impostazioni ed effettuare al suo interno la ricerca delle forme *uomo* e *uomini*, arrivando ad ottenerne le occorrenze e a stampare su carta i relativi contesti.



Per cominciare, avviare GATTO utilizzando l'icona  se presente sul video o partendo dal pulsante **Avvio** (o **Start**) in basso a sinistra fino a selezionare la scritta GATTO.

Nella finestra di avvio di GATTO (SCELTA AMBIENTE) cliccare su [**RICERCHE**] (primo menu a sinistra) per entrare nell'ambiente omonimo, quindi usare [**CORPUS**] per accedere alla finestra APERTURA DEI CORPUS. Cliccare su un punto qualunque della riga relativa al corpus **Demo** (che apparirà evidenziata), quindi aprire il corpus con [**APRI**]. L'apertura del corpus è confermata dall'abilitazione del menu [**INFORMAZIONI..**].

LETTURA DI STATISTICHE E IMPOSTAZIONI DEL CORPUS DEMO

Per visualizzare le statistiche del corpus **Demo** utilizzare [**INFORMAZIONI.. + STATISTICHE**]: si vede così che il corpus è formato da 12 testi, per complessive 145.659 occorrenze di 17.903 forme diverse; per quanto concerne la lemmatizzazione, questa è stata applicata a 21.379 occorrenze e fa uso di 5.276 lemmi diversi; sono anche presenti 9 iperlemmi.

Cliccando sull'icona di chiusura **x** in alto a destra la finestra si chiude.

Per vedere le impostazioni iniziali del corpus **Demo** usare invece [**INFORMAZIONI.. + IMPOSTAZIONI**]: si trova che in **Demo** i segni **?**, **!** e **...** sono considerati punteggiatura debole, viene fatto uso dei numeri di pagina e di riga, ma non delle cifre arabe all'interno di forme e lemmi; si possono anche vedere le lunghezze massime consentite per le forme (**26** caratteri), i lemmi (**26**), le categorie grammaticali (**10**), ecc.. Usare (**Esci**) per chiudere questa finestra.

Siamo così tornati alla finestra APERTURA DEI CORPUS. Il corpus in uso attualmente è "**C:demo**", come si può leggere nella testata della finestra: a questo punto è possibile procedere con l'effettuazione di ricerche all'interno di **Demo**. Per farlo, uscire dalla finestra con [**CHIUDI FINESTRA**].

RICERCA DELLE FORME *UOMO* E *UOMINI*

Ci proponiamo ora di trovare tutte le occorrenze delle forme *uomo* e *uomini* all'interno di **Demo**.

Per cominciare, entri nella finestra MODALITÀ RICERCA impostare [RICERCHE.. + PER FORME]. Sotto la dicitura in giallo **forme da cercare** compaiono 10 rettangoli a sfondo bianco (*selettori*) nei quali si possono digitare appunto le parole da cercare. Digitare *uomo* nel primo selettore, poi portarsi nel secondo col tasto <↓> o cliccandovi col mouse e digitare *uomini*; avviare la ricerca con [AVVIO RICERCA]. Si arriva così alla finestra FORME LOCALIZZATE la quale mostra le forme, presenti nel corpus, soddisfacenti la richiesta formulata e le relative occorrenze. Si trova così che in **Demo** ci sono 11 occorrenze di *uomini* e 56 di *uomo*.

Prima di procedere nella ricerca, a titolo di esercizio facciamo un passo indietro e torniamo alla precedente finestra RICERCA PER FORME tramite [VAI A.. + IMPOSTAZIONE RICERCA]; cancelliamo il contenuto dei due selettori usati con [CANCELLA TUTTI I DATI] e proviamo ad impostare la stessa ricerca utilizzando, in forma più compatta, un solo selettore nel quale andrà digitata la stringa: **uom<o,ini>** la quale predispone la ricerca di due forme di uguale radice *uom*, la prima completata da *o* e la seconda da *ini*; riavviamo la ricerca col tasto <Invio> (alternativo al menu [AVVIO RICERCA]): come si può vedere si ottiene lo stesso risultato di prima.

Tramite (**Mostra lemmi**) possiamo anche avere un quadro più completo per quanto riguarda la lemmatizzazione di queste occorrenze: si vede così che una delle 11 occorrenze di *uomini* e 3 delle 56 di *uomo* sono lemmatizzate col lemma *uomo* (*s.m.*). A titolo di esempio, supponiamo di scartare le 10 occorrenze non lemmatizzate di *uomini* e di procedere nella ricerca con le altre: per fare questo cliccare, nella colonna **sel.**, in corrispondenza della seconda, terza e quarta riga della tabella, avendo cura di tenere sempre premuto il tasto <Ctrl>; come risultato comparirà in ognuna delle tre righe una **x**, ad indicare le coppie forma-lemma selezionate.

A questo punto si passa alla fase successiva, ovvero al trasferimento delle coppie forma-lemma selezionate nell'ACCUMULATORE, tramite [COPIA IN ACC.]; le tre coppie costituiscono l'intero contenuto attuale dell'ACCUMULATORE, poiché non erano state fatte altre ricerche in precedenza.

Rimane l'ultimo passaggio, ovvero la localizzazione e presentazione dei contesti relativi alle occorrenze di queste coppie forma-lemma. Selezionare le tre righe dell'ACCUMULATORE con [SEL/DESEL.. + SELEZIONA TUTTO] (il campo **sel.** delle tre righe contiene ora **x**) e procedere con [MOSTRA CONTESTI].

Si giunge così alla finestra RISULTATI RICERCA, contenente i contesti cercati; con [CONTESTI.. + NUM. CONT. TROVATI, SELEZ.] vediamo che i contesti localizzati sono 57 (dato consistente con i valori trovati in precedenza: 1 + 53 + 3 = 57); cliccando su **x** la finestra si chiude. Ogni contesto è descritto su 3 righe, una di testata e due per il contesto, con la parola localizzata evidenziata in rosso.

Tramite [CONTESTI.. + TIPO + KWIC (RIF. TOP.)] si può ottenere una diversa visualizzazione dei contesti, nella quale ad ognuno di essi è riservata una sola riga a video e le parole localizzate sono mostrate incolonnate al centro; con [CONTESTI.. + TIPO + KWIC (RIF. ORG.)] si passa ad una rappresentazione analoga, mentre con [CONTESTI.. + TIPO + STANDARD] si torna a quella di partenza.

Il contesto n. 2 è il primo tratto da un'opera il cui titolo abbreviato è "Novellino, XIII u.v...". Cliccando su un punto qualunque di questo titolo abbreviato si apre una finestra contenente i dati bibliografici e statistici dell'opera: si vede così che il testo, individuato dalla sigla **od**, è di autore anonimo e si colloca nell'area toscana, più precisamente fiorentina; facendo scorrere i dati della finestra le statistiche ci dicono tra l'altro che esso consta di 25.983 occorrenze di 4.573 forme diverse; col tasto <Esc> la finestra si richiude.

Sempre restando al contesto n. 2, si nota che all'estremo destro della sua riga di testata compare il simbolo **@**, la cui presenza sta ad indicare che questa occorrenza di *uomini* è stata lemmatizzata. Cliccando su tale simbolo si ritrova l'informazione relativa al lemma usato (*uomo*, *s.m.*); col tasto <Esc> o cliccando su **x** la finestra si richiude.

Se ci portiamo sul contesto n. 5 e clicchiamo in un punto qualunque delle due righe di contesto, si aprirà una finestra diversa, in cui questo contesto è mostrato con una diversa geometria: in particolare è mostrato il periodo cui appartiene l'occorrenza trovata. Cliccando sul bottone (**Allarga**) verranno aggiunti al contesto due periodi, uno all'inizio e uno alla fine, portando la lunghezza complessiva del contesto a 3 periodi, come indicato nel box in basso. Se a questo punto decidiamo di stampare il contesto lungo 3 periodi, cliccare sul 5 (in alto a sinistra) indicante il numero di contesto o alla sua sinistra: il numero di contesto verrà visualizzato in rosso, preceduto da una **P** che indica che il contesto è stato selezionato per la stampa con una configurazione a periodi. A questo punto possiamo tornare alla finestra RISULTATI RICERCA col bottone (**Esci**) in basso a destra: come si può vedere, la selezione del contesto è rimasta. Selezioniamo per la stampa anche i contesti n. 2, 3 e 4, sempre cliccando sui numeri di contesto o alla loro sinistra; in questi casi ai

contesti viene associata la lettera **N**, ad indicare che il formato di stampa è quello a parole (in questo caso 31 parole).

Per procedere alla stampa effettiva dei quattro contesti fin qui selezionati usare il menu **[STAMPA]**, cliccare sul *check box* **stampante** per indicare il tipo di uscita e infine sul bottone **(OK)** per eseguire l'operazione.

Si può notare che i 4 contesti sono stati stampati mantenendo per ciascuno la configurazione del contesto attiva al momento della sua selezione (la numerazione progressiva dei contesti stampati non è in relazione con quella riportata a video e ha valore puramente indicativo del numero di contesti stampati).

Per terminare la sessione, chiudere GATTO utilizzando **[ESCI. + FINE]**.

PARTE I - PREPARAZIONE DEI DATI

CAP. 3 - PREPARAZIONE DI UN FILE TESTO

PREMESSE

Ricordando quanto detto nell'introduzione un *testo* (individuato univocamente nella procedura mediante una sigla formata da 1 a 3 caratteri alfanumerici) per GATTO corrisponde al contenuto di un *file testo*: normalmente quest'ultimo include, in forma digitale, un documento letterario completo. Niente impedisce però di ripartire in più file, e quindi in più *testi* per GATTO, un'unica opera: si potrebbe ad esempio avere interesse a ripartire in tre file la Commedia, assegnando ad ognuna delle tre cantiche una diversa sigla. Al contrario, si potrebbe inserire in uno stesso file, trattandola quindi come un unico documento, una raccolta di novelle di uno stesso autore.

Una volta operata la scelta iniziale di cosa si intenda esattamente per *testo* da inserire in GATTO si può passare alla trasposizione su file del materiale di cui si è in possesso. Qualora questo sia già in una versione elettronica, quindi su file, sincerarsi che questa sia scritta in ANSI (i caratteri accettati da Windows) e che sia priva di codifiche introdotte da programmi di videoscrittura o di altra natura; un modo per rendersene conto può essere quello di aprire il file con Notepad o Wordpad: GATTO vedrà esattamente quello che vedono e mostrano questi due editori.

Sempre nell'introduzione sono stati spiegati i motivi che sconsigliano l'uso di programmi di videoscrittura come Microsoft Word, che introducono nei testi modifiche in conflitto con quanto richiesto da GATTO; è opportuno fare uso di semplici editori, come ad esempio Wordpad.

In ogni caso i file testo prodotti devono contenere solo **testo non formattato**, scritto cioè in ANSI puro; per questo motivo le istruzioni di assegnazione di specifici tipi e dimensioni ai caratteri (es. Times New Roman 12) saranno inutili, in quanto destinate ad essere perdute al momento del salvataggio del file. Stesse considerazioni valgono circa il ricorso a parti sottolineate, in grassetto, in corsivo o barrate, anch'esse inutili.

Il file testo, ovvero il file destinato a contenere in formato elettronico il testo da inserire in un corpus, dovrà essere denominato obbligatoriamente **filgat.<sigla>.txt**, dove <sigla> indica una stringa formata da 1, 2 o 3 caratteri alfanumerici: tale stringa identificherà sempre, in maniera univoca, tale testo all'interno del corpus. Ad esempio, il testo **a6** sarà contenuto nel file **filgat.a6.txt**.

In questo capitolo verrà mostrato come digitare il testo da inserire in un file testo: in particolare verrà spiegato in dettaglio quali sono i caratteri ammessi all'interno delle parole e come codificare quelli non presenti nella tavola ANSI, sulla quale si basa la costruzione dei testi.

Si suggerisce di utilizzare i file testo (filgat.<sigla>.txt) riportati nella directory dei testi del corpus **Demo** come esempi di applicazione delle codifiche descritte nel seguito.

ELEMENTI PRESENTI IN UN TESTO

Un testo può essere considerato come un insieme costituito da elementi di natura diversa, schematizzabile come segue:

- titolo, autore ed altri dati bibliografici,
- riferimenti organici,
- parti di testo immediatamente successive ai singoli riferimenti organici e afferenti ad essi.

Mentre titolo, autore e altri dati bibliografici vanno collocati nell'archivio bibliografico, riferimenti organici e parti di testo associate costituiscono gli elementi da inserire fisicamente nel file testo.

A sua volta il testo afferente ad un certo riferimento organico può essere strutturato in:

- testo costituito da parole che si vogliono memorizzare per farne successivamente oggetto di lemmatizzazione e ricerca (questa è ovviamente la parte essenziale del testo),
- testo escluso dall'indicizzazione (ad esempio numeri romani),
- codici diacritici non appartenenti al testo originale ma necessari per fornire le opportune informazioni a GATTO.

NOTE RELATIVE ALLA SCRITTURA DI UN FILE TESTO

Nel preparare un file testo è necessario rispettare alcune regole:

- ricordarsi di inserire il codice % (riferimento organico) all'inizio,
- terminare ogni riga con <Invio>: GATTO non impagina automaticamente il testo adeguandolo alle finestre di Windows, ma vuole che sia espressamente indicata la fine di ogni riga di testo,
- subito dopo la prima riga², avviata dal codice %, iniziare il testo col codice \$nn\$, dove nn è il primo numero di pagina (a meno che all'atto della creazione del corpus non si sia escluso l'uso di pagine e righe),
- il titolo non fa parte del file testo, quindi non vi va inserito,
- tutte le scritte presenti nelle pagine di copertina, le prefazioni, le introduzioni ecc. non devono essere inserite a meno che non si intenda trattarle come parti integranti del testo e come tali interrogabili per forme, lemmi, ecc.,
- le righe del file dovranno coincidere, come struttura, con quelle del testo, salvo nel caso di parola spezzata tra due righe: questa situazione non è ammessa e la parola in oggetto dovrà essere mantenuta intera, scrivendola come parola finale della prima delle due righe o come iniziale della seconda,
- un singolo periodo, ovvero la parte di testo racchiusa tra due interpunzioni forti, come punto, due punti, punto e virgola, non può essere costituito da più di 32767 parole; questa regola va tenuta attentamente presente nel caso si vogliano gestire testi privi di interpunzioni,
- figure, grafici, tabelle e didascalie non devono essere inseriti,
- evitare l'uso di indentazioni e tabulazioni, sostituendole con sequenze di spazi,
- nel *file testo* si può inserire una gerarchia nelle ripartizioni del testo (tipo capitoli divisi in paragrafi a loro volta divisi in sottoparagrafi): i *riferimenti organici* possono essere organizzati in una struttura a più livelli (massimo 5),
- fare attenzione, nell'applicare le codifiche, all'impostazione iniziale del corpus relativa alla ammissibilità o meno delle cifre arabe all'interno delle forme da indicizzare,
- le righe iniziali di un file testo contengono i codici che forniscono le informazioni iniziali all'interprete: se qualcuno di questi è mancante o inserito nell'ordine sbagliato si possono generare messaggi di errore con conseguente rifiuto di inserire il testo nel corpus. Vedere a questo proposito la parte finale della Lezione 5, *Inizio del file testo.*

RIFERIMENTI TOPOGRAFICI

In questo manuale si farà sempre l'ipotesi che i numeri di pagina e di riga vengono utilizzati, in quanto è la scelta più pratica nell'uso di GATTO. Quando, come risultato di una ricerca, vengono visualizzati o stampati dei contesti, insieme ad ognuno di essi vengono riportati anche i relativi riferimenti organici e topografici; i primi derivano direttamente dal testo scritto dall'autore, mentre i secondi variano in funzione dell'edizione. Qualora per motivi di rispetto dei diritti di editore si ritenga di non visualizzare i riferimenti topografici (pagina e riga nella pagina), la cosa è ottenibile impostando opportunamente un certo parametro all'atto della creazione del corpus: in quel caso i riferimenti topografici non verranno visualizzati se presenti e non verrà segnalato alcun errore se non saranno stati inseriti nei file testo. Se invece si decide di fare uso dei riferimenti topografici nel corpus, questi debbono essere obbligatoriamente inseriti nel file testo.

Riassumendo:

Impostazione del corpus	Riff. top. nel file testo	Risultato
visualizza riferimenti. topog.	presenti	i riff. top. vengono mostrati
	assenti	errore: <i>file testo</i> rifiutato
non visualizza riferimenti top.	presenti	i riff. top. non compaiono
	assenti	

² O le prime righe.

IMPAGINAZIONE DEL FILE TESTO

Vale la pena di ripetere che GATTO nel memorizzare il file testo non tiene conto dell'impaginazione automatica delle finestre di Windows (anche se il programma di videoscrittura usato per scrivere il file testo ne dà una rappresentazione visiva costantemente adeguata alla larghezza della finestra), ma si aspetta che l'utente imposti manualmente la fine di ogni riga tramite il tasto <Invio> (o <Enter>); in caso contrario il testo verrà considerato inserito tutto in un'unica riga, dando luogo a malfunzionamenti.

LEZIONE 2. CARATTERI VALIDI NEI FILE TESTO ALL'INTERNO DELLE PAROLE. FORME POLIREMATICHE. ORDINAMENTO ALFABETICO DELLE FORME. DIGITAZIONE DELLE LETTERE ACCENTATE. SEGNI DI INTERPUNZIONE

CARATTERI AMMESSI NELLE PAROLE DEL TESTO

Non tutti i 255 caratteri previsti dalla tabella ANSI sono accettati come validi all'interno di un file testo; bisogna inoltre distinguere tra caratteri accettati all'interno di una parola, segni di interpunzione validi e caratteri ammessi nel testo ma non utilizzabili all'interno di una parola: questi ultimi verranno visualizzati nei contesti ma non potranno fare parte di parole indicizzabili e quindi cercabili; appartengono a questa categoria, ad esempio, i simboli di frazione e i caratteri speciali come €, ®, ©.

Lettere con e senza accenti

All'interno di una forma indicizzabile sono accettate le 26 lettere dell'alfabeto inglese, con tutte le accentazioni, dieresi, cediglie, ring previsti nella tavola ANSI.

J e y con accento grave

In alcuni testi compaiono j e y, maiuscole e minuscole, con accento grave. Queste combinazioni non sono previste nella tavola ANSI; per indicarle bisogna far seguire la y o la j dal carattere "`" (ANSI 96) (quindi ad esempio per indicare *Y con accento grave*, digitare nel file testo "Y`"); anche nella visualizzazione di forme e contesti verrà utilizzata questa modalità di rappresentazione.

T con cediglia

Un'altra combinazione presente in alcune edizioni, ma non nella tavola ANSI, è la lettera t maiuscola o minuscola con cediglia. Per indicarla all'interno del file testo far seguire la t da un "6" o da "," (ANSI 184); quindi, ad esempio, "t6", "t,", "T6" o "T," verranno riconosciute come t con cediglia e rappresentate visivamente nei contesti rispettivamente con "t," e "T,", dove il secondo carattere è ancora l'ANSI 184. Nel caso di corpus impostati per accettare le cifre arabe all'interno delle forme le codifiche "t6" e "T6" non possono essere usate col significato sopra descritto e bisogna utilizzare necessariamente quelle basate sul carattere "," (ANSI 184).

Punto di abbreviazione. Punto sostitutivo di carattere illeggibile

In certe circostanze il carattere *punto* non è considerato come segno di interpunzione ma come un normale carattere appartenente alla parola con cui è a contatto. Questo si verifica quando il punto ha il significato di punto di abbreviazione (es. "sig.") oppure quando è usato come simbolo sostitutivo di un carattere illeggibile (si possono quindi avere anche più punti consecutivi; es.: "Bar...lo del' Isstorna"). In questi casi ogni punto va digitato sostituendolo col carattere '8' (es. "sig8" e "Bar888lo del' Isstorna"); nelle uscite di GATTO il testo verrà peraltro visualizzato correttamente con i caratteri punto (es. "sig." e "Bar...lo del' Isstorna").

Se la codifica è inserita in un corpus impostato per accettare le cifre arabe all'interno delle forme, in luogo della codifica "8" va usata "©" (ANSI 169).

Se il punto di abbreviazione o sostitutivo viene inserito in una parola inclusa in un campo formula la codifica è diversa ("&."): vedere l'argomento *Campi formula* nella Lezione 5.

Notazione anglosassone

Uno dei segni utilizzati all'interno delle parole, ovvero l'*apice*, è caratterizzato dalla proprietà di poter avere due effetti alternativi sulla struttura delle parole con cui è a contatto. Per indicare al programma quale effetto si intende applicare si ricorre ad un campo speciale che va sotto il nome di *notazione anglosassone*:

questo campo è definito da due codifiche diverse che ne marcano inizio e fine. L'apice avrà un effetto diverso a seconda che si trovi in parole incluse in una zona di testo in notazione anglosassone o meno.

Il codice di inizio di una zona di testo in notazione anglosassone, che può includere singole parole, parti di testo o interi testi, è **&A**; il codice di fine è **&I**.

Apice

L'apice (ANSI 39) può essere utilizzato come normale segno di interpunzione (vedere *Interpunzioni* nella Lezione 5). Altrimenti questo carattere appartiene alla parola con cui è a contatto, per cui non vi sono ambiguità interpretative se l'apice si trova all'inizio o alla fine di una parola, o anche se costituisce una parola da solo.

Diversa è la situazione quando l'apice si trova all'interno di una parola: in questo caso costituisce con essa un'unica forma se si è in notazione anglosassone; altrimenti la parola viene spezzata in due, formate dai caratteri rispettivamente precedenti e successivi all'apice, e l'apice diviene l'ultimo carattere della forma precedente.

La situazione è descritta dalla tavola seguente:

testo	forme estratte
'	'
inver'	inver'
'namoramento	'namoramento
domanda'lo	domanda' lo
&A...domanda'lo...&I	domanda'lo

Trattino

Il trattino "-" può comparire come normale segno di interpunzione agente come separatore tra due parole (e in questo caso andrà digitato come un normale trattino). Questo stesso segno è talvolta usato con significato opposto, cioè come convenzione grafica per tenere unite parti di una stessa parola o per segnalare un prolungamento grafico della consonante finale. In questi casi il trattino va codificato digitando al suo posto un "7"; nella successiva visualizzazione la parola apparirà scritta con un normale trattino. Ad esempio "voleva-ly" va inserito nel file testo con "voleva7ly" e verrà poi visualizzato come "voleva-ly".

Se la codifica è inserita in un corpus impostato per accettare le cifre arabe all'interno delle forme, in luogo della codifica "7" va usata "_" (ANSI 95).

Il trattino unificatore non è previsto all'interno dei campi formula.

Punto in alto

È il simbolo "·" (ANSI 183). Si tratta di un diacritico che indica la caduta di una consonante. Il punto in alto appartiene alla parola con cui è a contatto: non vi sono ambiguità interpretative se il punto in alto si trova all'inizio o alla fine di una parola, o anche se costituisce una parola da solo; può avere invece due effetti alternativi se si trova all'interno di una parola, ovvero può spezzarla in due (e in questo caso il punto in alto si accoda alla parola precedente), oppure può costituire un'unica parola con i caratteri circostanti; per indicare che si intende utilizzare questo secondo comportamento, il punto in alto va codificato facendolo precedere da un "&", digitando cioè "&·".

La situazione è descritta dalla tavola seguente:

testo	forme estratte
·	·
·l	·l
i·	i·
co·llui	co· llui
ê&·llo	ê·llo

FORME POLIREMATICHE

Si intendono con questo termine forme costituite da due o più parole. Si tratta spesso di nomi geografici o di persona. Normalmente le parole costituenti verrebbero trattate da GATTO come forme distinte, come tali memorizzate e cercabili. Se invece la sequenza di parole viene racchiusa tra i codici "@/" e "/@", viene trattata come un tutt'uno, dando origine ad un'unica forma.

Ad esempio **Bartolomeo di Bonaccorso** dà origine ad occorrenze delle 3 forme *bartolomeo*, *di* e *bonaccorso*, mentre **@/Bartolomeo di Bonaccorso/@** genera l'unica forma *bartolomeo di bonaccorso*.

Per quanto concerne le forme polirematiche valgono le regole seguenti:

- è ammesso il carattere spazio, alfabeticamente precedente ogni altro carattere. Lo spazio non è però ammesso a inizio e fine parola;
- il trattino di unione può essere indicato, indifferentemente, con i codici "-" o "7" (nei corpus impostati per accettare le cifre arabe all'interno delle forme il trattino può essere indicato solo col codice "-", per evitare ambiguità);
- il punto di abbreviazione può essere indicato, indifferentemente, con i codici "." o "8": ne consegue che il carattere *punto* è sempre interpretato come abbreviazione, mai come interpunzione; il punto debole "&." non è ammesso; nei corpus impostati per accettare le cifre arabe all'interno delle forme il punto di abbreviazione può essere indicato solo col codice ".", per evitare ambiguità;
- il punto in alto è accettato solo nella codifica "&" (altrimenti spezzerebbe la forma, il che sarebbe in conflitto col concetto di polirematica);
- se la polirematica deve essere trattata in notazione anglosassone, i codici relativi a questa notazione, ovvero **&A** e **&I**, vanno collocati esternamente ai codici di inizio e fine polirematica: un esempio di codifica corretta è **&A@/Bartolomeo di Bonaccorso/@&I**;
- l'apice (o apostrofo) è ammesso solo se la polirematica è inserita in una parte di testo in notazione anglosassone (altrimenti spezzerebbe la forma, il che sarebbe in conflitto col concetto di polirematica);
- una forma polirematica non può essere inserite all'interno di un campo formula;
- una forma polirematica non può essere ripartita su due (o più) righe di testo.

Un uso particolare dei codici di forma polirematica

I codici di forma polirematica possono essere applicati, in modo forse un po' improprio ma sintatticamente e funzionalmente corretto, per memorizzare in archivio, così come compaiono nel testo, parole evidenziate tramite l'inserimento di uno o più spazi tra i caratteri che le costituiscono. Ad esempio, volendo memorizzare la parola "R o m a", basta codificarla con @/R o m a/@. La parola verrà memorizzata come "r o m a" e come tale potrà essere ricercata, completa di spazi. Tale parola non avrà niente a che vedere con l'eventuale parola "roma", che, se presente negli archivi, verrà considerata come una forma diversa dalla precedente.

ORDINAMENTO ALFABETICO DELLE FORME

Quando si ottengono, come risultato di una ricerca, liste di forme o serie di contesti, l'ordine di presentazione dei risultati dipende dall'ordinamento alfabetico delle forme coinvolte. Da questo punto di vista non esistono differenze tra lettere maiuscole e minuscole, per cui parole come *Casa*, *casa* e *CASA* sono indicizzate con la comune grafia *casa* e quindi trattate come un'unica forma. L'ordinamento alfabetico tiene invece conto, come è ovvio, delle lettere costituenti le forme e delle eventuali aggettivazioni applicate (accenti, cediglie, altri simboli). Da questo punto di vista gli elementi costituenti una forma possono essere di tre tipi:

1. caratteri non accentati,
2. caratteri accentati,
3. *pseudocaratteri*.

Nella terminologia di GATTO col termine *pseudocaratteri* si intendono speciali caratteri diversi dalle lettere dell'alfabeto, a se stanti e distinti dagli altri, a differenza ad esempio delle normali accentazioni che costituiscono invece un unico carattere con la lettera cui si applicano (ad esempio **á** costituisce un unico carattere, di codice ANSI 225).

Gli pseudocaratteri sono 5:

- ` (ANSI 96) utilizzato per rappresentare j o y con accento grave,

- , (ANSI 184) utilizzato per rappresentare t con cediglia,
- (ANSI 183) punto in alto,
- ' apice,
- trattino.

I primi due sono utilizzati in realtà solo per rappresentare accentazioni mancanti nella tavola ANSI.

L'ordinamento alfabetico non può basarsi semplicemente sull'uguaglianza o diversità dei caratteri inclusi nelle forme. Si considerino infatti, a titolo di esempio, le tre forme **li**, **lima**, **li**: la prima lettera, "l", è uguale per tutte, la seconda è "i" (ANSI 236) per la prima e "i" (ANSI 105) per le altre due. Basando quindi l'ordinamento semplicemente sui caratteri costituenti le forme (e i relativi codici ANSI) otterremmo l'ordinamento:

- 1°. **li**
- 2°. **lima**
- 3°. **li**

che non è certo quello desiderato.

L'ordinamento seguito da GATTO per risolvere questo tipo di difficoltà opera in due tempi:

1. dapprima le parole vengono confrontate sulla base delle lettere componenti, escludendo sia gli accenti che gli pseudocaratteri,
2. se a questo punto due parole risultano uguali si reintroducono accenti e pseudocaratteri e si procede ad un ulteriore confronto tra esse.

Applicando questo metodo al caso in questione otterremmo, dopo la prima fase:

- 1°. **li = li**
- 3°. **lima**

dopo la seconda fase:

- 1°. **li**
- 2°. **li**
- 3°. **lima**

che è l'ordinamento voluto.

Per quanto riguarda la gerarchia alfabetica complessiva, fermo restando che non esiste una stessa lettera alla quale possano essere associati tutti gli accenti e gli pseudocaratteri elencati, essa è la seguente:

- 1°. nessun accento
- 2°. accento grave
- 3°. accento acuto
- 4°. accento circonflesso
- 5°. dieresi
- 6°. tilde
- 7°. ring (es.: Å)
- 8°. cediglia
- 9°. punto in alto
- 10°. apice
- 11°. trattino

Per quanto riguarda infine l'ordinamento tra le diverse lettere vale ovviamente il normale ordinamento alfabetico, con la regola aggiuntiva che il *punto* (di abbreviazione, codificato nei file testo con "8") precede alfabeticamente la "a". Nel caso di forme polirematiche lo spazio precede ogni altro carattere.

DIGITAZIONE DELLE LETTERE ACCENTATE

Non tutte le lettere accentate sono disponibili sulle tastiere dei PC; per quelle non presenti si può ricorrere alla digitazione mediante il corrispondente codice numerico ANSI. In tal caso procedere come segue.

Supponiamo di voler digitare la lettera di codice ANSI 200 (È):

1. premere il tasto <Alt> (quello a sinistra),
2. sempre tenendo premuto <Alt>, digitare sul tastierino numerico <0> seguito dalle cifre costituenti il codice, nel caso in questione <2>, poi <0>, poi <0>,
3. rilasciare <Alt>.

Per l'elenco completo dei codici ANSI, vedere l'Appendice G - *Caratteri ANSI*.

SEGNI DI INTERPUNZIONE ACCETTATI

La punteggiatura può essere costituita da:

- punto, punto e virgola, due punti (punteggiatura forte, cioè che funge da separatore tra due periodi),
- virgolette, virgola, trattino, parentesi tonda aperta e chiusa, caratteri « e », punto debole (punteggiatura debole, cioè che non funge da separatore tra due periodi),
- punto esclamativo, punto interrogativo, puntini di sospensione (punteggiatura che può essere definita come debole o forte a scelta dell'utente)

LEZIONE 3. COME PREPARARE UN FILE TESTO USANDO I SOLI CODICI BASE: TESTO, RIFERIMENTI ORGANICI, NUMERI DI PAGINA, DI VOLUME E DI COLONNA, NUMERI DI RIGA

PANORAMICA

Affinché un'opera trascritta in un file possa essere correttamente utilizzata da GATTO, questo deve includere tutte le informazioni aggiuntive al testo vero e proprio e necessarie alla procedura per strutturare il sistema di coordinate da usare all'interno di quel testo.

Le codifiche che devono necessariamente comparire all'interno del file sono due: riferimenti organici e numeri di pagina (in realtà questi ultimi non sono richiesti se il loro uso non è stato previsto all'atto della creazione del corpus - ved. capitolo 6).

riferimento organico: consiste nel titolo del capitolo, paragrafo, sonetto, cantica o altro che caratterizza la sezione di testo seguente. Si può dire che il file testo è costituito da blocchi di testo ognuno preceduto dal proprio riferimento organico. Si possono definire riferimenti organici di 5 livelli diversi: ad ogni contesto verrà associato un unico riferimento organico complessivo, costituito dalla concatenazione dei riferimenti organici non nulli definiti al momento, a partire da quello di livello 5 fino a quello di livello 1, separati tra loro dal carattere "-". Nel file testo un riferimento organico è preceduto da uno o più caratteri % (N caratteri % per un riferimento organico di livello N), inseriti come primi caratteri nella riga del file ed è costituito da tutti i successivi caratteri presenti nella stessa riga, la quale non può quindi contenere testo.

numero di pagina: è costituito dal numero della pagina racchiuso da una coppia di caratteri \$, posto all'inizio del testo presente in quella pagina. Questa codifica va inserita all'inizio del rigo. I caratteri che seguono il secondo \$ costituiscono la prima riga della pagina.

Poiché ad ogni occorrenza localizzata in un testo viene assegnato un riferimento organico e un numero di pagina, il file testo dovrà cominciare con un riferimento organico, seguito dal primo numero di pagina e quindi dal testo.

Nota

In questa Lezione e nella successiva Lezione 4, per semplicità, tutti gli esempi verranno fatti utilizzando solo riferimenti organici di livello 1, ovvero quelli introdotti da un singolo codice %. Al termine della Lezione 4 verrà descritto in dettaglio l'uso di riferimenti organici a più livelli.

Nota

In questa Lezione, così come nel resto del manuale, l'uso di corsivi, grassetto ed altro nel rappresentare il contenuto di testi ha solo funzioni di evidenziazione tipografica: i file testo preparati devono essere assolutamente privi di queste come di altre formattazioni.

Supponiamo di dover inserire nel corpus il testo seguente:

Introduzione

(inizio pagina 3) *Al nome di Dio amen. Incomincia il libro che si chiama il Dyalagho. Questa opera si è facta per mano di Frate Giovanne Campoli di Messina dell' ordine de frati minori, ad istantia e devotione di nostra signiora Madonna Alionora regina di Sicilia; sì che ognia persona che leggerà nella dicta opera prieghi Iddio per lei devotamente,*

e per lo nostro signiore il Re Fridericho suo marito, et anche per tucta loro herede.

(inizio pagina 4) che era de sapere per che fine è facto chisto libro, dico che lo fine si è a ciò, che chilli chi legino sì se sforzino de secutare, co l' aiuto de la gratia de Dio, la vita et la sanctitate de li patri antichi, e specialmente minesprezando li delecti corporali e vani de chisto mundo, per li quali se inpèdicano li delecti spirituali e devoti de l' anima.

Prologo

«Quadam die.» La intencione de chisto prolago si è, che reconta sancto Gregorio che uno jurno, stando illo multo agravato per la sollicitudine la (inizio pagina 5) quale avia per lo officio suo, chè era papa et avia generale cura de tucti, volcese dare alcuno riposo in uno loco solitario, in lo quale loco potesse ben pensare de se medemi et pensare bene quanto li desplachia la sollicitudine de lo mundo, per la quale illo era destracto de la quiete [de] contemplacione.

(inizio pagina 6) La tercza cosa de che se actristava sancto Gregorio si era, quando illo pensava supra alcune sancte persune, che da l' in tucto aveano habandonato lo mundo et erano tucti coniunti con Dio; lo quale Dio volceli conservare in chella suavitate de contemplacione, no li dando sollicitudine de chisto mundo.

Il testo include due riferimenti organici e va da pagina 3 a pagina 6.

Il file testo dovrà iniziare con il primo riferimento organico:

%Introduzione

subito dopo andrà inserito il codice indicante l'inizio della pagina 3, ovvero

\$3\$

a seguire, il testo della prima pagina (in evidenza i codici di controllo):

%Introduzione

\$3\$*Al nome di Dio amen. Incomincia il libro che si chiama il Dyalagho. Questa opera si è facta per mano di Frate Giovanne Campoli di Messina dell' ordine de frati minori, ad instantia e devotione di nostra signiora Madonna Alionora regina di Siciglia; sì che ogra persona che leggerà nella dicta opera prieghi Iddio per lei devotamente, e per lo nostro signiore il Re Fridericho suo marito, et anche per tucta loro herede.*

Il file testo completo risulterà come segue:

%Introduzione

\$3\$*Al nome di Dio amen. Incomincia il libro che si chiama il Dyalagho. Questa opera si è facta per mano di Frate Giovanne Campoli di Messina dell' ordine de frati minori, ad instantia e devotione di nostra signiora Madonna Alionora regina di Siciglia; sì che ogra persona che leggerà nella dicta opera prieghi Iddio per lei devotamente, e per lo nostro signiore il Re Fridericho suo marito, et anche per tucta loro herede*

\$4\$*che era de sapere per che fine è facto chisto libro, dico che lo fine si è a ciò, che chilli chi legino sì se sforzino de secutare, co l' aiuto de la gratia de Dio, la vita et la sanctitate de li patri antichi, e specialmente minesprezando li delecti corporali e vani de chisto mundo, per li quali se inpèdicano li delecti spirituali e devoti de l' anima.*

%Prologo

«Quadam die.» La intencione de chisto prolago si è, che reconta sancto Gregorio che uno jurno, stando illo multo agravato per la sollicitudine la
\$5\$*quale avia per lo officio suo, chè era papa et avia generale cura de tucti, volcese dare alcuno riposo in uno loco solitario, in lo quale loco potesse ben pensare de se*

medemi et pensare bene quanto li desplachia la sollicitudine de lo mundo, per la quale illo era destructo de la quiete [de] contemplacione.

\$6\$La tercza cosa de che se actristava sancto Gregorio si era, quando illo pensava supra alcune sancte persune, che da l' in tucto aveano habandonato lo mundo et erano tucti coniunti con Dio; lo quale Dio volceli conservare in chella suavitate de contemplacione, no li dando sollicitudine de chisto mundo.

Il file testo è completo e non richiede altri interventi per essere inserito in un corpus.

Obbligatorietà o facoltatività dell'inserimento dei numeri di pagina

All'atto della costruzione di un nuovo corpus dovrà essere indicato se nei testi che ne faranno parte i numeri di pagina e di riga saranno *obbligatori, disabilitati* o *facoltativi*.

Corpus con numeri di pagina obbligatori

Se si è optato per la prima soluzione, tutti i testi inclusi nel corpus devono riportare i numeri di pagina, a partire dall'inizio del documento. Un testo privo di numeri di pagina o contenente il codice \$0\$ (pagina 0) non viene accettato in fase di inserimento nel corpus.

Volendo proprio inserire in un corpus di questo tipo un documento privo di numeri di pagina (ad esempio un manoscritto), si può ricorrere al trucco di inserire all'inizio un numero di pagina fittizio, ad esempio \$1\$ (non si può utilizzare \$0\$ perché non accettato); in questo caso, poiché il numero di riga nella pagina viene incrementato automaticamente, a partire da 1 con un massimo di 999, ricordarsi di inserire nel testo altri numeri di pagina fittizi ad intervalli opportuni, ovvero prima di arrivare alla riga 1.000 (ogni volta il conteggio del numero di riga ricomincia). Volendolo, si può ribadire sempre lo stesso numero di pagina (es.: \$1\$...\$1\$...\$1\$...).

Corpus con numeri di pagina facoltativi

In questo caso i numeri di pagina possono essere inseriti in alcuni testi e omessi in altri; uno stesso testo può contenere i numeri di pagina in alcune sue parti ed esserne privo in altre.

Perché un intero testo venga assunto come privo di numeri di pagina è sufficiente non inserire al suo interno nessun codice del tipo \$nnn\$.

Per far iniziare i numeri di pagina non all'inizio del testo ma in uno specifico punto al suo interno, non inserire all'inizio del filgato nessun codice \$nnn\$ ma iniziare a introdurli, con i corretti numeri di pagina, a partire dal punto desiderato.

Se in un testo corredato da numeri di pagina si vuole interrompere tale indicazione a partire da un certo punto (per una parte del testo o per tutta la parte rimanente), inserire in tale punto il codice \$0\$, che indica pagina 0 (ovvero, per questo tipo di corpus, *assenza di pagina*). In qualunque punto successivo la numerazione può essere ripresa, eventualmente di nuovo interrotta con la stessa tecnica e così via.

Corpus con numeri di pagina disabilitati

Quanto detto circa l'obbligatorietà dell'inserimento dei numeri di pagina non si applica nel caso di corpus costruiti specificando, all'atto della loro costruzione, che non si intende farne uso (uso *disabilitato*): in questo caso i numeri di pagina possono essere tranquillamente omessi. La loro eventuale presenza in alcuni testi o in alcune parti di testi viene semplicemente ignorata da GATTO che non farà uso di tali dati.

SCRITTURA DI UN FILE TESTO CONTENENTE SOLO TESTO, RIFERIMENTI ORGANICI (DI LIVELLO 1) E NUMERI DI PAGINE

L'esempio precedente chiarisce come inserire testi contenenti solo riferimenti organici, numeri di pagine e testo. Le informazioni relative a questi punti verranno ora completate con accenni alle possibili situazioni di errore, alle regole seguite da GATTO nel generare automaticamente i numeri di riga e all'eventuale inserimento di numeri di volume o di colonna.

Avvertenze nella scrittura di riferimenti organici

Come detto in precedenza, un riferimento organico occupa una riga del file testo ed è preceduto dal codice %. Nell'inserire dei riferimenti organici in un file testo tenere presenti i punti seguenti:

- fare attenzione al fatto che il primo riferimento organico deve essere collocato prima dell'inizio del testo, in quanto tutte le parole del testo devono poter essere associate ad un riferimento organico precedente ad esse,
- il riferimento organico può essere formato al massimo da 250 caratteri, inseriti in un'unica riga,
- le parole incluse nei riferimenti organici non sono ricercabili: le ricerche operano solo sulle parole del testo,
- eventuali spazi inseriti tra % e il riferimento organico verranno considerati facenti parte di quest'ultimo,
- è accettato un riferimento organico vuoto, cioè formato dal solo carattere % con l'eventuale aggiunta di spazi,
- non spezzare su più righe il riferimento organico: le righe successive verrebbero interpretate come appartenenti al corpo del documento (se si iniziassero anche le righe di continuazione con %, queste verrebbero interpretate come altrettanti riferimenti organici diversi)
- un file testo deve contenere obbligatoriamente almeno un riferimento organico. Non volendo inserire riferimenti organici in un file testo, inserire come prima riga:
%
ovvero una riga contenente solo un riferimento organico fittizio costituito da nessun carattere o da soli spazi.
- come detto, le parole che formano un riferimento organico non sono indicizzabili; volendo poter ricercare anche le parole costituenti i riferimenti organici, duplicare questi ultimi in normali righe di testo (è consigliabile terminare la parte duplicata con un punto, per separarla dalle frasi successive).

Esempio.

Nel testo che segue, si vogliono poter cercare anche le parole *problemi* e *finanziari*:

Problemi finanziari

In questo capitolo tratteremo la risoluzione di...

Soluzione: il file testo va scritto come segue:

%Problemi finanziari

Problemi finanziari. In questo capitolo tratteremo la risoluzione di...

La prima riga contiene il riferimento organico vero e proprio: La seconda ne duplica il contenuto, permettendo in tal modo che esso venga indicizzato in quanto parte del testo; la parte duplicata termina con un punto per separarla dal periodo successivo nella visualizzazione a periodi. Il resto del testo mantiene i corretti numeri di riga.

Esempi di riferimenti organici corretti:

1. *%Risoluzione di problemi*
2. *%*
3. *% Risoluzione di problemi*

Esempi di riferimenti organici errati:

1. *Risoluzione di problemi* (errore: manca il %, la riga verrebbe considerata come testo)
2. *%Modifica del formato, del tipo, della dimensione e delle altre caratteristiche di un elemento del testo o delle figure* (errore: il secondo rigo, privo di %, verrebbe considerato parte del testo)
3. *%Modifica del formato, del tipo, della dimensione e delle altre caratteristiche di un elemento %del testo o delle figure* (errore: il secondo rigo verrebbe considerato come un secondo riferimento organico, separato dal primo; al primo riferimento organico non risulterebbe associata nessuna parte del testo)

NUMERI DI VOLUME E DI COLONNA

È stato già detto che il numero di pagina va posto all'inizio del primo rigo presente nella pagina ed è formato da un numero intero compreso tra 1 e 32.000, racchiuso tra due caratteri \$.

Un testo deve contenere obbligatoriamente almeno un numero di pagina.

Numero di volume: prima codifica ammessa

Tra il primo \$ e il numero di pagina può essere inclusa una lettera dell'alfabeto, maiuscola o minuscola, per indicare il numero di volume, secondo la codifica "a" o "A" = volume 1, "b" o "B" = volume 2,.. Nei contesti ottenuti come risultati delle ricerche verrà riportato il numero di volume prima del numero di riga. Ad esempio, inserendo nel file testo **\$c34\$**, otterremo nella testata dei contesti tratti da quella pagina di testo la dicitura "**3.34**" (col numero di volume in azzurro, seguito da un punto e dal numero di pagina in blu). Fa eccezione la lettera minuscola "r", che specifica invece che il numero di pagina è riportato nel testo come numero romano: nei contesti, per indicare questa situazione, il numero di pagina è visualizzato preceduto da "r". Ad esempio, **\$R12\$** dà luogo alla scrittura "**18.12**" (pagina 12 del volume 18) mentre **\$r12\$** dà luogo alla scrittura "**r12**" (pagina XII).

Numero di colonna: prima codifica ammessa

Tra il numero di pagina e il secondo \$ può essere inclusa una lettera dell'alfabeto, maiuscola o minuscola, per indicare il numero di colonna nella pagina, secondo la codifica "a" o "A" = colonna 1, "b" o "B" = colonna 2,.. Nei contesti verrà riportato il numero di colonna dopo il numero di pagina. Ad esempio, inserendo nel file testo la codifica **\$29B\$**, otterremo nella testata dei contesti tratti dal testo successivo la dicitura "**29.2**" (col numero di pagina in blu, seguito da un punto e dal numero di colonna in blu). Fanno eccezione le lettere minuscole "r" e "v" che indicano rispettivamente *recto e verso*. Ad esempio, una occorrenza nella riga 1 della pagina **\$30R\$** dà luogo alla scrittura "**30.18.1**" (pagina 30, colonna 18, riga 1) mentre **\$30r\$** dà luogo alla scrittura "**30r.1**" (pagina 30 recto, riga 1) e **\$42v\$** dà luogo alla scrittura "**42v.1**" (pagina 42 verso, riga 1).

Numero di volume e colonna: seconda codifica ammessa

Un secondo modo di indicare il numero di volume consiste nell'inserire, subito dopo il \$ iniziale, la sequenza **VnnnV**, dove **nnn** (numero di volume) indica un numero compreso tra 1 e 999. L'eventuale numero di colonna viene codificato come descritto precedentemente. Ad esempio, **\$V87V34c\$** indica la colonna 3 della pagina 34 del volume 87, e darà luogo alla scrittura "**87.3.34**".

Non occorre riportare il numero di volume in corrispondenza di ogni numero di pagina: dopo la prima assegnazione esso rimane infatti invariato, per le pagine successive, fino a che non viene ridefinito; per evitare che da un certo punto del testo in poi il numero di volume continui a essere riportato, associare al numero di pagina con cui comincia tale parte del testo il volume 0 (es. **\$V0V45\$**).

NUMERI DI RIGA

Quando l'interprete trova un numero di pagina, il numero di riga (nella pagina) viene automaticamente riportato a 1 e tutte le parole che si trovano nella stessa riga vengono assegnate alla riga 1 di quella pagina; ogni volta che il testo va a capo il contatore delle righe viene automaticamente incrementato di una unità, fino ad incontrare il successivo numero di pagina.

Valgono le regole seguenti:

- ogni riga del file testo corrisponde ad una riga di testo,
- le righe vuote o contenenti solo segni di interpunzione sono conteggiate come righe a tutti gli effetti,
- una riga che contiene solo la codifica del numero di pagina non viene conteggiata come riga e le parole inserite nella riga successiva del file testo sono assegnate alla riga 1 della pagina,
- le righe dei riferimenti organici (cioè quelle che cominciano con %), non sono conteggiate come righe,
- un numero di riga nella pagina non può superare 999. Questo limite non verrà ovviamente mai raggiunto inserendo regolarmente i numeri di pagina (che reinizializzano il conteggio), ma potrebbe essere raggiunto se, ad esempio, venisse inserito solo un unico numero di pagina fittizio all'inizio del documento; in un caso del genere ricordarsi quindi di inserire, di tanto in tanto, ulteriori numeri di pagina. Tenere peraltro presente che, se i numeri di pagina e di riga non interessano, è possibile escluderli del tutto all'atto della costruzione del corpus.

Esempio (le righe indicate con "-" vanno intese come righe di testo vuote o contenenti solo segni di interpunzione):

%Primo capitolo
\$1\$*In questo capitolo tratteremo*
problemi particolari
\$2\$*che si possono presentare*
in certe situazioni
\$3\$
di cui mostreremo
\$4\$
 -
esempi significativi
 -
che descriveremo in dettaglio.
%Secondo capitolo
 -
Passiamo ora a considerare

Le parole delle frasi incluse nell'esempio precedente, in base a quanto detto, riceveranno i seguenti numeri di pagina e di riga nella pagina:

testo	pagina	riga	commento
In questo capitolo tratteremo	1	1	le parole inserite a destra della codifica del numero di pagina sono assegnate alla prima riga di tale pagina
problemi particolari	1	2	il numero di riga è incrementato automaticamente a 2
che si possono presentare	2	1	la presenza di un nuovo numero di pagina reinizializza il conteggio del numero di riga
in certe situazioni	2	2	il numero di riga è incrementato automaticamente a 2
di cui mostreremo	3	1	una riga contenente solo il numero di pagina non conta come riga, per cui la riga 1 di tale pagina è quella successiva
esempi significativi	4	2	come sopra, e le righe vuote contano come righe
che descriveremo in dettaglio.	4	4	le righe vuote contano come righe
Passiamo ora a considerare	4	6	le righe contenenti riferimenti organici non sono conteggiate; le righe vuote contano come righe

Esempi di codifiche errate:

1. *quella sera \$2\$ era tardi* (errore: la codifica deve stare all'inizio della riga)
2. **\$2\$ \$3\$** *non siamo stati* (errore: la seconda codifica è in posizione errata)
3. **\$a\$** (errore: tra i simboli \$ deve essere incluso un numero)
4. **\$0\$** (errore: il numero di pagina non può essere 0)
5. **\$3\$ %Capitolo primo** (errore: il numero di pagina non può essere posto in corrispondenza di un riferimento organico)
6. Il testo inizi con:
%Capitolo primo.
In questo capitolo
\$1\$ *tratteremo vari argomenti* (errore: le prime parole del documento, ovvero "In questo capitolo", risulterebbero poste a pagina 0).

Codice &/Lnn& per forzare il numero di riga ("nn" è il numero della prossima riga)

Come detto in precedenza i numeri di riga vengono gestiti in maniera automatica. Vi sono però casi in cui si può voler forzare il numero di riga nella pagina ad un diverso valore; questo può capitare, ad esempio, quando nel file testo vengono volutamente omesse delle righe di testo oppure quando il riferimento organico nell'opera originale cade all'interno di una riga di testo (es. riga 12): in questo caso, poiché nel file testo il riferimento organico va collocato in una riga a se stante, le parole che lo seguono immediatamente verrebbero scritte materialmente nella riga successiva del file e attribuite automaticamente (ed erroneamente)

al rigo seguente (13 nell'esempio); collocando all'inizio della riga successiva al riferimento organico la sequenza **&/L12&** si indica che le parole che seguono appartengono ancora alla riga 12. Al termine di tale riga la numerazione automatica riprende regolarmente (nell'esempio, a partire dalla 13).

Altro esempio: per indicare che la riga all'inizio della quale è collocata la codifica è la 25, la sequenza da inserire **&/L25&**; le righe successive verranno automaticamente numerate 26, 27, ...

Questo codice si può usare anche in presenza di molte righe bianche consecutive presenti nel testo originale: anziché inserire fisicamente altrettante righe vuote nel file testo si può inserire semplicemente questo codice all'inizio del primo rigo successivo contenente testo per aggiornare manualmente il conteggio delle righe, tenendo così conto delle righe saltate.

Per compatibilità con le versioni precedenti di GATTO viene accettata, con la stessa funzione, la codifica **&/Pnn&** nella quale però "nn" è il numero della prossima riga meno uno. Ad esempio: per indicare che la riga all'inizio della quale è collocata la codifica è la 25, la sequenza da inserire **&/P24&**; le righe successive verranno automaticamente numerate 26, 27, ...

Esempi di applicazione di questo codice si trovano, tra l'altro, in *filgat.kaa.txt* e *filgat.p48.txt* (file testo del corpus **Demo**).

Il massimo numero di riga specificabile è 999.

LEZIONE 4. COME PREPARARE UN FILE TESTO: PARTI IN VERSI E PARTI IN PROSA, NUMERI DI VERSO, CONTEGGIO DEI VERSI PER DISTICI, TERZINE, ECC..., CODIFICA DELLA RIMA AL MEZZO, RIFERIMENTI ORGANICI A PIÙ LIVELLI

PARTI IN VERSI E PARTI IN PROSA

Non vi sono motivi di fondo per trattare diversamente le parti di testo in prosa e quelle in versi; l'opportunità di fare una distinzione deriva essenzialmente da scopi di natura pratica:

1. nei contesti ottenuti con una ricerca evidenziare il numero di verso,
2. consentire di fare ricerche di forme limitatamente al rimario del testo,
3. evidenziare la separazione tra versi successivi nei contesti visualizzati.

In particolare, nella finestra **RISULTATI RICERCA** il fine riga delle parti in versi (fine verso) viene visualizzato con " / ".

NUMERI DI VERSO

GATTO assume in partenza che un testo sia in prosa. Per indicare che una parte del testo è in versi, bisogna inserire l'informazione all'inizio della prima riga in versi mediante un codice apposito che, nella sua forma più semplice, è **&V**. Se il testo dopo un certo numero di versi torna ad essere in prosa, la cosa va segnalata ponendo una codifica **&P** all'inizio del primo rigo in prosa. Un ulteriore **&V** posto all'inizio di uno dei righe successivi indicherà l'inizio di una nuova parte in versi a partire da tale rigo e così via. All'inizio di un file testo, se il testo stesso è in prosa, non occorre inserire nessun codice specifico in quanto come si è detto *per default* un testo è assunto essere in prosa.

Alla prima comparsa di un codice **&V** all'interno di un riferimento organico il numero di verso è inizializzato a 1 e viene incrementato di 1 ad ogni riga successiva fino ad incontrare, eventualmente, il codice **&P**; a quel punto il conteggio dei versi viene congelato e sono possibili due prosecuzioni:

1. all'interno dello stesso riferimento organico è presente una seconda parte in versi, dopo quella intermedia in prosa: alla comparsa del nuovo codice **&V** il conteggio del numero di verso riparte da dove si era interrotto; se la parte in versi precedente terminava ad esempio col verso 22 (compreso), la nuova partirà dal verso 23;
2. si incontra un nuovo riferimento organico: il numero di verso ripartirà da 1.

Tenere presente che una riga vuota, così come incrementa il numero di riga, incrementa anche il numero di verso.

L'esempio che segue riassume le cose dette:

%Prologo. Explicit

\$13\$*Qui comincia il libro primo di Anizio Mallio Torquato*

Severino Boezio - ex7consolo ordinario, patrizio -
de la Filosofica Consolazione, ridotto in volgare;
%Componim. 01

&V Io, che compuosi già versi e cantai

Con studio fiorito, son costretto

\$14\$Di scriver canti di tristizia e guai.

Ecco le laceranti nel mio petto

&PS15\$Quando meco queste cose tacito ripensava e la lagrimosa
lamentanza con lo stile scrivea, sopra 'l mio capo essere
gittato, con questi versi della turbazione di nostra mente
fece lamento:

&VAimè! come la mente attuffata

Nel basso strabocchevole profondo.

-

Sta impigrita, di virtù privata;

%Componim. 02

E lasciata la luce del giocondo

Proprio lume, che le dà natura,

Rovina fuor in tenebre del mondo.

&P «Ma di medicina è tempo, non di lamentanza». E allora...

Alle righe di questo file testo vengono assegnati i seguenti numeri di pagina, riga, verso:

inizio riga	pagina	riga	verso	commento
Prologo...	-	-	-	il riferimento organico non è riga di testo
Qui comincia...	13	1	-	prima riga di parte in prosa
Severino...	13	2	-	riga: +1; (notare "ex7consolo" per avere "ex-consolo")
de la Filosofica...	13	3	-	parte in prosa; riga incrementata di 1
Componim. 01	-	-	-	nuovo rif.org.; il riferimento organico non è riga di testo
Io, che...	13	4	1	parte in versi; riga +1; verso inizializzato a 1
Con studio...	13	5	2	parte in versi; incrementati di 1 numeri di riga e di verso
Di scriver...	14	1	3	nuova pagina; la riga torna a 1; verso incrementato di 1
Ecco le...	14	2	4	parte in versi; incrementati di 1 numeri di riga e di verso
Quando meco...	15	1	-	nuova pagina; nuova parte in prosa: la riga torna a 1
lamentanza...	15	2	-	parte in prosa; riga incrementata di 1
gittato, con...	15	3	-	parte in prosa; riga incrementata di 1
fece lamento...	15	4	-	parte in prosa; riga incrementata di 1
Aimè...	15	5	5	nuova parte in versi; incrementati di 1 riga e verso
Nel basso...	15	6	6	parte in versi; incrementati di 1 numeri di riga e di verso
- (riga vuota)	15	7	7	una riga vuota incrementa numero di riga e di verso
Sta impigrita...	15	8	8	parte in versi; incrementati di 1 numeri di riga e di verso
Componim. 02	-	-	-	nuovo rif.org.; il riferimento organico non è riga di testo
E lasciata...	15	9	1	riga incrementata di 1; verso reinizializzato a 1
Proprio lume...	15	10	2	parte in versi; incrementati di 1 numeri di riga e di verso
Rovina fuor...	15	11	3	parte in versi; incrementati di 1 numeri di riga e di verso
«Ma di medicina...	15	12	--	nuova parte in prosa; riga incrementata di 1

CONTEGGIO DEI VERSI PER DISTICI, TERZINE, ECC..

Per semplicità finora è stato fatto riferimento al codice \$V per indicare la presenza di una parte di testo in versi.

In realtà questo codice indica più specificamente una parte di testo *in versi numerati singolarmente*. È possibile utilizzare il codice &V con una sintassi più generale per effettuare il conteggio anche per distici, terzine, quartine, ottave o stanze di lunghezza qualunque; ad esempio, nel caso di numerazione per distici,

all'occorrenza localizzata verrà associato il numero di distico, il quale verrà incrementato di 1 ogni due righe in versi.

Tutte le altre considerazioni fatte nel paragrafo precedente, circa i meccanismi di interruzione del conteggio e di reinizializzazione ad 1 ad ogni nuovo riferimento organico si applicano anche a questo caso più generale.

La tabella che segue riporta il quadro completo dei codici &V e il loro significato.

Codice	Modalità di conteggio dei versi
&V	numerazione per versi singoli
&V@2@	numerazione per distici
&V@3@	numerazione per terzine
&V@4@	numerazione per quartine
&V@8@	numerazione per ottave
&V@nnS@	numerazione per stanze di nn versi (nn compreso tra 1 e 99)

Note

- La "S" (stanza) se usata in combinazione con 2, 3, 4 o 8 (nn = 2, 3, 4, 8) sta ad indicare che il raggruppamento in versi, pur coincidendo con tali lunghezze, non si configura come distico, terzina, quartina o ottava (es. &V@2S@);
- se nn è un valore diverso da 1, 2, 3, 4 o 8, la "S" può essere omessa, in quanto non vi è comunque ambiguità con i casi precedenti (es. &V@12@ = &V@12S@);
- per compatibilità di sintassi, la numerazione per versi singoli può essere indicata, oltre che con &V, anche con **&V@1@**, **&V@S@**, **&V@1S@**, con identico significato.
- all'interno del testo afferente ad uno stesso riferimento organico non possono coesistere raggruppamenti diversi, come ad esempio distici e terzine. Ad esempio se in un punto del testo vi sono versi definiti col codice &V@2@ e questi versi sono seguiti da un parte in prosa, ulteriori versi nel testo relativo allo stesso riferimento organico potranno essere definiti da un nuovo &V@2@ ma non, ad esempio, da &V o da &V@4@ (in un caso di questo genere, se non è possibile un'altra soluzione, ripartire il testo in due riferimenti organici distinti).

Rappresentazione della numerazione in versi nelle testate dei contesti localizzati nelle ricerche

Tra le altre informazioni riportate nelle testate dei contesti ottenuti come risultato di una ricerca vi è anche il numero di verso, distico, ecc., nel quale l'occorrenza localizzata è inserita, in colore verde. Tale dato ha la forma: <lettera indicante il tipo di numerazione> + "." + <numero>, sulla base della tabella che segue:

dicitura	significato
v.	verso
d.	distico
t.	terzina
q.	quartina
o.	ottava
s.	stanza

Es.: "**v. 34**" = verso numero 34; "**q. 5**" = quartina numero 5.

Modifica del numero di verso, distico, ecc.. generato automaticamente

Come detto in precedenza i numeri di verso, distico, terzina o altro raggruppamento di versi vengono gestiti in maniera automatica. Vi sono però casi in cui si può voler forzare questi dati a valori diversi da quelli generati automaticamente. Inserire in questo caso il codice **&Vnn&** all'inizio del primo rigo del raggruppamento di cui si vuole ridefinire il numero: a tale raggruppamento di versi verrà attribuito il numero nn. Ad esempio: avendo in precedenza definito la parte di testo corrente come composta da terzine (codice &V@3@) e volendo ad un certo punto cambiare la numerazione delle terzine, per indicare che il verso corrente è il primo della terzina 151 inserire all'inizio del verso il codice **&V151&**; le terzine successive verranno automaticamente numerate 152, 153, ...

Per compatibilità con le versioni precedenti di GATTO viene accettata, con la stessa funzione, la codifica **&Rnn&** nella quale però "nn" è il numero del prossimo raggruppamento di versi meno uno. Ad

esempio: per indicare che il verso corrente è il primo della terzina 151 inserire all'inizio del verso il codice **&/R150&**.

Il massimo numero di raggruppamento specificabile è 32.000.

CODIFICA DELLA RIMA AL MEZZO

Volendo indicare che la parola immediatamente precedente costituisce rima al mezzo, inserire subito dopo il codice **&T** (separato o meno da spazi).

Vedere ad esempio nel testo filgat.btb.txt (pagina 11):

*No, ma lo core meo
more più spesso e forte
che no faria di morte&T - naturale,*

La parola che costituisce rima al mezzo è *morte*; il successivo trattino compare nell'edizione ma non è richiesto da GATTO.

RIFERIMENTI ORGANICI A PIÙ LIVELLI

Finora, per semplicità, in tutti gli esempi sono stati utilizzati solo riferimenti organici *di livello 1*, ovvero quelli introdotti da un singolo codice **%**. In realtà, come anticipato in precedenza, GATTO consente di fare uso di 5 diversi livelli di riferimenti organici, così da consentire una strutturazione più articolata del testo.

Quando si parla di *riferimento organico associato a un'occorrenza o a un contesto* si intende una stringa formata dalla concatenazione dei riferimenti organici definiti in quel punto, separati dal carattere "-"; più precisamente la stringa è costituita dal rif. org. di livello 5 (se esiste), seguito da un trattino, più il rif. org. di livello 4 (se esiste), seguito da un trattino, e così via fino al rif. org. di livello 1.

Per inserire nel testo un riferimento organico di livello n (con n compreso tra 1 e 5) è sufficiente inserire una riga che inizia con n caratteri **%**, seguiti dal testo del riferimento organico di quel livello.

Vediamo un esempio

*%%%Parte prima
%%Capitolo terzo
%Prima strofa
Questo è il primo verso...*

A ognuna delle parole che compongono il testo "Questo è il primo verso" verrà associato il riferimento organico:

Parte prima-Capitolo terzo-Prima strofa

Le regole da seguire per inserire i riferimenti organici nei testi sono le seguenti:

- ogni parola di testo deve essere associata ad un riferimento organico. Questo significa che prima della prima parola di testo deve essere inserito almeno uno dei cinque livelli di riferimento organico;
- ogni riga di riferimento organico può contenere solo un riferimento organico di un unico livello;
- l'inserimento di un riferimento organico di livello n in un punto del testo:
 1. lascia inalterati gli eventuali riferimenti organici di livello superiore preesistenti,
 2. sostituisce l'eventuale riferimento organico di uguale valore preesistente,
 3. azzerà gli eventuali riferimenti organici di livello inferiore preesistenti.

Vediamo un esempio per meglio chiarire l'effetto di queste regole. Consideriamo il testo seguente (le righe costituite da un trattino sono state inserite solo al fine di migliorare la leggibilità dell'esempio):

*%%%Parte prima
%%Capitolo terzo
%Prima strofa
Questo è il primo verso.
-
%Seconda strofa*

Questo è il secondo verso.

-

%%Capitolo quarto

Questo è il terzo verso.

-

%Quarta strofa

Questo è il quarto verso.

-

%%Capitolo quinto

%Quinta strofa

Questo è il quinto verso.

-

%%%Parte seconda

%Sesta strofa

Questo è il sesto verso.

-

%%Capitolo sesto

%Settima strofa

Questo è il settimo verso.

-

%%Capitolo sesto

Questo è l'ultimo verso.

Alle parole *primo, secondo, terzo,...., settimo, ultimo* risultano associati i riferimenti organici riportati in tabella:

parola	riferimento organico associato
primo	Parte prima-Capitolo terzo-Prima strofa
secondo	Parte prima-Capitolo terzo-Seconda strofa
terzo	Parte prima-Capitolo quarto
quarto	Parte prima-Capitolo quarto-Quarta strofa
quinto	Parte prima-Capitolo quinto-Quinta strofa
sesto	Parte seconda-Sesta strofa
settimo	Parte seconda-Capitolo sesto-Settima strofa
ottavo	Parte seconda-Capitolo sesto

Fare attenzione che per inserire consecutivamente più righe di riferimenti organici (di livelli diversi) in modo corretto bisogna sempre partire da quello di livello più alto e via via scendere. L'esempio precedente iniziava con

%%%Parte prima

%%Capitolo terzo

%Prima strofa

Questo è il primo verso.

ottenendo per la parola *primo* il riferimento organico **Parte prima-Capitolo terzo-Prima strofa**; se si fosse ad esempio invertito l'ordine delle righe dei riff. orgg, un questo modo:

%Prima strofa

%%Capitolo terzo

%%%Parte prima

Questo è il primo verso.

si sarebbe ottenuto per la parola *primo* il riferimento organico **Parte prima** (basta ricordare che un riferimento organico di livello 3 azzera i preesistenti riff. orgg. di livello inferiore).

LEZIONE 5. COME PREPARARE UN FILE TESTO: CAMPI FORMULA, SCIOGLIMENTI, ESPUNZIONI, CORSIVI, CITAZIONI, INTERPUNZIONI, PARLANTE. LEMMI E IPERLEMMI. INIZIO DEL FILE TESTO

Questa Lezione descrive l'uso di altri codici previsti da GATTO, ampliando le informazioni relative alla trascrizione su file dei testi a stampa.

CAMPI FORMULA

Un *campo formula* è destinato a contenere parti di testo che devono essere ignorate dal programma, ovvero che non saranno oggetto delle procedure di lemmatizzazione ed interrogazione. Tali sono, ad esempio, note dell'editore, brani citati da altre opere, formule matematiche (da cui il nome della codifica) e via dicendo. Per realizzare un *campo formula* è sufficiente racchiudere il testo che ne fa parte tra due codici speciali. A differenza del resto del testo queste aree possono contenere pressoché qualunque carattere. Non ci sono limiti alla lunghezza di un campo formula.

Un riferimento organico (definito dai codici %) non può essere inserito in un campo formula. Altri codici, come il numero di pagina, vengono invece accettati al suo interno e gestiti correttamente.

Apertura e chiusura di un campo formula

Un campo formula è aperto dal carattere "#" (ANSI 35) ed è chiuso dal carattere "@" (ANSI 64). In alternativa³ un campo formula può essere aperto con "&L-" e chiuso con "&L+": l'unica differenza tra i due modi per codificare i campi formula riguarda il trattamento dei puntini di sospensione collocati al loro interno.

Esempio di applicazione di un campo formula

Si consideri il brano

...nel secolo XIII si verificò una serie di eventi...

se venisse inserito nel file testo così com'è, verrebbe erroneamente indicizzata tra le altre anche la forma *xiii*, generata dalla presenza del numero romano. Per evitare che questo accada è sufficiente inserire il numero romano in un campo formula, digitando il brano come segue:

...nel secolo #XIII@ si verificò una serie di eventi...

Caratteri accettati all'interno dei campi formula

In un campo formula è accettato qualunque carattere o sequenza di caratteri, ad eccezione dei caratteri di controllo dei campi formula e di 4 stringhe *vietate*, in quanto il loro uso genererebbe erroneamente nei contesti la visualizzazione di "/" , sequenza usata invece per rappresentare un fine riga nei brani in versi.

Le 4 sequenze *vietate* sono:

1. spazio + "/" + spazio
2. spazio + "/" + "@"
3. "#" + "/" + spazio
4. "#" + "/" + "@"

Le parole che compaiono all'interno di un campo formula vengono regolarmente riconosciute e conteggiate; contribuiscono, ad esempio, a dimensionare i contesti definiti in numero di parole. Tuttavia queste parole non sono indicizzabili, non possono quindi essere lemmatizzate né cercate.

Se una parola inclusa in un campo formula contiene un punto di abbreviazione, codificato nel testo normale con "8" (o "©" nei corpus impostati per accettare le cifre nelle forme), tale carattere andrà codificato in questo caso con "&.", ovvero con un punto preceduto da "&": in caso contrario verrebbe erroneamente interpretato come la cifra 8, ammessa e riconosciuta come tale all'interno di un campo formula. Più in generale, ogni volta che un carattere *punto* va inserito all'interno di un campo formula senza però che gli si voglia attribuire il significato di punteggiatura forte, esso va codificato con "&.".

CAMPI DIVERSI

Parentesi tonda interna (nelle codifiche OVI indica scioglimento interno)

³ Per compatibilità con codifiche usate in precedenza all'Opera del Vocabolario Italiano.

Racchiude uno o più caratteri di una parola di testo; per non spezzare la forma deve essere codificata con "&(" e "&)".

Esempi:

q(uesto&) (errato)
q(uest)o (errato)
q&(ue&)sto (corretto)

Parentesi angolare esterna (nelle codifiche OVI indica espunzione esterna)

Racchiude una o più parole di testo; è codificata con "<" e ">".

Esempi:

qu<esto mio> (errato)
qu<esto> mio (errato)
<questo mio> (corretto)

Parentesi angolare interna (nelle codifiche OVI indica espunzione interna)

Racchiude uno o più caratteri di una parola di testo; per non spezzare la forma deve essere codificata con "&<" e "&>".

Esempi:

q<uesto&> (errato)
q<uest>o (errato)
q&<ue&>sto (corretto)

Parentesi quadrata esterna (nelle codifiche OVI indica integrazione esterna)

Racchiude una o più parole di testo; è codificata con "[" e "]".

Esempi:

sag[g]ia donna (errato)
[saggia don]na (errato)
[saggia donna] (corretto)

Parentesi quadrata interna (nelle codifiche OVI indica integrazione interna)

Racchiude uno o più caratteri di una parola di testo; per non spezzare la forma deve essere codificata con "&[" e "&]".

Esempi:

sag&[g]ia (errato)
sa[g]gia&] (errato)
sa&[g&]gia (corretto)

Corsivo esterno

Si estende a una o più parole di testo; è compreso tra le codifiche "&C" e "&c".

Corsivo interno

Si estende a uno o più caratteri di una parola di testo; è compreso tra le codifiche "&K" e "&k".

Citazione

È racchiusa tra le codifiche "}" (ANSI 125 e 124) e "{" (ANSI 124 e 123).

INTERPUNZIONI

Apice come segno di interpunzione

Quando l'apice ' (ANSI 39) viene usato come segno di interpunzione e non come elemento appartenente ad una parola, va codificato nel file testo con "&" (e commerciale più virgolette).

Ad esempio '**viso**', nel senso di parola tra apici va digitata **&"viso&"**; il testo visualizzato in seguito sarà '**viso**'; la corrispondente forma indicizzata sarà **viso**.

Punto non costituente segno di punteggiatura forte (punto debole)

Normalmente il punto costituisce un segno di punteggiatura forte e va digitato utilizzando proprio il carattere punto.

Sono stati già trattati i casi in cui il punto è di abbreviazione o sostitutivo di un carattere illeggibile.

Vi sono altri casi in cui il punto, pur non appartenendo a una parola e potendosi con ciò considerare segno di interpunzione, non ha però il significato di punteggiatura forte; può essere ad esempio un elemento di numerazione. In questi casi il punto va codificato con "&.", ovvero con un punto preceduto da "&". Come è stato già detto in precedenza, la stessa codifica si applica ai punti di abbreviazione e sostitutivi di caratteri illeggibili se collocati all'interno di campi formula. Nella visualizzazione all'interno dei contesti verrà in ogni caso mostrato un semplice punto. Un altro caso in cui si deve ricorrere al punto debole è quando un carattere *punto* è collocato prima della prima parola del testo appartenente ad un riferimento organico: in quella posizione non è accettato un punto forte, per cui l'unico modo per lasciarvelo è di anteporgli il codice &, trasformandolo in punto debole.

Questa codifica non è ammessa all'interno delle forme polirematiche, trattandosi comunque di un segno di punteggiatura.

Puntini di sospensione

Sono rappresentati da una successione di 2 o più punti consecutivi (non preceduti da "&"). Possono costituire punteggiatura forte o debole, a seconda dell'impostazione scelta al momento della costruzione del corpus e del loro inserimento o meno all'interno di campi formula o parentesi quadre. Valgono le regole seguenti:

- **Puntini nel testo.** Sono considerati punteggiatura forte o debole a seconda della scelta operata all'atto della costruzione del corpus.
- **Puntini in campo formula #...@ o in campo parlante \...\.** Sono sempre e comunque punteggiatura debole.
- **Puntini in campo formula &L-...&L+.** Sono considerati punteggiatura forte o debole a seconda della scelta operata all'atto della costruzione del corpus. Se però vengono racchiusi tra parentesi quadre e sono quindi codificati [...] costituiscono comunque punteggiatura debole.

Punto all'inizio del testo di un riferimento organico

Il testo immediatamente successivo ad un riferimento organico non può contenere un segno di punteggiatura forte prima della prima parola. Quindi in tale posizione non è ammesso neanche il punto. Avendo la necessità di collocarvi un punto (ad esempio per una sorta di numerazione) lo si può fare utilizzando la codifica di *punto debole* &.

Esempio di testo errato:

```
% elenco delle proprietà  
#.1@ casa di proprietà di messer Lapo;
```

Esempio di testo corretto:

```
% elenco delle proprietà  
#&.1@ casa di proprietà di messer Lapo;
```

PARLANTE

Il campo *parlante* indica colui del quale viene riportato il discorso diretto; come tale può non appartenere propriamente al testo. Il parlante è racchiuso tra una coppia di codici "\" (es.: "...Luca...").

Se questo campo è posto all'esterno dei campi formula il suo contenuto verrà indicizzato, mentre non lo sarà nel caso contrario. Se il campo parlante viene inserito all'interno di un campo formula, va codificato con #...\@ (oppure &L-...\&L+), ovvero con i codici "\" a contatto con i codici di inizio e fine campo formula; occorrenze del carattere "\" all'interno di campi formula ma con collocazioni diverse vengono interpretate come presenze di un carattere di interpunzione. All'interno di questo campo non può trovarsi un campo formula, neanche in parte; esempi di sequenze errate:

```
...\#...\@...\...  
...\&L-...\&L+...\...  
...\#...\@...\...
```

CODICI DI LEMMA E IPERLEMMA

La descrizione di questi elementi a rigore non andrebbe collocata in questa Lezione in quanto si tratta di codici inseriti automaticamente da GATTO all'interno dei file testo al momento dell'estrazione da un corpus e non di codifiche disponibili per esservi inserite da parte dell'utente. Tuttavia se quest'ultimo consulta ad esempio i file testo inclusi nel corpus **Demo** incontrerà frequentemente questi codici, ed è quindi opportuno spiegarne il significato.

Questi codici possono comparire con un doppio formato:

1. =nnn=
2. =nnn,mmm=

con **nnn** e **mmm** numeri interi compresi tra 1 e 32.000 racchiusi tra simboli di uguaglianza. In particolare **nnn** è il codice che designa il lemma associato all'occorrenza immediatamente successiva nel testo; **mmm** dove presente, è il codice dell'iperlemma associato alla stessa occorrenza.

Ad esempio, nel testo

"...parla de cose multo =47,112=suctile..."

all'occorrenza della parola *suctile* sono stati associati il lemma di codice 47 e l'iperlemma di codice 112.

INIZIO DEL FILE TESTO

Al momento di inserire i codici iniziali in un file testo può capitare di non avere ben chiaro l'ordine in cui questi vanno introdotti. Non esiste un'unica sequenza da rispettare rigidamente ma in alcuni casi l'ordine adottato può dar luogo ad errori.

Sia gli esempi precedenti che i file filgat contenuti nel corpus **Demo** possono contribuire a rimuovere eventuali dubbi a questo riguardo. In ogni caso la situazione dovrebbe essere chiarita con gli esempi che seguono.

Esempio 1

La parte di documento che si intende riportare nel file testo comincia col capitolo "Introduzione", iniziante a pagina 17 del libro, a partire dalla riga 13 di tale pagina. Il file testo comincerà come segue:

```
%Introduzione
$17$&/L13&inizio del testo...
```

Si inserisce cioè prima di tutto il riferimento organico; segue poi (comunque sempre prima dell'inizio del testo) il numero di pagina ed eventualmente subito dopo la riassegnazione del numero di riga (che altrimenti partirebbe da 1).

Esempio 2

Il testo in oggetto potrebbe essere un commento ad alcune quartine di un'opera in versi e si intenda riportare nel file testo solo le più significative tra queste ultime, rispettando però i riferimenti topografici dell'edizione. Supponiamo che la parte di interesse inizi dalla terza quartina commentata nell'"Introduzione"; questa quartina inizia alla riga 8 di pagina 19. Avremo:

```
%Introduzione
&V@4@$19$&/L8&&/V3&inizio della quartina...
Nell'ordine, sono state introdotte le informazioni seguenti:
```

1. riferimento organico (%Introduzione)
2. indicazione che il testo è in versi, più precisamente in quartine (&V@4@)
3. il testo comincia a pagina 19 (\$19\$)
4. il testo comincia alla riga 8 della pagina (&/L8&)
5. il testo comincia con la terza quartina (&/V3&)

Altre soluzioni sarebbero ugualmente lecite, come ad esempio la seguente, in cui numero di pagina e indicazione della ripartizione in quartine sono scambiati:

```
%Introduzione
```

\$19\$&V@4@&/L8&&/V3&inizio della quartina..

CAP. 4 - USO DI NOTE E DI EDIZIONI DIVERSE O TRADUZIONI

Il testo può contenere delle note da inserire anche nella versione elettronica, al fine di visualizzarle assieme alle parole o alle parti di testo cui si riferiscono. Può trattarsi di note collocate in calce alle pagine, al termine delle sezioni del testo o in fondo al documento. In tutti i casi, per poterne fare uso all'interno di GATTO, è necessario definire con precisione quale è la parte di testo cui ogni nota si riferisce.

Va anche detto che il contenuto delle note non è indicizzabile, ma può solo essere mostrato insieme al testo cui si riferiscono. Qualora si volesse indicizzarne il contenuto, una soluzione potrebbe consistere nel creare un paragrafo fittizio, posto magari al termine dell'opera, intitolato "Note" o qualcosa del genere, nel quale includere i testi delle note; queste andranno comunque trattate anche col normale meccanismo previsto in GATTO che verrà illustrato in questo capitolo, al fine di farle apparire unitamente al testo cui fanno riferimento.

Quando si parla di *mostrare una nota insieme al testo cui si riferisce* si intende in particolare di dare all'utente la possibilità di prendere visione di una nota ogni volta che, come risultato di una ricerca, sia stata localizzata una occorrenza inclusa in una parte di testo collegata alla nota: da qui la necessità di definire con precisione le parti di testo associate alle note, in modo da far sapere esattamente al programma in quali casi mostrarle.

Per associare ad un testo delle note sono necessarie 3 operazioni:

1. inserimento all'inizio del file testo di uno speciale codice che avverte l'interprete che quel file testo è destinato a contenere delle note;
2. inserimento all'interno del file testo di speciali marcatori per indicare al programma l'inizio e la fine di ogni parte di testo interessata da una nota;
3. scrittura di un altro file, ancora di tipo testuale (ANSI puro) e di nome predefinito, contenente i corpi delle note.

Il nome del file destinato a contenere le note ha, come il file testo, una parte fissa e una variabile, dipendente dalla sigla assegnata al testo. Come il nome del file testo ha la struttura "**filgat.<sigla>.txt**", dove <sigla> indica una stringa formata da 1, 2 o 3 caratteri alfanumerici, così il file destinato a contenere le note dovrà avere nome "**note.<sigla>.txt**", dove <sigla> è ancora la sigla del testo.

Un testo può contenere un massimo di 32.000 note. Ogni nota è identificata da un numero intero compreso tra 1 e 32.000; non è ammessa la presenza di due note indicate con lo stesso numero; i numeri assegnati alle note non devono essere per forza consecutivi e nemmeno crescenti: ad esempio un testo potrebbe contenere 3 note, la prima delle quali indicata dal numero 152, la seconda dal 4 e la terza dal 41.

Per quanto riguarda l'ordine dei codici posti all'inizio del file testo, vale quanto illustrato al termine della Lezione precedente, salvo il fatto che il codice che denota presenza di testo associato ovvero **@@NT**, **@@TR** o **@@A2** (per i dettagli vedere le Lezioni seguenti) va posto da solo nel rigo iniziale, prima di ogni altro codice. Ad esempio, riprendendo l'ultimo caso trattato nella Lezione precedente, se il testo contenesse delle note, si dovrebbe scrivere:

```
@@NT
%Introduzione
$19$&V@4@&/P7&&/R2&inizio della quartina..
```

LEZIONE 6. COME UTILIZZARE DELLE NOTE IN UN FILE TESTO.

Supponiamo di dover preparare il file testo corrispondente al seguente testo in versi:

(il testo manca di un titolo)

(inizio pagina 9) ... *va cercando.*
Al nome di Dio cantiam dell'amore
a laude ed a gloria^(A) del nostro salvatore
e della beata ave^(B) vergine fiore
cantiam dell'amore di Giesù gloriando.

Or che fai cor meo, che non prehendi gioia
del più vago amore ke sia senza noia
e sta nella mente sì^(C) pensoso e noia [sic]
d'amore non sente ke non va giubilando.
Fra tucta la gente lo mio cor si muta
di Christo^(D) amor dolcie in cui ò posta la speme
fioriscono le pratora per ogni rivera
e vegnaci il mio amore infra lli fior dilectando.
A te lo dico, bella: farai mech'un^(E) pacto,
ked io pur vuglo Ihesu Christo intrazacto,
e mettermene voglio ad ogni baracto:
lo cielo colla terra ne vo gire invitando.

note

^(A) ed a gloria (u.f.)] et da gloria (ed.)

^(B) aue (u.f.)] ave (ed.)

^(C) sì (u.f.)] si (ed.)

^(D) Christo (u.f.)] Xpo (ed.)

^(E) farai mech'un (u.f.)] faraime chun (ed.)

(le parti sottolineate sono quelle cui si riferiscono le note)

([sic] costituisce una annotazione dell'editore a conferma della lezione trascritta dal manoscritto)

Nel testo le note sono indicate con delle lettere, mentre GATTO accetta solo numeri; potremo quindi sostituire le lettere A,...,E con i numeri dall'1 al 5.

Immediatamente prima di una zona di testo interessata da una nota andrà inserito un codice costituito da un asterisco seguito dal numero della nota. Ad esempio, la nota numero 1 dovrà essere preceduta dal codice "*1". Al termine della parte di testo interessata dalla nota andrà inserito il codice di fine nota "*/". Come risultato, la zona di testo interessata dalla nota n. 1 risulterà "*1ed a gloria*/", per la n. 2 avremo "*2ave*/", ecc..

Affinché l'interprete accetti questi codici la prima riga del file testo dovrà essere costituita solo dal codice "@@NT".

Per completare la codifica del testo dell'esempio bisogna tenere presenti i seguenti elementi:

- manca un titolo da attribuire al riferimento organico,
- la pagina di testo è la n. 9,
- il testo è in versi,
- i puntini di sospensione iniziali costituiscono una integrazione,
- l'annotazione dell'editore non costituisce testo indicizzabile.

Il risultato finale sarà:

@@NT

%

\$9\$&V[...] va cercando.

Al nome di Dio cantiam dell'amore
a laude *1ed a gloria*/ del nostro salvatore
e della beata *2ave*/ vergine fiore
cantiam dell'amore di Giesù gloriando.

Or che fai cor meo, che non prehendi gioia
del più vago amore ke sia senza noia
e sta nella mente *3sì*/ pensoso e noia #[sic]@
d'amore non sente ke non va giubilando.

Fra tucta la gente lo mio cor si muta
di *4Christo*/ amor dolcie in cui ò posta la speme
fioriscono le pratora per ogni rivera
e vegnaci il mio amore infra lli fior dilectando.

A te lo dico, bella: *5farai mech'un*/ pacto,

*ked io pur voglio Ihesu Christo intrazacto,
e mettermene voglio ad ogni baracto:
lo cielo colla terra ne vo gire invitando.*

Con differenze minime, legate principalmente alla presenza di occorrenze lemmatizzate, il risultato mostrato coincide col file filgat.ddt.txt, posto nella directory Demotesti contenente i file testo del corpus **Demo**.

Il file testo è completo in quanto, come detto, esso deve contenere solo i riferimenti alle note e non le note stesse.

Come rendere indicizzabile il contenuto delle note

Qualora si volessero indicizzare anche i contenuti delle note, un modo potrebbe essere quello di includerle in un paragrafo fittizio *note*, inserito alla fine del testo. In questo paragrafo potrebbero essere scritti i testi delle note, eventualmente preceduti da diciture che indichino ad esempio il numero di ogni nota. Per quanto riguarda il numero della pagina nella quale inserire il paragrafo fittizio *note*, potrebbe trattarsi di un unico numero di pagina arbitrario, magari maggiore dei numeri di pagina effettivamente esistenti nel testo per palesare che si tratta di una pagina fittizia (soluzione adottata nell'esempio sottostante) oppure si potrebbe attribuire ad ogni nota il numero di pagina del testo cui la nota stessa si riferisce:

@@NT

%

\$999\$[...] *va cercando.*

*Al nome di Dio cantiam dell'amore
a laude *1ed a gloria*/ del nostro salvatore
e della beata *2ave*/ vergine fiore
cantiam dell'amore di Giesù gloriando.*

*Or che fai cor meo, che non prendi gioia
del più vago amore ke sia senza noia
e sta nella mente *3si*/ pensoso e noia #[sic]@
d'amore non sente ke non va giubilando.*

*Fra tuata la gente lo mio cor si muta
di *4Christo*/ amor dolcie in cui ò posta la speme
fioriscono le pratora per ogni rivera
e vegnaci il mio amore infra lli fior dilectando.*

*A te lo dico, bella: *5farai mech'un*/ pacto,
ked io pur voglio Ihesu Christo intrazacto,
e mettermene voglio ad ogni baracto:
lo cielo colla terra ne vo gire invitando.*

%note

\$999\$#nota1@ed a gloria (u8f8) #]@ et da gloria (ed8).

#nota 2@aue (u8f8) #]@ ave (ed8).

#nota 3@si (u8f8) #]@ si (ed8).

#nota 4@Christo (u8f8) #]@ Xpo (ed8).

#nota 5@farai mech'un (u8f8) #]@ faraime chun (ed8).

Nel paragrafo 'note' sono stati inseriti alcuni codici per motivi pratici:

- il numero di pagina fittizio "**\$999\$**" per palesare che non si tratta di una pagina di testo vero e proprio,
- le indicazioni "#nota n@", destinate ad essere visualizzate ma non indicizzate, per riconoscere a quale nota appartengano le parole incluse al suo interno e localizzate in una ricerca,
- i codici "8" per i punti di abbreviazione,
- un punto alla fine di ogni nota, per inserire un fine periodo,
- campi formula attorno alle parentesi quadre chiuse, per evitare messaggi di errore relativi alla chiusura di integrazioni esterne mai aperte.

In ogni caso per poter funzionare come tali le note, a prescindere dal duplicato indicizzabile descritto in questa Lezione, vanno comunque inserite anche seguendo il meccanismo illustrato nella Lezione seguente.

LEZIONE 7. COME PREPARARE IL TESTO ASSOCIATO CONTENENTE LE NOTE

Facendo riferimento all'esempio trattato nella Lezione precedente, vediamo ora come completare l'operazione di inserimento delle note, digitando materialmente queste ultime.

Le note in questione sono quelle riportate di seguito, indicate dai numeri 1,...,5.

⁽¹⁾ *ed a gloria (u.f.)] et da gloria (ed.)*

⁽²⁾ *aue (u.f.)] ave (ed.)*

⁽³⁾ *sì (u.f.)] si (ed.)*

⁽⁴⁾ *Christo (u.f.)] Xpo (ed.)*

⁽⁵⁾ *farai mech'un (u.f.)] faraime chun (ed.)*

Le codifiche che delimitano il corpo delle note sono identiche a quelle inserite nelle corrispondenti zone del testo base. Perciò una nota dovrà essere iniziata da un asterisco seguito dal numero di nota e dovrà terminare con un asterisco seguito da una barra. Ad esempio, la nota numero 1 andrà digitata:

**1ed a gloria (u.f.)] et da gloria (ed.)*/*

Rispetto all'esempio della Lezione precedente il corpo della stessa nota è ora privo dei codici aggiuntivi precedentemente inseriti al suo interno. Infatti il numero di nota (racchiuso in campo formula) era stato introdotto per palesare a quale parte del testo (la nota 1 appunto) appartenessero ad esempio le occorrenze di *gloria* qualora localizzate a seguito di una ricerca; inoltre nel *testo* non è ammessa la presenza di una parentesi quadra chiusa ma non aperta in precedenza e il punto di abbreviazione va codificato con un "8".

Per contro il contenuto delle note, non essendo indicizzabile, viene mostrato così come è ed ogni nota viene visualizzata in connessione col testo corrispondente.

Il testo associato comprendente le note sarà quindi, semplicemente:

**1ed a gloria (u.f.)] et da gloria (ed.)*/*

**2aue (u.f.)] ave (ed.)*/*

**3sì (u.f.)] si (ed.)*/*

**4Christo (u.f.)] Xpo (ed.)*/*

**5farai mech'un (u.f.)] faraime chun (ed.)*/*

Notare che, quando una certa nota viene chiamata in causa, da questo file viene prelevato solo il contenuto della nota, quindi solo il testo racchiuso tra i codici **nn* e **/*. Di conseguenza tutto ciò che eventualmente sia presente nel file al di fuori di queste coppie di codici non verrà utilizzato e sarà ignorato da GATTO.

Si può quindi, volendo, utilizzare questa caratteristica per inserire nel file dei commenti che ne rendano più agevole l'interpretazione per l'utente, magari per successivi interventi di modifica. Notare anche che le note possono essere inserite nel file in ordine qualunque. Vediamo come potrebbe essere riscritto il file delle note, tenendo conto di queste considerazioni:

Questo file contiene le note relative al testo DDT.

Quella che segue è la nota relativa a...

**1ed a gloria (u.f.)] et da gloria (ed.)*/*

**5farai mech'un (u.f.)] faraime chun (ed.)*/*

La note seguenti sono state suggerite da...

**2aue (u.f.)] ave (ed.)*/*

**3sì (u.f.)] si (ed.)*/*

**4Christo (u.f.)] Xpo (ed.)*/*

Il testo così costruito dovrà infine essere salvato in un file di nome predefinito, ovvero **note.<sigla>.txt**, dove <sigla> è, al solito, la sigla alfanumerica da uno a tre caratteri che identifica il testo base. Anche questo file dovrà essere in formato ANSI puro, ovvero contenente solo testo, senza codici di formattazione.

Immaginando quindi che **abc** sia la sigla del testo cui si sta lavorando nell'esempio, il testo base andrà salvato in un file di nome **filgat.abc.txt** e il testo contenente le note nel file **note.abc.txt**.

LEZIONE 8. COME PREPARARE UN FILE TESTO COMBINANDOLO CON UNA DIVERSA EDIZIONE O UNA TRADUZIONE IN ALTRA LINGUA

Una delle possibilità offerte da GATTO consiste nel visualizzare, parallelamente ai contesti localizzati nelle ricerche, i contesti corrispondenti tratti da una diversa edizione del testo o da una sua traduzione in altra lingua.

Tale funzionalità viene descritta in questo punto del manuale perché sfrutta un meccanismo molto simile a quello utilizzato per gestire le note.

È importante aver chiaro che il testo *parallelo*, ovvero quello contenuto in un'altra edizione o versione, viene visualizzato ma non è indicizzabile, al pari delle note.

Il punto di partenza consiste nel ripartire il testo base e il testo associato in *brani* corrispondenti, intendendo con questo termine partizioni di testo lunghe a piacere ma caratterizzate dal fatto di corrispondersi. Ciò significa, ad esempio, che la traduzione del *brano* numero 19 del testo base è costituita esattamente dal *brano* numero 19 del testo associato.

Per separare un brano dal successivo, sia nel testo base che nel testo associato, si fa uso di un asterisco, ovvero del codice **"*"**. Gli asterischi possono essere liberamente inseriti nei file testo, a condizione di rispettare i seguenti vincoli:

1. Non collocare un asterisco come primo o ultimo carattere del testo (con lo scopo di indicare, rispettivamente, l'inizio del primo brano e la fine dell'ultimo) in quanto l'asterisco è un *separatore* tra *brani* successivi, quindi si presuppone che sia preceduto e seguito da un brano: la parte di testo che va dall'inizio del testo al primo asterisco inserito viene automaticamente considerata come primo brano del testo; analogo meccanismo vale per la parte finale del testo.
2. Non inserire un asterisco immediatamente prima di una cifra, in quanto il codice verrebbe interpretato come "numero di nota" (vedere la Lezione 6): dovendo inserire un asterisco come separatore prima di un numero, separarli con uno spazio o collocare l'asterisco immediatamente a ridosso della parola precedente (ad esempio, volendo inserire un asterisco tra **presenti** e **10** nel testo "...erano presenti 10 persone...", si può scrivere "...erano presenti* 10 persone...").
3. Non inserire il codice "*" all'interno di un campo formula, in quanto sarebbe considerato un semplice segno di interpunzione e non verrebbe considerato un separatore tra brani.
4. Non inserire il codice "*" all'interno del titolo di un riferimento organico (righe che iniziano col codice %), in quanto sarebbe considerato un carattere facente parte del titolo e non verrebbe considerato un separatore tra brani.

Se ad un file testo è destinato ad essere associato un testo *parallelo* contenente una diversa edizione o una traduzione la prima riga del file testo dovrà essere costituita solo dal codice **@@TR**.

Vediamo un esempio concreto:

(il testo manca di un titolo)

(inizio pagina 3)*Queste sonno le storie
de Troia e de Roma.*

*Ad honore de lo omnipotente Dio
e ad utilitate de li homini ke questo
libro legeraco e userannolo de
legere ke lo faza sapio*

*De Iason e de lo pecorone e
de Laumedoth, rege de Troia.*

*In quello tempo in Grecia foro doi
fratri, Eson e Pelias. Pelias non
avea filio masculo, ma presore filie.*

Eson avea filio Iasone, lo quale era dicto filio de dea Cereris e avea bona agura ne li sementi de la terra. Pelias avenno pagura de Iasone suo nepote, ke era molto sapio e ardito, sottrasselo e gioli ad tradimento komo devesse morire. Et dixit: «Filio mio ne l'isola de Colcho ene una ventura de uno pecorone, ke ao la lana de l'auro e ene facto ad honore de dio Iovis. Se tu me la duci, io te donno la midate de lo regno mio». Extimanno ka potea morire de la ventura de lo pecorone...

Segue una (libera) versione latina.

Incipiunt multe ystorie et troiane et romane De Iasone et Thosone aureo. Eo tempore fuerunt duo fratres Eson et Pelias. Pelias non habuit masculam prolem set filias plurimas. Eson vero habuit filium Iasonem . qui dictus est filius Cereris eo quod multo habundavit in frugibus terrenis . Pelias timens ne Iason nepos eius sibi regnum auferret . eo quod vir probus erat et strenuus . ipsi calide persuasit ut iret in Pontum insulam et vellus aureum inde auferret . quod erat simulacrum ad honorem Iovis . ubi revera ostendabantur miracula . existimans illum propter ferocitatem Scitarum periturum . quod si faceret dicebat medietatem sui regni sibi daturum .

Inseriamo ora una ragionevole ripartizione in brani nel testo base:

(il testo manca di un titolo)

(inizio pagina 3) *Queste sonno le storie de Troia e de Roma.*

**Ad honore de lo omnipotente Dio e ad utilitate de li homini ke questo libro legeraco e userannolo de legere ke lo faza sapio*

**De Iason e de lo pecorone e de Laumedoth, rege de Troia.*

**In quello tempo in Grecia foro doi fratri, Eson e Pelias. Pelias non avea filio masculo, ma presore filie.*

**Eson avea filio Iasone, lo quale era dicto filio de dea Cereris e avea bona agura ne li sementi de la terra. *Pelias avenno pagura de Iasone suo nepote, ke era molto sapio e ardito, sottrasselo e gioli ad tradimento komo devesse morire. Et dixit: «Filio mio ne l'isola de Colcho ene una ventura de uno pecorone, ke ao la lana de l'auro e ene facto ad honore de dio Iovis. Se tu me la duci, io te donno la midate de lo regno mio». Extimanno ka potea morire de la ventura de lo pecorone...*

e corrispondentemente nella traduzione:

Incipiunt multe ystorie et troiane et romane

**De Iasone et Thosone aureo.*

** *Eo tempore fuerunt duo fratres Eson et Pelias. Pelias non habuit masculam prolem set filias plurimas. *Eson vero habuit filium Iasonem . qui dictus est filius Cereris eo quod multo habundavit in frugibus terrenis . *Pelias timens ne Iason nepos eius sibi regnum auferret . eo quod vir probus erat et strenuus . ipsi calide persuasit ut iret in Pontum insulam et vellus aureum inde auferret . quod erat simulacrum ad honorem Iovis . ubi revera ostendabantur miracula . existimans illum propter ferocitatem Scitarum periturum . quod si faceret dicebat medietatem sui regni sibi daturum .*

Come si può vedere, entrambi i testi contengono 5 asterischi e sono quindi divisi in 6 *brani*, peraltro di lunghezza alquanto diversa. Notare anche che i due testi non cominciano e non finiscono con asterischi.

Al terzo brano del testo base "*De Iason...rege de Troia.*" non corrisponde in realtà alcuna traduzione: questo caso si gestisce inserendo al suo posto, nel secondo testo, un semplice spazio; in effetti, nel secondo testo dopo il brano numero 2, ovvero quello che termina con "...*et Thosone aureo.*", compare la sequenza "* *", che di fatto definisce con un semplice spazio vuoto il brano numero 3, passando poi al brano numero 4 che comincia con "*Eo tempore...*", traduzione del corrispondente quarto brano del testo base "*In quello tempo...*".

Il testo associato contenente la traduzione è pronto.

Rimane da completare il testo base, corredandolo dei codici mancanti:

@@TR

%

\$3\$*Queste sonno le storie
de Troia e de Roma.*

**Ad honore de lo omnipotente Dio
e ad utilitate de li homini ke questo
libro legeraco e userannolo de
legere ke lo faza sapio*

**De Iason e de lo pecorone e
de Laumedoth, rege de Troia.*

**In quello tempo in Grecia foro doi
fratri, Eson e Pelias. Pelias non
avea filio masculo, ma presore filie.*

**Eson avea filio Iasone, lo quale era
dicto filio de dea Cereris e avea bona
agura ne li sementi de la terra. *Pelias
avenno pagura de Iasone suo nepote,
ke era molto sapio e ardito, sottrasselo
e gioli ad tradimento komo devesse
morire. Et dixit: «Filio mio ne l'isola
de Colcho ene una ventura de
uno pecorone, ke ao la lana de l'auro
e ene facto ad honore de dio Iovis.*

*Se tu me la duci, io te donno la midate
de lo regno mio». Extimanno ka
potea morire de la ventura de lo pecorone...*

Analogamente al caso del file contenente delle note, il testo contenente la traduzione (o una diversa edizione) dovrà infine essere salvato in un file di nome predefinito, ovvero **trad.<sigla>.txt**, dove <sigla> è, al solito, la sigla alfanumerica formata da uno a tre caratteri che identifica il testo base. Anche questo file dovrà essere in formato ANSI puro, ovvero contenente solo testo, senza codici di formattazione.

Immaginando quindi che **abc** sia la sigla del testo cui si sta lavorando nell'esempio, il testo base andrà salvato in un file di nome **filgat.abc.txt** e il testo contenente la traduzione nel file **trad.abc.txt**.

LEZIONE 9. COME GESTIRE LA PRESENZA CONTEMPORANEA DI NOTE E DI UNA DIVERSA EDIZIONE O TRADUZIONE IN ALTRA LINGUA

Gli esempi considerati nelle due Lezioni precedenti si possono combinare per trattare il caso di un testo che preveda la presenza di una versione tradotta e che allo stesso tempo includa riferimenti a note.

L'unica avvertenza di natura tecnica è relativa al caso in cui il separatore tra brani "*" e il codice di inizio zona interessata da una nota (ad esempio "*5") si trovino nello stesso punto di testo (ciò si verifica quando la parte iniziale di un brano è interessata da una nota): se nel file testo venisse digitata la sequenza "***5" verrebbe segnalato un errore, in quanto non sono accettati due asterischi consecutivi; la soluzione consiste nell'inserire semplicemente uno spazio tra i due asterischi, digitando "* *5": il primo asterisco viene interpretato come separatore tra brani, il secondo come inizio di codifica relativa alla nota numero 5.

Cambia il codice da porre nella prima riga del file testo, che diviene in questo caso **@@A2**.

Come esempio pratico si consideri il testo dell'esempio precedente, con l'aggiunta di alcune note:

(il testo manca di un titolo)

(inizio pagina 3)*Queste sonno le storie
de Troia e de Roma.*

*Ad honore de lo omnipotente Dio
e ad utilitate de li homini ke questo
libro legeraco e userannolo de
legere ke lo faza sapio*

*De Iason e de lo pecorone e
de Laumedoth, rege de Troia.*

*In quello tempo in Grecia foro doi
fratri, Eson e Pelias. Pelias non
avea filio masculo, ma presore filie.*

*Eson avea filio Iasone, lo quale era
dicto filio de dea Cereris e avea bona
agura ne li sementi de la terra. Pelias
avenno pagura de Iasone suo nepote,
ke era molto sapio e ardito, sottrasselo
e gioli ad tradimento komo devesse
morire. Et dixit: «Filio⁽¹⁾ mio ne l'isola⁽²⁾
de Colcho ene una ventura de
uno pecorone, ke ao la lana de l'auro⁽³⁾
e ene facto ad honore de dio Iovis.*

*Se tu me la duci, io te donno la midate
de lo regno mio». Extimanno ka
potea morire de la ventura de lo pecorone...*

note

⁽¹⁾ dixit: «Filio (u.f.)] dixit filio (ed.)

⁽²⁾ l'isola (u.f.)] lisola (ed.)

⁽³⁾ l'auro (u.f.)] lauro (ed.)

Corredato della stessa traduzione:

*Incipiunt multe ystorie et troiane et romane
De Iasone et Thosone aureo.*

*Eo tempore fuerunt duo fratres Eson et Pelias. Pelias non habuit
masculam prolem set filias plurimas. Eson vero habuit filium
Iasonem . qui dictus est filius Cereris eo quod multo habundavit
in frugibus terrenis . Pelias timens ne Iason nepos eius sibi regnum
auferret . eo quod vir probus erat et strenuus . ipsi calide persuasit*

ut iret in Pontum insulam et vellus aureum inde auferret . quod erat simulacrum ad honorem Iovis . ubi revera ostendabantur miracula . existimans illum propter ferocitatem Scitarum periturum . quod si faceret dicebat medietatem sui regni sibi daturum .

Seguendo quanto illustrato nelle due Lezioni precedenti è facile vedere che i tre file cui si arriva, una volta inserite le opportune codifiche, sono:

filgat.<sigla>.txt:

@@A2

%

\$3\$*Queste sonno le storie de Troia e de Roma.*

**Ad honore de lo omnipotente Dio e ad utilitate de li homini ke questo libro legeraco e userannolo de legere ke lo faza sapio*

**De Iason e de lo pecorone e de Laumedoth, rege de Troia.*

**In quello tempo in Grecia foro doi fratri, Eson e Pelias. Pelias non avea filio masculo, ma presore filie.*

**Eson avea filio Iasone, lo quale era dicto filio de dea Cereris e avea bona agura ne li sementi de la terra. *Pelias avenno pagura de Iasone suo nepote, ke era molto sapio e ardito, sottrasselo e gioli ad tradimento komo devesse morire. Et *1dixe: «Filio*/ mio ne *2l'isola*/ de Colcho ene una ventura de uno pecorone, ke ao la lana de *3l'auro*/ e ene facto ad honore de dio Iovis.*

Se tu me la duci, io te donno la midate de lo regno mio». Extimanno ka potea morire de la ventura de lo pecorone...

note.<sigla>.txt:

**1dixe: «Filio (u.f.)] dixe filio (ed.)*/*

**2l'isola (u.f.)] lisola (ed.)*/*

**3l'auro (u.f.)] lauro (ed.)*/*

trad.<sigla>.txt:

Incipiunt multe ystorie et troiane et romane

**De Iasone et Thosone aureo.*

** *Eo tempore fuerunt duo fratres Eson et Pelias. Pelias non habuit masculam prolem set filias plurimas. *Eson vero habuit filium Iasonem . qui dictus est filius Cereris eo quod multo habundavit in frugibus terrenis . *Pelias timens ne Iason nepos eius sibi regnum auferret . eo quod vir probus erat et strenuus . ipsi calide persuasit ut iret in Pontum insulam et vellus aureum inde auferret . quod erat simulacrum ad honorem Iovis . ubi revera ostendabantur*

*miracula . existimans illum propter ferocitatem Scitarum periturum .
quod si faceret dicebat medietatem sui regni sibi daturum .*

Immaginando al solito che **abc** sia la sigla del testo cui si sta lavorando nell'esempio, il testo base andrà salvato nel file **filgat.abc.txt**, quello contenente le note in **note.abc.txt** e quello contenente la traduzione in **trad.abc.txt**.

Il materiale utilizzato per gli esempi illustrati nelle ultime 3 Lezioni è tratto dal testo R3, incluso nel corpus **Demo**.

CAP. 5 - RACCOLTA DELLE INFORMAZIONI BIBLIOGRAFICHE

Ogni testo incluso in un corpus viene caratterizzato da una serie di informazioni bibliografiche che vanno collocate in un archivio apposito. Queste informazioni sono richiamabili nel corso delle successive operazioni sui testi.

I dati bibliografici previsti sono diversi, ma solo due vanno inseriti obbligatoriamente in quanto necessari in alcune operazioni; l'inserimento di tutti gli altri è facoltativo.

I dati bibliografici previsti verranno ora elencati in ordine alfabetico; in tutti i casi, con esclusione della data codificata, lettere maiuscole e minuscole sono equivalenti.

Abbreviazione titolo

Contiene in forma sintetica informazioni su titolo, autore o altro elemento che faciliti il riconoscimento dell'opera: infatti i primi 20-30 caratteri di questo dato compariranno nelle testate di tutti i contesti localizzati, con lo scopo di evidenziarne la provenienza.

Questa informazione viene visualizzata anche nei casi in cui è richiesto di indicare un testo del corpus a partire dalla sigla, onde renderne più sicura l'identificazione.

Qualora questo dato non venga inserito, gli verrà attribuito automaticamente il valore "=".

Anno finale

Anno finale di datazione dell'opera; se questa è collocata cronologicamente in un unico anno, l'*anno finale* coincide con l'*anno iniziale*. Questo dato è generato in maniera automatica da GATTO sulla base della data codificata.

Anno iniziale

Anno iniziale di datazione dell'opera; se questa è collocata cronologicamente in un unico anno, l'*anno iniziale* coincide con l'*anno finale*. Questo dato è generato in maniera automatica da GATTO sulla base della data codificata.

Area generica

Si intende l'ambiente linguistico esteso in cui si colloca il testo (ad esempio regione).

Area specifica

Si intende l'ambiente linguistico ristretto in cui si colloca il testo (ad esempio città).

Autore

Collocazione interna

Collocazione assegnata all'opera.

Curatore

Data codificata

Quando GATTO come risultato di una ricerca fornisce dei contesti tratti dalle opere incluse in un corpus, tali contesti vengono normalmente ordinati in base alla collocazione cronologica dei testi da cui i contesti provengono. Data l'importanza che la collocazione temporale riveste nell'analisi di documentazione relativa alla lingua antica, sono stati previsti meccanismi alquanto sofisticati per fornire un ordinamento dei contesti ben definito e soggetto a una serie di regole molto precise.

Col termine *data codificata* si intende una speciale stringa di caratteri che viene associata ad ogni testo sulla base della sua collocazione cronologica: la generazione della stringa è studiata in modo che l'ordinamento alfabetico delle date codificate rispecchi rigorosamente l'ordinamento cronologico dei testi corrispondenti: in questo modo, poiché i contesti ottenuti nel corso di una ricerca sono presentati secondo l'ordine alfabetico delle date codificate associate ai relativi testi, si ha la certezza di ottenere i contesti nella sequenza cronologica corretta.

Questa definizione di data codificata verrà ripresa e ampliata nel corso della Lezione che segue. Al suo interno sono illustrati due metodi per generare le date codificate; se però non si ha particolare interesse ad ordinare cronologicamente i testi in maniera precisa, si possono inserire date codificate semplificate ma sintatticamente corrette, a seconda che il testo sia assegnato ad un anno preciso o ad un periodo temporale, seguendo le seguenti indicazioni:

- **anno preciso:** predisporre come *data codificata* una stringa formata dall'anno, immediatamente seguito dalla lettera maiuscola "L". Ad esempio, 1325 => **1325L**;
- **periodo temporale:** predisporre una stringa formata dalle cifre dell'anno finale, immediatamente seguite da quelle dell'anno iniziale e dalla sigla "BL". Ad esempio: 950-1050 => **1050950BL**.

Nella Lezione 12 è riportato un altro modo semplificato per introdurre la *data codificata*, tramite l'inserimento guidato da GATTO.

La data codificata va inserita obbligatoriamente.

Data descrittiva

Collocazione cronologica del testo, liberamente definita in modo discorsivo (es. "fine XIII sec.").

Data di inserimento nel corpus

Data alla quale il testo è stato inserito nel database. Il formato utilizzato è mm/gg/aa. Es.: 3 aprile 2001 => 04/03/01.

Edizione

Forma

Denota la forma del testo, indicata da una lettera. Ad esempio, all'Istituto Opera del Vocabolario Italiano sono utilizzati i seguenti codici: **M** = misto versi/prosa, **P** = prosa, **V** = versi.

Genere

Genere letterario.

Indice di qualità

Indicatore della qualità del testo, nel senso di attendibilità e pertinenza linguistica, congruità cronologica, ecc. Indica in sostanza un testo significativo per la documentazione della specifica varietà linguistica. Per i testi dell'Istituto Opera del Vocabolario Italiano valgono le seguenti codifiche:

- TS** testi significativi per la documentazione della specifica varietà linguistica,
- TPC** testi da citare solo in casi particolari per ragioni di cronologia,
- TPL** testi da citare solo in casi particolari per ragioni linguistiche.

Nota

Disponibile per inserire annotazioni di natura varia.

Sigla

È una stringa alfanumerica, unica nell'ambito del corpus (ovvero non possono esservi due testi con la stessa sigla in uno stesso corpus) che individua il testo. Non c'è differenza tra lettere maiuscole e minuscole. Sono caratteri accettati le 26 lettere dell'alfabeto inglese e le cifre. Data la sua natura, questo dato va inserito obbligatoriamente. Attenzione: dopo che un testo è stato caricato nel corpus non tentare di modificarne la sigla in bibliografia: si potrebbe danneggiare il corpus stesso; per cambiare la sigla di un testo inserito nel corpus vedere all'inizio del capitolo 25.

Tipo

Denota il tipo del testo, indicato da una lettera. Ad esempio, all'Istituto Opera del Vocabolario Italiano sono utilizzati i seguenti codici: **M** = misto originale/volgarizzamento, **O** = originale, **P** = parafrasi, **T** = traduzione, **V** = volgarizzamento.

Titolo

La tabella che segue riporta le lunghezze massime ammesse, in numero di caratteri, per i dati elencati.

Dato bibliografico	Num. max. caratteri
Abbreviazione titolo	70
Area generica	20
Area specifica	23
Autore	50
Collocazione interna	100
Curatore	30
Data descrittiva	35
Edizione	248
Forma	1
Genere	16
Indice di qualità	3
Nota	255
Sigla	3
Tipo	1
Titolo	200

LEZIONE 10. DATAZIONE DEI TESTI. CONCETTO DI DATA CODIFICATA, PASSAGGIO DALLA DATA DESCRITTIVA ALLA DATA CODIFICATA, COSTRUZIONE DELLA DATA CODIFICATA

DATAZIONE DEI TESTI. CONCETTO DI DATA CODIFICATA

Data l'importanza che la collocazione temporale riveste nell'analisi di documentazione relativa alla lingua antica, GATTO esige che sia attribuita ai testi una datazione, al limite anche fittizia. Sono stati previsti meccanismi alquanto sofisticati per fornire un ordinamento dei contesti ben definito e soggetto a una serie di regole molto precise.

Col termine *data codificata* si intende una speciale stringa di caratteri che viene associata al testo sulla base della sua collocazione cronologica: la generazione della stringa è studiata in modo che l'ordinamento alfabetico delle date codificate rispecchi rigorosamente l'ordinamento cronologico desiderato dei testi corrispondenti. Per consentire ordinamenti cronologici sofisticati e che tengano conto anche di incertezze di datazione sono state fissate a priori alcune convenzioni che permettono di tradurre in stringhe utilizzabili dal programma la varietà delle formule che si usano in genere discorsivamente (per esempio "tra la fine del XII secolo e il primo trentennio del XIII"). Queste stringhe convenzionali, nella terminologia di GATTO, sono chiamate *date codificate*.

NOTA

Quella che segue è una descrizione dettagliata dei criteri adottati all'Opera del Vocabolario per convertire una data descrittiva nella corrispondente *data codificata*, rispettando gli specifici criteri di ordinamento temporale adottati nell'Istituto. Tale descrizione, che occupa il resto di questa Lezione, non è di facile lettura e può essere trascurata se non si è direttamente coinvolti nell'attività dell'Opera del Vocabolario.

Per introdurre le date codificate si può semplicemente fare ricorso al meccanismo molto semplificato riportato al termine della parte introduttiva del capitolo 7, oppure utilizzare l'*inserimento guidato* da GATTO, descritto ed esemplificato nella Lezione 12: quest'ultima tecnica richiede all'utente di conoscere solo la data descrittiva che si intende associare al testo, mentre tutte le conversioni necessarie per generare, a partire da essa, la *data codificata* così come descritta in questa e in altre Lezioni, vengono svolte automaticamente dal programma.

PASSAGGIO DALLA DATA DESCRITTIVA ALLA DATA CODIFICATA

Si tratta, come detto, di una codifica speciale che permette di collocare cronologicamente un testo in corrispondenza di un anno specifico o di un intervallo temporale, affiancando a tale datazione ulteriori valutazioni relative al grado di certezza o ai margini di approssimazione dell'attribuzione. La data codificata è usata dal programma non solo per collocare cronologicamente la singola opera ma anche per definire una gerarchia di presentazione dei contesti, a parità di collocazione temporale. Una volta assegnata ad un testo una data codificata, il programma genererà automaticamente, sulla base di questo dato, anche i contenuti dei campi bibliografici **anno iniziale** e **anno finale**, che individuano il lasso di tempo associato al testo; nel caso di anno singolo, i due valori coincidono. Questi due dati sono utilizzati sia per includere o meno un testo in sottocorpus definiti su base temporale sia per collocare i contesti estratti in una precisa sequenza temporale: in questo caso viene utilizzato il valore di **anno finale**.

Per ottenere la data codificata il punto di partenza è, ovviamente, la data descrittiva associata al testo. Dal punto di vista logico il passaggio da data descrittiva a *data codificata* si articola in tre fasi successive:

1. conversione della data descrittiva in un anno o in un periodo temporale ben definito, con valutazione dei margini di approssimazione della datazione,
2. definizione di criteri con cui ordinare testi sulla base delle datazioni,
3. traduzione degli anni, degli intervalli temporali e dei margini di approssimazione citati nel punto 1 in stringhe alfanumeriche il cui ordinamento sposti i criteri stabiliti nel punto 2.

Il punto 1 è piuttosto semplice e basteranno pochi esempi per chiarirne il senso.

Il punto 2 dipende in linea di principio dalle opinioni dell'utente: qui vi è però un limite tecnico: GATTO opera in base al criterio definito ed utilizzato all'Istituto Opera del Vocabolario Italiano e non è possibile modificarne la logica, che verrà comunque descritta in dettaglio.

Il punto 3 può risultare un po' più ostico, data la sua natura strettamente tecnica.

I prossimi paragrafi descriveranno in dettaglio queste tre fasi.

Come già detto, per superare le difficoltà legate al tecnicismo insito in questo meccanismo, GATTO offre in alternativa alla costruzione diretta della *data codificata* un sistema di generazione assistita di quest'ultima, meccanismo che richiede all'utente di fare solo quanto indicato nel punto 1 e di rispondere poi ad alcune semplici domande poste dal programma: GATTO provvederà poi a generare automaticamente la data codificata sulla base delle risposte fornite.

COSTRUZIONE DELLA DATA CODIFICATA

1) Conversione della data descrittiva in un anno o in un periodo temporale definito, con indicazione dei margini di approssimazione dell'attribuzione

Quando la stesura di un testo viene attribuita ad un anno ben preciso non ci sono conversioni da effettuare.

Diverso il caso degli intervalli temporali: qui bisogna definire una convenzione per fissarne gli estremi in funzione dei termini usati nella data descrittiva (es. "primi decenni" può essere interpretata come "i primi 20 anni", o "i primi 30 anni",...). La tabella che segue riporta degli esempi facenti riferimento ai criteri adottati all'OVI (il separatore "/" indica intervallo noto con approssimazione, il separatore "-" indica intervallo noto con precisione).

intervalli espressi con diciture descrittive	conversione in intervalli temporali
sec. XIII in.	1200/1210
sec. XIII secondo decennio	1211/1220
sec. XIII primo quarto	1201/1225
sec. XIII primi decenni	1201/1230
sec. XIII prima metà	1201/1250
sec. XIII secondo quarto	1226/1250
sec. XIII metà	1246/1255
sec. XIII terzo quarto	1251/1275
sec. XIII metà - ante 1276	1246-1276

sec. XIII terzultimo decennio	1271/1280
sec. XIII	1201/1300
sec. XIII metà/seconda metà	1246/1300
sec. XIII seconda metà	1251/1300
sec. XIII ultimi decenni	1271/1300
sec. XIII ultimo quarto	1276/1300
sec. XIII ultimo ventennio	1281/1300
sec. XIII ex.	1291/1300
1291, 1305, 1310	1291-1310
sec. XIII ex. / XIV in.	1291/1310
sec. XIII ex. - XIV in.	1291/1310
sec. XIII ex. / 1310?	1291-1310
sec. XIII ex. - ante 1310	1291-1310
sec. XIII / XIV	1291/1310

Nel caso di attribuzione del testo ad un anno preciso, questo può essere noto con esattezza oppure caratterizzabile da *attributi* che meglio specifichino il senso dell'attribuzione; gli attributi previsti in GATTO sono (si possono anche combinare tra loro):

- <nessun attributo>
- **ante**
- **post**
- **circa**
- **?**

Nel caso di attribuzione del testo ad un intervallo temporale, si possono meglio precisare le caratteristiche di quest'ultimo; le tipologie di intervallo previste in GATTO sono:

- intervallo noto **con approssimazione** (opera composta in un qualche periodo compreso tra i due estremi)
- intervallo noto **con precisione** (opera iniziata e terminata ai due estremi)
- intervallo del tipo **sec.zzz ex/xxxx** oppure **sec.zzz in/xxxx**
- intervallo del tipo **sec.zzz ex-xxxx** oppure **sec.zzz in-xxxx**
- intervallo del tipo **sec.zzz/sec.vvv**

dove "xxxx" indica un anno, "vvv" e "zzz" indicano secoli, "in" = ineunte, "ex" = exeunte.

Sempre nel caso di attribuzione del testo ad un intervallo temporale, all'anno finale di quest'ultimo (l'anno cronologicamente più significativo in quanto più direttamente collegabile all'uscita dell'opera) possono anche venire associati gli stessi attributi previsti nel caso di assegnazione del testo ad un anno preciso, ovvero <nessun attributo>, **ante**, **post**, **circa**, **?**.

2) Criteri di ordinamento cronologico dei testi adottati all'OVI

Questo paragrafo è puramente informativo, nel senso che spiega i criteri, peraltro non modificabili, adottati all'Opera del Vocabolario Italiano e quindi all'interno di GATTO.

A. Confronto tra due testi entrambi datati con un anno preciso:

- si confrontano i due anni;
- a parità, si tiene conto degli eventuali attributi, secondo la gerarchia (xxxx è l'anno):
 - 1°. **ante xxxx**
 - 2°. **ante xxxx circa**
 - 3°. **ante xxxx ?**
 - 4°. **xxxx**
 - 5°. **xxxx circa**
 - 6°. **xxxx ?**
 - 7°. **post xxxx**
 - 8°. **post xxxx circa**
 - 9°. **post xxxx ?**

B. Confronto tra un testo datato con un anno preciso e uno datato con un intervallo:

- si confronta l'anno preciso con l'anno finale dell'intervallo;
- a parità, viene prima l'intervallo.

C. Confronto tra due testi datati con intervalli:

- si confrontano gli anni finali;
- a parità si confrontano gli anni iniziali;
- a parità si confrontano i tipi di intervallo, secondo la gerarchia:
 - 1°. intervallo noto con approssimazione (opera composta in un qualche periodo compreso tra i due estremi)
 - 2°. ovvero intervallo noto con precisione (opera iniziata e terminata ai due estremi)
 - 3°. **sec.zzz ex/yyyy** oppure **sec.zzz in/yyyy**
 - 4°. **sec.zzz ex-yyyy** oppure **sec.zzz in-yyyy**
 - 5°. **sec.zzz/sec.vvv**
dove "xxxx" indica un anno, "vvv" e "zzz" indicano secoli, "in" = ineunte, "ex" = exeunte.
- a parità, si tiene conto degli eventuali attributi, secondo la gerarchia (xxxx è l'anno finale):
 - 1°. **ante xxxx**
 - 2°. **ante xxxx circa**
 - 3°. **ante xxxx ?**
 - 4°. **xxxx**
 - 5°. **xxxx circa**
 - 6°. **xxxx ?**
 - 7°. **post xxxx**
 - 8°. **post xxxx circa**
 - 9°. **post xxxx ?**

3) Passaggio alle date codificate

Una volta passate in rassegna le tipologie di datazione attribuibili ai testi e le opzioni accessorie, si può passare a convertire le diverse combinazioni di questi elementi in sequenze di caratteri il cui mutuo ordinamento alfabetico riproduca fedelmente l'ordinamento cronologico dei testi corrispondenti.

3a) Traduzione degli anni precisi in date codificate

Quando al testo è assegnato un anno preciso, la data codificata è una stringa costituita dalle cifre dell'anno seguite da una lettera maiuscola che tiene conto degli eventuali attributi, secondo la tabella:

attributo	lettera da inserire
ante xxxx	F
ante xxxx circa	G
ante xxxx ?	H
xxxx	L
xxxx circa	M
xxxx ?	N
post xxxx	P
post xxxx circa	Q
post xxxx ?	R

dove xxxx è l'anno.

Ad esempio:

post 1320 circa => **1320Q**
1276 => **1276L**
ante 1310 => **1310F**
post 1295 ? => **1295R**

3b) Traduzione degli intervalli temporali in date codificate

In questo caso la data codificata è formata dalle cifre relative all'anno finale, seguite dalle cifre relative all'anno iniziale, seguite da una lettera maiuscola che definisce il tipo di intervallo secondo lo schema riportato nella tabella seguente, seguite ancora da una lettera maiuscola che tiene conto degli eventuali attributi dell'anno finale, secondo lo schema riportato nella tabella precedente.

tipo di intervallo	lettera da inserire
noto con approssimazione	A
noto con precisione	B
sec.zzz ex/yyyy oppure sec.zzz in/yyyy	C
sec.zzz ex-yyyy oppure sec.zzz in-yyyy	D
sec.zzz/sec.vvv	E

dove "yyyy" indica un anno, "vvv" e "zzz" indicano secoli, "in" = ineunte, "ex" = exeunte.

Ad esempio:

sec XIII metà - ante 1276 => **12761246BF**
 1275-1320 circa => **13201275BM**

3c) Caso di testi con identica data codificata

Nei casi in cui l'applicazione dei criteri suesposti porti a testi aventi la stessa data codificata, è possibile ordinarli ulteriormente accordando alle loro date codificate la stringa "-nnnnn", ovvero un numero arbitrario di 5 cifre preceduto da un trattino. In questo modo l'ordinamento, a parità di data codificata basata su criteri strettamente temporali, ha luogo per valori crescenti dei numeri così inseriti.

Ad esempio:

13201275BM-00055

viene prima di

13201275BM-00112

4) Generazione assistita della data codificata

Come accennato in precedenza, per ovviare al fatto che la costruzione della data codificata è alquanto complessa, GATTO offre la possibilità di procedere alla sua generazione automatica, sulla base delle seguenti informazioni che l'utente dovrà predisporre insieme a tutte le altre (dati bibliografici ecc.):

- nel caso di datazione mediante anno preciso, l'anno ed eventuali informazioni sull'attendibilità del dato (vedere Lezione 12),
- nel caso di datazione mediante intervallo temporale, l'anno iniziale, l'anno finale, le eventuali informazioni sull'attendibilità dei dati e il tipo di intervallo (vedere Lezione 12).

CAP. 6 - DEFINIZIONE DEI PARAMETRI DA ASSEGNARE AL CORPUS

Una volta che sono stati preparati i file testo da introdurre in un corpus, insieme alle relative informazioni bibliografiche, la predisposizione dei dati necessari per operare con GATTO è quasi completa.

Restano da definire alcuni altri dati che caratterizzeranno non tanto questo o quel testo da inserire nel corpus ma proprio quest'ultimo; si tratta di dati che andranno inseriti al momento della costruzione del corpus e che verranno ora analizzati singolarmente.

Nome del corpus

Il nome da assegnare al corpus è una stringa lunga da 1 a 12 caratteri alfanumerici. Quando il corpus verrà costruito, verrà creata una cartella avente lo stesso nome ed *estensione gat*. Ad esempio, il corpus **Demo** si trova nella cartella "demo.gat". A causa di questo meccanismo, vi può essere un solo corpus con un certo nome in un disco, mentre vi possono essere corpus diversi nello stesso disco (ciascuno nella sua cartella) o corpus con lo stesso nome ma collocati in dischi diversi. Inoltre, la cartella del corpus viene collocata al livello più alto, come ad esempio "C:\demo.gat" o "D:\demo.gat"; se il corpus venisse trasferito manualmente in una sottocartella (ad esempio "C:\corpora\demo.gat"), GATTO non sarebbe più in grado di localizzarlo e utilizzarlo.

Unità disco

È il disco nel quale verrà collocata la cartella del corpus. Nella maggior parte dei casi sarà il disco "C".

Descrizione del corpus

Si tratta di una stringa identificativa che verrà visualizzata ogni volta che si vuole aprire un corpus; può contenere, in forma compatta, informazioni sul contenuto del corpus, la data della sua creazione, la versione o altri dati utili per distinguerlo tra i corpus localizzati da GATTO.

Directory dei testi

Cartella destinata a contenere i file testo, i file associati del tipo note.<sigla>.txt o trad.<sigla>.txt, i file contenenti le lemmatizzazioni, insomma i file che GATTO utilizza in associazione con i testi. Questa cartella può trovarsi su qualunque disco e può essere liberamente collocata dal punto di vista gerarchico (può cioè essere anche una sottocartella, o una sottocartella di sottocartella, ecc.).

I file testo, inclusi in questa cartella e relativi a testi che, una volta introdotti nel corpus, vengono poi lemmatizzati, non devono essere modificati in alcun modo: in caso contrario non sarà possibile estrarre successivamente i testi lemmatizzati dal corpus. Qualora si debbano apportare modifiche a questi testi, seguire scrupolosamente la procedura illustrata nelle Lezioni 49, 50 e 51.

Se questa cartella non esiste, verrà creata al momento della costruzione del corpus. Più corpus possono condividere la stessa directory dei testi, la quale può contenere anche file che nulla hanno a che vedere con GATTO.

Non utilizzare per nessun motivo come directory dei testi la directory di un corpus esistente, ad esempio c:\demo.gat.

Archivio bibliografico

Archivio destinato a contenere i dati bibliografici dei testi inclusi nel corpus. Anche l'archivio bibliografico può trovarsi in qualunque cartella o sottocartella di ogni livello. Se l'archivio bibliografico non esiste verrà creato al momento della costruzione del corpus; stessa cosa dicasi per la cartella destinata a contenerlo. Questa, volendo, può coincidere con la directory dei testi. Un archivio bibliografico può contenere dati di testi afferenti a corpus diversi o non inseriti in nessun corpus.

In ogni caso l'archivio bibliografico deve essere creato da GATTO, in modo che durante la sua generazione venga dotato della struttura interna compatibile con quanto richiesto dal programma.

Non inserire per nessun motivo un archivio bibliografico nella directory di un corpus esistente, ad esempio c:\demo.gat.

Opzione ' ? e ! considerati punteggiatura debole'

Al momento della costruzione del corpus verrà richiesto se i segni di interpunzione "?" e "!" debbano essere considerati all'interno dei testi segni di punteggiatura forte, operanti cioè come separatori tra periodi, o meno. Bisogna pertanto decidere cosa si risponderà a quel momento. La risposta di default è sì.

Opzione '... considerati punteggiatura debole'

Opzione del tutto analoga alla precedente e relativa ai puntini di sospensione. La risposta di default è sì.

Opzione 'uso di cifre arabe in forme e lemmi'

In assenza di questa opzione tutti i numeri arabi presenti nel testo dovranno essere racchiusi in campo formula, per evitare segnalazioni di errore da parte dell'interprete.

Se l'opzione è attiva tutti i numeri presenti nel testo verranno indicizzati al pari delle forme; verranno indicizzate anche stringhe costituite da combinazioni di lettere e cifre. Questa opzione può trovare applicazione, ad esempio, volendo indicizzare anche le date presenti nel testo.

La stessa opzione opera in maniera analoga sui lemmi, che potranno quindi contenere cifre al loro interno.

Per quanto riguarda iperlemmi e disambiguatori non vi sono particolari controlli sui caratteri costituenti, a prescindere dall'impostazione corrente di questa opzione.

Opzione 'gestione e uso di pagina e riga'

I riferimenti topografici (volume, pagina, colonna, riga) possono essere riportati o no nei contesti estratti dal corpus a video, a stampa e su file. Si dovrà decidere se farne uso o meno, avendo ben presente che questi dati si riferiscono non tanto all'opera originaria quanto all'edizione presa in considerazione; il loro uso, comunque, agevola il lavoro con GATTO in varie fasi del suo funzionamento e nel manuale si farà conto che questa opzione sia stata attivata. Tenere anche presente che se si decide di farne uso i corrispondenti codici dovranno essere preventivamente inseriti nei file testo. Verranno ora illustrate le alternative disponibili. La scelta di default è **disabilitato**, a tutela dei diritti di editore.

- **disabilitato.** In questo caso i testi possono contenere o meno le indicazioni dei numeri di pagina; queste possono comparire in tutti i testi, solo in alcuni o in nessuno; possono inoltre essere presenti solo in alcune parti di uno stesso testo: in tutti i casi GATTO ignorerà tali informazioni e non le visualizzerà;
- **facoltativo.** Anche in questo caso i testi possono trovarsi in tutte le diverse condizioni descritte nel caso precedente: la differenza sta nel comportamento di GATTO, che visualizzerà automaticamente pagina e riga per tutti i contesti il cui numero di pagina sia dichiarato e diverso da zero. In altre parole, tutte le occorrenze tratte da testi o parti di testi prive di numero di pagina o caratterizzate dal codice \$0\$ (pagina 0) non daranno luogo alla visualizzazione di questi dati; tutte le altre invece ne saranno automaticamente corredate;
- **obbligatorio.** Tutti i testi dovranno contenere al loro interno, a partire dall'inizio del file filgat, codici di pagina \$nnn\$, con nnn diverso da 0. In caso contrario il testo non verrà accettato nel corpus.

Lunghezze massime, in caratteri, dei dati destinati a finire negli archivi.

Al momento della costruzione del corpus verrà chiesto di fissare le dimensioni massime che certe categorie di dati potranno avere al suo interno: valori troppo piccoli potrebbero risultare incompatibili con forme, lemmi o altro utilizzati all'interno del corpus, mentre valori inutilmente elevati sprecherebbero spazio disco e potrebbero rallentare gli accessi agli archivi senza che ve ne sia motivo.

La tabella che segue riporta le categorie di dati interessate, unitamente ai valori assegnati da GATTO per default.

Notare che nel caso di questi dati, così come per quanto concerne i default delle precedenti opzioni e altri valori preimpostati in GATTO, questi valori iniziali possono essere sostituiti con altri a piacere dell'utente, come sarà meglio visto in seguito.

dato	lunghezza massima in car.
forme	26
lemmi	26
categorie grammaticali	10
disambiguatori lemmi	26
commenti lemmi	50
iperlemmi	26
disambiguatori iperlemmi	26

Il valore ricorrente "26" deriva dalla parola "precipitevolissimevolmente", forse la più lunga parola della lingua italiana, e non tiene conto dell'opportunità di prevedere dimensioni maggiori qualora si vogliano considerare forme polirematiche.

Sia questi valori che le corrispondenti impostazioni iniziali possono essere modificati, per tutti potendo scegliere lunghezze massime comprese tra 10 e 250 caratteri.

Limiti alla possibilità di modificare successivamente queste impostazioni

È importante avere ben chiaro che la maggior parte delle impostazioni sopradescritte, una volta assegnata, non può essere più modificata. Fanno eccezione "descrizione del corpus", "directory dei testi" e "archivio bibliografico", i cui valori potranno essere cambiati in seguito (in pratica anche l'unità disco è di fatto modificabile trasferendo manualmente il corpus in un altro disco).

Tutte le altre impostazioni una volta introdotte verranno congelate e, trovandosi nella necessità di cambiarle, l'unico modo sarà quello di costruire un nuovo corpus, dotato delle nuove impostazioni, e di ritrasferirvi successivamente i testi desiderati.

PARTE II - COSTRUZIONE DI UN CORPUS E INSERIMENTO DI TESTI

CAP. 7 - COSTRUZIONE DI UN CORPUS E DEL RELATIVO ARCHIVIO BIBLIOGRAFICO

Una volta predisposti tutti i dati necessari per costruire un corpus e introdurre i testi si può procedere all'effettuazione di queste operazioni. Prima però è opportuno riassumere il quadro delle informazioni che andranno fornite al computer, onde sincerarsi di disporre prima di cominciare; questo anche in considerazione del fatto che molti di questi dati una volta assegnati non saranno più modificabili se non, ovviamente, cancellando il corpus e costruendone uno nuovo con le caratteristiche modificate.

I dati da fornire in relazione al nuovo corpus e all'archivio bibliografico sono:

1. nome del corpus (max. 12 caratteri)
2. unità disco destinata a contenere il corpus ("C", "D",...)
3. breve descrizione del corpus (20 - 30 caratteri)
4. directory destinata a contenere i file testo
5. nome dell'archivio bibliografico associato al corpus e directory destinata a contenerlo
6. valore di punteggiatura debole o forte da attribuire ai segni ? e !
7. valore di punteggiatura debole o forte da attribuire ai puntini di sospensione
8. utilizzo o meno dei numeri pagina e di riga nei testi del corpus
9. lunghezze massime in caratteri per: forme, lemmi, categorie grammaticali, disambiguatori per lemmi, commenti ai lemmi, iperlemmi, disambiguatori per iperlemmi

Se non si hanno idee precise in merito ai punti 6, 7 e 9 ci si può affidare in partenza ai valori di default suggeriti da GATTO. A questo riguardo va notato che il default proposto per i punti 6 e 7 (punteggiatura debole) è pensato per i testi in lingua antica, mentre per quelli in lingua moderna potrebbe essere più opportuna la scelta opposta. Circa il punto 9 va ribadito che il valore di 26 caratteri proposto per le forme nasce dalla parola 'precipitevolissimevolmente', probabilmente la più lunga parola della lingua italiana; per analogia lo stesso valore è stato proposto per le lunghezze massime di altri elementi, quali i lemmi. Questa scelta non tiene conto dell'eventuale presenza di forme o lemmi polirematici, la quale potrebbe suggerire il ricorso a lunghezze maggiori.

Per ogni testo da inserire nel corpus si dovrà disporre del file testo `filgat.<sigla>.txt` e degli eventuali file associati `note.<sigla>.txt` e `trad.<sigla>.txt`. Inoltre si dovranno conoscere i dati bibliografici da associare al testo; alcuni di questi (scritti in rosso) vanno inseriti obbligatoriamente:

1. sigla
2. data codificata
3. abbreviazione titolo

Per la data codificata, che è una stringa utilizzata dal programma per presentare nel corretto ordine cronologico i contesti localizzati, è stato già indicato un modo semplicissimo per costruirla, qualora non si abbia interesse ad un ordinamento cronologico particolarmente rigoroso; questo criterio semplificato viene ora ribadito:

- se la datazione del testo fa riferimento ad un anno preciso, la data codificata può essere costituita dalle cifre di tale anno seguite dalla lettera maiuscola "L". Ad esempio, 1325 => **1325L**;
- se invece la datazione del testo fa riferimento ad un periodo compreso tra due anni estremi, la data codificata può essere formata con le cifre dell'anno finale, seguite da quelle dell'anno iniziale e dalla sigla "BL". Ad esempio: 950-1050 => **1050950BL**.

L'"abbreviazione titolo" a rigore non è obbligatoria, ma il suo utilizzo è decisamente consigliato poiché in certe situazioni il testo cui GATTO fa riferimento è riconoscibile proprio da questo dato.

Le Lezioni che seguono mostreranno come costruire il corpus di nome **Prova** posto nel disco **C**; la directory dei testi dovrà essere **C:\gattesti** e dovrà contenere anche la bibliografia **esebib.mdb**.

Nel corpus verranno introdotti i testi denotati dalle sigle **es1**, **es2** e **es3**, l'ultimo dei quali prevede anche due file associati contenenti rispettivamente delle note e una traduzione in latino del testo base. I file da

utilizzare sono tutti inclusi nella cartella **esempi** del CD di GATTO. Nella prossima Lezione verrà mostrato come costruire **esebib.mdb**: in ogni caso una copia già generata si trova nella cartella **esempi**.

LEZIONE 11. COME COSTRUIRE UN CORPUS COMPLETO DI ARCHIVIO BIBLIOGRAFICO

Ci proponiamo di costruire nel disco **C** un corpus testuale di nome **Prova**.

A questo corpus verrà associata la descrizione **corpus di prova contenente 3 testi**.

I file testo cui il corpus farà riferimento dovranno essere contenuti nella cartella **C:\gattesti**: si tratta della cartella indicata col termine *directory dei testi* (di quel corpus).

La bibliografia cui il corpus **Prova** farà riferimento dovrà essere **esebib.mdb** e verrà collocata nella *directory dei testi* **C:\gattesti**. Naturalmente avrebbe potuto essere collocata in qualunque altra *directory*, già esistente o meno, così come la *directory dei testi* avrebbe potuto coincidere con qualunque cartella già esistente.

In questo corpus i segni **?**, **!** e **...** verranno considerati punteggiatura debole e verrà fatto uso di pagine e righe. Inoltre la lunghezza massima delle categorie grammaticali verrà estesa dal default proposto di 10 caratteri al valore di 20 caratteri.

Per cominciare, aprire GATTO e portarsi nell'*ambiente Gestione base dati* mediante [**GEST. BASE DATI**]. Selezionare le operazioni sui corpus con [**CORPUS**]. Probabilmente come unico corpus presente apparirà il corpus **Demo**, se installato.

Per avviare la costruzione di un nuovo corpus usare [**GESTIONE.. + CREA**]: apparirà una finestra per selezionare nome e collocazione del nuovo corpus. Nella casella *corpus* digitare **Prova**; nel sottostante box a tendina se il disco proposto è diverso da **C** cliccare sulla freccia a destra per visualizzare i dischi disponibili, quindi cliccare sul disco **C**. Procedere con (**OK**) (o fermare tutto con (**Annulla**)). Apparirà una nuova finestra.

Nella casella *descrizione del corpus*: digitare **corpus di prova contenente 3 testi**; cancellare l'attuale contenuto della casella *directory dei testi*: e digitarvi **c:\gattesti**; se la *directory* fosse stata già presente, per selezionarla andavano utilizzate le caselle sottostanti, scegliendo prima il disco e poi la *directory* voluta (una volta scelto il disco, per espandere il contenuto del sottostante riquadro fare un doppio clic sulla sua prima riga, indicante l'unità disco scelta). Al solito, volendo procedere usare (**OK**). Apparirà il messaggio **la directory c:\gattesti non esiste. Vuoi crearla ?**; ovviamente cliccare su (**Si**). Apparirà la finestra successiva.

Volendo è ora possibile modificare la dicitura precedentemente inserita nella casella *descrizione del corpus*:. Digitare nella casella *bibliografia*: nome e collocazione dell'archivio bibliografico da creare, ovvero **c:\gattesti\esebib.mdb**; notare che l'estensione **mdb** può anche essere omessa, nel qual caso verrà aggiunta automaticamente dal programma (se la bibliografia e la sua *directory* fossero state già presenti, bastava selezionarle usando le caselle sottostanti: prima selezionare l'unità disco, poi la *directory* nel riquadro in basso a sinistra con un doppio clic, poi la bibliografia nel riquadro in basso a destra; se si fosse trattato di una nuova bibliografia in una *directory* già esistente bastava selezionare disco e *directory* aggiungendo poi a mano nella casella *bibliografia*: il nome del nuovo archivio bibliografico). Procedere con (**OK**). Apparirà il messaggio **l'archivio bibliografico c:\gattesti\esebib.mdb non esiste. Vuoi crearlo ?**; ovviamente cliccare su (**Si**). Apparirà il messaggio di conferma dell'avvenuta creazione dell'archivio bibliografico. Cliccando su (**OK**) si passa all'ultima finestra relativa alla creazione di un corpus.

Come detto in precedenza ai segni **?**, **!** e **...** verrà lasciato il valore di punteggiatura debole, così come proposto dal programma; rimarrà anche escluso l'uso di cifre arabe all'interno di forme e lemmi; si cliccherà invece su **obbligatorio** alla destra del quarto riquadro per attivare gestione e uso dei numeri di pagina e di riga della pagina. Sulla parte destra della finestra sono riportate le lunghezze massime ammesse per gli elementi destinati ad essere inseriti negli archivi di GATTO: queste scelte riguardano esclusivamente il corpus in fase di costruzione. Come detto in precedenza, modificheremo la lunghezza massima ammessa per le categorie grammaticali, portandola da 10 a 20 caratteri: ciò si ottiene cliccando nella casella contenente **10**, cancellando questo valore e digitando **20** al suo posto. Va tenuto presente che per tutti i tipi di dati possono essere attribuiti valori compresi tra **10** e **250**; conviene cercare di mantenersi su valori bassi per non aumentare inutilmente le dimensioni degli archivi e anche perchè le caselle che si aprono nelle diverse finestre di GATTO sono di dimensioni limitate e renderebbero un po' scomoda la lettura di dati sistematicamente molto lunghi.

È importante notare che tutte le scelte fatte in questa finestra, una volta rese effettive col tasto (OK) non potranno più essere modificate, se non cancellando e ricostruendo il corpus. La stessa considerazione vale per il nome e la collocazione assegnati al corpus.

Invece nota descrittiva, directory dei testi e archivio bibliografico potranno, volendolo, essere modificati in un secondo tempo.

Tornando alla finestra delle impostazioni del nuovo corpus, se non si è sicuri delle modifiche effettuate si può tornare ai valori preimpostati cliccando su **(Mantieni i default)** e ripartire quindi con nuove modifiche. Quando si è sicuri, completare la costruzione del corpus con **(OK)**. Notare che il bottone **(Annulla)** non è più presente perchè a questo punto l'operazione di costruzione del nuovo corpus verrà comunque effettuata.

Cliccando su **(OK)** apparirà il messaggio di conferma della creazione del nuovo corpus. Cliccando di nuovo su **(OK)** il messaggio sparirà e il contenuto della finestra con la lista dei corpus localizzati verrà modificato per includere il nuovo. Sarà così visibile la riga includente il corpus **Prova**, la stringa descrittiva e, usando la barra di scorrimento orizzontale posta in basso la relativa bibliografia e la directory dei testi; in questo modo è possibile verificare se i dati digitati sono esattamente quelli voluti in partenza.

A titolo di esercizio, provare ad aprire il nuovo corpus cliccando in un punto qualunque della sua riga e quindi su **[APRI]**: l'avvenuta apertura sarà confermata dalla abilitazione del menu **[INFORMAZIONI..]** e dalla modifica del titolo della finestra, che conterrà ora anche disco e nome del nuovo corpus tra parentesi, ovvero **(C:prova)**. Cliccando su **[INFORMAZIONI.. + STATISTICHE]** vedremo che il corpus non contiene testi nè alcun altro dato, a conferma che un corpus è un semplice contenitore per i testi che vi saranno introdotti in un secondo tempo; uscire dalla finestra col tasto **<Esc>** o cliccando sul simbolo **x** in alto a destra. Cliccando invece su **[INFORMAZIONI.. + IMPOSTAZIONI]** si ritroveranno le impostazioni assegnate al corpus, ovvero opzioni e lunghezze massime degli elementi: si può verificare, ad esempio, che la lunghezza massima accettata per le categorie grammaticali è 20 caratteri. Questa finestra si chiude cliccando su **(Esci)**.

Per uscire dal programma, chiudere la finestra APERTURA E GESTIONE GENERALE DEI CORPUS con **[CHIUDI FINESTRA]** e quindi uscire con **[FINE]**.

Come già detto in precedenza non utilizzare per nessun motivo come directory dei testi la directory di un corpus, ad esempio c:\demo.gat, né inserirvi un archivio bibliografico. Il perché di questa avvertenza è semplice: qualora in futuro si decidesse di cancellare il corpus, magari per ricostruirlo diversamente, l'operazione di cancellazione rimuoverebbe le directory in questione e tutto il suo contenuto, provocando la perdita di quanto vi fosse stato incluso, compresi quindi file testo e bibliografia.

LEZIONE 12. COME INSERIRE E CORREGGERE I DATI IN UN ARCHIVIO BIBLIOGRAFICO

Inserimento dei dati bibliografici

Una volta costruito il corpus destinato ad includere i testi **es1**, **es2** e **es3** e prima di procedere al loro inserimento effettivo è necessario introdurre nell'archivio bibliografico del corpus i record relativi.

Questi i dati bibliografici proposti:

es1

Autore: anonimo

Titolo: Lauda in lucchese

Titolo abbreviato: Poes. an. lucch. XIV in.

Sigla: es1

Area generica: tosc.

Data descrittiva: XIV in. (da tradursi nell'intervallo 1300 - 1310, approssimato)

es2

Autore: anonimo

Titolo: =

Titolo abbreviato: Doc. venez. 1315

Sigla: es2

Area specifica: venez.

Data descrittiva: 1315

es3

Autore: anonimo

Titolo: Storie de Troia e de Roma

Titolo abbreviato: St. de Troia e de Roma Laur. 1252/58

Sigla: es3

Data descrittiva: 1252/58 (da tradursi nell'intervallo 1252 - 1258, approssimato)

Per procedere, avviare GATTO e portarsi nell'*ambiente Gestione base dati* mediante [GEST. BASE DATI].

Tramite [BIBLIOGRAFIE.. + APRI] e la successiva finestra di ricerca localizzare la bibliografia **c:\gattesti\esebib.mdb** e selezionarla con (Apri). Apparirà la finestra GESTIONE ARCHIVIO BIBLIOGRAFICO, col nome dell'archivio aperto nella riga del titolo; l'archivio al momento non contiene alcun record.

Tramite [MODIF. DATI.. + INSERISCI] attivare l'operazione di inserimento di un nuovo record nell'archivio (unica operazione consentita in un archivio vuoto): viene presentata la scheda di immissione dati.

Digitare **Anonimo** nel campo *autore*, **Lauda in lucchese** nel campo *titolo*, **Poes. an. lucch. XIV in. in abbr. tit., es1** in *sigla*, **tosc.** in *area gen.* e **XIV in.** in *data desc.*.

Passare ora alla data codificata. Per quanto visto nella prima parte del capitolo si potrebbe digitare direttamente **13101300AL** ma, volendo introdurre un dato più accurato, si può utilizzare l'*inserimento guidato* offerto da GATTO. Per fare questo cliccare sull'etichetta (**genera data cod.**) posta in basso a destra (per evidenziare che questo oggetto grafico, pur avendo lo stesso aspetto delle etichette associate agli altri campi, è in realtà un *pulsante* cliccabile gli è stata data una colorazione differente). Cliccandolo appare una finestra che chiede di indicare il tipo di data: nel caso in questione cliccare su **Periodo**. Procedere con (**Avanti**) o annullare con (**Annulla**). La nuova finestra chiede di digitare l'anno iniziale del periodo (**1300**) e di scegliere il tipo di periodo (**da anno... ad anno... (periodo approssimato)**). Procedere con (**Avanti**), annullare con (**Annulla**) o tornare alla finestra precedente con (**Indietro**). Supponendo di procedere, si arriva all'ultima finestra, nella quale andrà semplicemente digitato l'anno finale (**1310**). Terminare l'operazione con (**OK**), annullarla con (**Annulla**) o tornare alla finestra precedente con (**Indietro**). Scegliendo (**OK**) la finestra verrà chiusa e si tornerà alla scheda di inserimento dati bibliografici, nella quale si troveranno 3 nuovi valori:

data codificata = **13101300AL**

anno iniziale = **1300**

anno finale = **1310**.

Questi dati completano l'inserimento delle informazioni bibliografiche relative al testo **es1**. Per rendere effettivo l'inserimento (per ora solo impostato) non rimane che utilizzare il menu [REGISTRA], mentre con [ANNULLA] l'inserimento verrebbe fatto abortire. Il record di **es1** viene inserito nell'archivio e la scheda di inserimento viene disabilitata (sfondo grigio).

In modo analogo si potrà procedere per quanto riguarda i testi **es2** e **es3**. Qui verrà illustrata solo la procedura di costruzione delle due date codificate.

Testo es2:

dopo aver cliccato su (**genera data cod.**) scegliere **Anno esatto** e (**Avanti**); nella finestra successiva in *anno?* digitare **1315** e poi (**OK**). Si otterranno nella scheda i valori:

data codificata = **1315L**

anno iniziale = **1315**

anno finale = **1315**.

Salvare la scheda con [REGISTRA].

Testo es3:

dopo aver cliccato su (**genera data cod.**) scegliere **Periodo** e (**Avanti**); nella finestra successiva in *anno iniziale?* digitare **1252** e come *tipo di periodo* scegliere **da anno... ad anno... (periodo approssimato)**, poi cliccare (**Avanti**); nella finestra successiva digitare **1258** in *anno finale?* e (**OK**). Si otterranno nella scheda i valori:

data codificata = **12581252AL**

anno iniziale = **1252**

anno finale = **1258**.

Salvare la scheda con [REGISTRA].

Correzione di dati bibliografici errati

Se uno dei dati risulta non corretto, è facile intervenire. Supponiamo ad esempio che il titolo di **es2**, per il quale è stato digitato semplicemente il simbolo "=", sia invece **Cedola di Angelo Odorigo**. Per cambiare questo dato, procedere così:

1. cliccare in un punto qualunque della riga relativa al testo **es2**: i suoi dati verranno riportati nella scheda, peraltro al momento disabilitata;
2. abilitare la scheda alle modifiche tramite [MODIF. DATI.. + MODIFICA] (le caselle della scheda abilitata hanno lo sfondo bianco);
3. nel campo *titolo* cancellare = e digitare **Cedola di Angelo Odorigo**;
4. registrare la modifica con [REGISTRA].

Uscire dalla finestra GESTIONE ARCHIVIO BIBLIOGRAFICO con [CHIUDI FINESTRA] e chiudere GATTO con [FINE].

CAP. 8 - INSERIMENTO DI UN TESTO IN UN CORPUS

Fino a questo punto è stata descritta la preparazione di un file testo (filgat.<sigla>.txt), formato dal testo propriamente detto e dai codici necessari al programma per comprenderne la struttura. È stato anche visto come costruire la directory in cui inserire il file filgat, come costruire il corpus e il suo archivio bibliografico e infine come inserire in quest'ultimo i dati bibliografici relativi al testo.

Al termine di queste operazioni, però, GATTO non è ancora in grado di interrogare il testo per estrarne i dati lessicografici, in quanto il testo stesso non è stato ancora *analizzato* dal programma e *inserito* nel corpus.

Per completare le operazioni preliminari è necessario effettuare *l'inserimento del testo nel corpus*, operazione che viene effettuata in due tempi: verifica della correttezza sintattica del contenuto del file filgat e, in caso di esito positivo di questo test, analisi del file e copia del suo contenuto in appositi archivi destinati a contenere il testo vero e proprio (trascritto in uno speciale formato ottimizzato ai fini delle operazioni di ricerca) e le informazioni relative alla sua struttura. La funzione di GATTO che svolge questo doppio passaggio si chiama appunto *inserimento testi*: essa presuppone che il file testo sia stato collocato nella directory dei testi e che il record relativo sia stato già inserito nell'archivio bibliografico; se queste operazioni preliminari sono state svolte correttamente l'unica cosa da fare è fornire al programma la sigla del testo da elaborare: tutto il resto dell'operazione procede automaticamente. Se il testo risulterà privo di errori verrà accettato e inserito nel corpus, in caso contrario l'operazione non avrà luogo e verranno segnalati gli errori individuati all'interno di filgat.<sigla>.txt.

Se non si è sicuri della correttezza sintattica del contenuto del file testo, soprattutto se questo ha dimensioni rilevanti, si può in alternativa effettuare inizialmente la sola operazione di controllo della correttezza sintattica del testo. Questo può essere fatto utilizzando la funzione *controllo correttezza*, la quale svolge sostanzialmente lo stesso controllo di validità che viene effettuato durante l'inserimento del testo, ma molto più velocemente.

ACCESSO AL FILE FILGAT.<SIGLA>.TXT DOPO L'INSERIMENTO DEL TESTO NEL CORPUS

Attenzione: una volta che un testo è stato inserito in un corpus, non si deve modificare in alcun modo la copia del relativo file filgat.<sigla>.txt posta nella directory dei testi del corpus.

In caso contrario, una eventuale successiva lemmatizzazione del testo (che verrebbe effettuata sulla base della versione originale di filgat.<sigla>.txt) potrebbe essere effettuata tranquillamente ma, se in seguito si tentasse di estrarre il testo lemmatizzato dal corpus (per correggere il testo stesso, per trasferirlo o copiarlo in un altro corpus o per altre finalità) si otterrebbe una segnalazione di errore e l'operazione non potrebbe essere eseguita.

Volendo apportare modifiche a un testo lemmatizzato seguire invece le procedure riportate nelle Lezioni 49 e 50. Se invece il testo non è lemmatizzato, e solo in questo caso, il filgat.<sigla>.txt può essere modificato, dopo di ch  si pu  procedere alla operazione di *sostituzione* del testo sulla base della nuova versione di quel file (vedere Lezione 51).

LOGICA DI FUNZIONAMENTO IN PRESENZA DI ERRORI NEL FILE TESTO

Durante il controllo di correttezza di un file testo il file filgat.<sigla>.txt viene analizzato riga per riga dall'interprete di GATTO. L'esame di ogni riga ha inizio dal suo primo carattere e normalmente procede fino al termine della stessa, dopo di che si continua con la riga successiva.

Spesso un *interprete*, ovvero un programma che ha anche il compito di analizzare la validità di un insieme di dati, opera analizzando ed accettando questi ultimi finché non si imbatte in una situazione riconosciuta come erronea; questa logica presuppone di conoscere a priori tutte le potenziali situazioni di errore, in modo da riconoscerle al loro verificarsi: in presenza di un errore non previsto l'interprete accetterebbe comunque i dati con conseguenze imprevedibili.

Per evitare il realizzarsi di situazione del tipo di quella descritta l'interprete di GATTO analizza e accetta il testo solo finché è in presenza di sequenze di caratteri incluse tra quelle dichiarate esplicitamente come valide: in questo modo ogni situazione erronea viene automaticamente bloccata.

La differenza tra le due logiche di funzionamento riguarda le sequenze di caratteri di testo che non sono state prese in considerazione nella scrittura dell'interprete: nel primo caso vengono accettate (in quanto non dichiaratamente erronee) con tutti i rischi del caso, mentre nel secondo vengono rifiutate anche se sintatticamente corrette e quindi prive di errori; il verificarsi di questo secondo caso porta alla segnalazione di errori in realtà inesistenti, situazione facilmente rimediabile intervenendo sul codice e che presenta il vantaggio di non incidere negativamente sull'affidabilità del sistema, a differenza di quanto potrebbe accadere facendo ricorso alla logica precedente.

Una conseguenza diretta della logica di funzionamento dell'interprete di GATTO è che, in presenza di una situazione di errore emersa durante l'analisi di una riga di testo, non è più possibile procedere analizzando la parte restante di tale riga, in quanto ormai al di fuori delle sequenze di caratteri valide. L'analisi del testo viene quindi fermata e riparte con la riga immediatamente successiva.

Da quanto detto discende che un ulteriore errore presente nella stessa riga del primo verrà al momento ignorato; se l'utente corregge il primo errore della riga e prova nuovamente ad inserire il testo nel corpus questa volta l'errore successivo verrà segnalato. Ciò spiega perché a volte ripetendo il tentativo di inserimento di un testo in un corpus vengano segnalati errori nuovi, non evidenziati in precedenza.

Consideriamo come esempio di quanto detto il caso seguente:

*...molti [di col+oro che] erano pres+enti
dissero [che] bisognava opporsi...*

una volta corretto l'errore dovuto alla presenza del primo simbolo "+" all'interno della forma "**coloro**" (e senza che sia stata segnalata la presenza del secondo errore), un successivo controllo di correttezza rivelerà la presenza del "+" nella forma "**presenti**", che non era stato localizzato in precedenza a causa dell'interruzione dell'analisi della riga in corrispondenza del primo "+".

OPZIONE RELATIVE ALL'INSERIMENTO DI TESTI

Opzione 'limita avvisi'

Un'altra conseguenza derivante dalla logica dell'interprete è che le informazioni contenute nella parte di riga non analizzata (in quanto successiva ad un errore) non vengono acquisite: questo può condurre ad ulteriori segnalazioni di errori peraltro inesistenti nel file testo.

Chiariamo la cosa con un esempio:

*...molti [di col+oro che] erano presenti
dissero [che] bisognava opporsi...*

In questo brano compaiono due integrazioni indicate dalle parentesi quadre; inoltre all'interno della parola *coloro* è stato inserito erroneamente un carattere non valido ("+"). Il controllo di correttezza di questo brano fornirà 2 segnalazioni di errore, la prima relativa al primo rigo, evidenziante la presenza di un carattere non ammesso all'interno di una forma (cioè il carattere "+"), la seconda relativa alla parentesi quadra aperta nel secondo rigo (di per sé corretta), segnalante *apertura di integrazione dentro integrazione*. Se si tiene presente quanto detto precedentemente (interruzione dell'analisi e passaggio al rigo successivo in presenza di un errore) si capisce che il testo **effettivamente analizzato** dall'interprete è stato:

*...molti [di col+
dissero [che] bisognava opporsi...*

e in questo caso il messaggio di errore è di per sé corretto, in quanto la seconda integrazione risulta aperta prima che sia chiusa la precedente e quindi risulta interna ad essa (non è ammessa, dato che è priva di senso, una integrazione dentro un'altra integrazione).

A causa di questo meccanismo un unico errore iniziale può in certi casi produrre più messaggi di errore in cascata. Supponiamo ad esempio che nella parte non analizzata di una riga vi sia il codice che indica apertura di un campo formula #; nelle righe successive al termine del campo formula nel testo comparirà il codice di chiusura @; poiché il codice di apertura di un campo formula non è stato incontrato dall'interprete, quest'ultimo non si è precedentemente accorto dell'esistenza di tale campo e, giunto in presenza del codice di chiusura @, segnalerà un errore del tipo *chiusura di campo formula mai aperto*: è evidente che questo secondo errore non esiste e, una volta corretto l'unico vero errore presente, non verrà più segnalato.

In particolari casi una situazione come quella descritta può dar luogo a sequenze anche molto lunghe di 'errori fantasma'. Ad esempio immaginiamo un testo in cui in una certa riga si apra un campo formula e in ciascuna delle righe successive si chiuda il campo formula aperto nella riga precedente e se ne apra uno nuovo. Una situazione cioè del tipo (presente ad esempio in certi antichi documenti contabili):

```
...<testo con errore>...#...<campo formula 1>  
...@...#... <campo formula 2>  
...@...#... <campo formula 3>  
...@...#... <campo formula 4>  
...@...  
...
```

L'interprete incontra l'errore e interrompe l'analisi della riga; in questo modo non incontra il codice # di inizio campo formula 1: giunto al codice @ (fine del campo formula 1) segnala l'errore *chiusura di campo formula mai aperto*, interrompe l'analisi della nuova riga, non intercetta il nuovo codice # (ponendo le basi per una nuova segnalazione) e passa alla riga successiva. La cosa si ripete per tutte le righe, dando luogo in tutto a 5 segnalazioni: la prima relativa all'errore realmente presente nel testo, le altre 4 (*chiusura di campo formula mai aperto*) fittizie.

Si possono avere in questo modo lunghe liste di errori non effettivi ma semplicemente *derivati* da una situazione di errore precedente. Per evitare l'ingenerarsi di tali liste si può utilizzare l'opzione **limita avvisi**, attiva per default, che limita il numero di segnalazioni di errore dello stesso tipo.

A dispetto di quello che può sembrare da quanto detto finora, questo meccanismo di interruzione del controllo di correttezza di una riga di testo in presenza di un errore riduce notevolmente le false segnalazioni che si avrebbero se l'analisi proseguisse.

Per semplificare l'analisi degli errori, basta attenersi a questi semplici criteri:

- in presenza di una o più segnalazioni di errore non comprensibili, sforzarsi di rimuovere almeno la prima (che potrebbe avere indotto le altre) e riprovare,
- una volta corretti tutti gli errori segnalati, ripetere comunque il controllo (che potrebbe evidenziarne altri), finché non viene esplicitamente dichiarata l'assenza di errori nel *file testo*.

Opzione 'uso memoria'

Nei limiti della memoria fisica disponibile nel computer questa opzione accelera notevolmente l'inserimento di un testo in un corpus.

Se però le dimensioni del file testo fossero tali da eccedere quelle della memoria fisica disponibile l'attivazione di **uso memoria** costringerà il sistema operativo ad effettuare una serie di operazioni di trasferimento dei dati in eccesso dalla memoria al disco fisso e di recupero dal disco fisso alla memoria, introducendo rallentamenti via via più consistenti nell'operazione e rendendo controindicato il ricorso a questa opzione. Non è possibile fissare dei criteri assoluti per decidere a priori se la memoria fisica disponibile è sufficiente o meno per un certo testo; molto a grandi linee si può dire che una memoria centrale di 64 megabyte può arrivare a contenere un testo dell'ordine di 500.000 parole: per dimensioni maggiori della memoria anche quelle del testo aumentano in proporzione. Ci si può rendere conto dell'avvenuto superamento dei limiti di memoria osservando l'avanzamento del numero della pagina analizzata: lo

scorrimento, prima a velocità più o meno costante, procede a scatti per gruppi di pagine, con intervalli di attesa sempre più lunghi tra un gruppo di pagine e il successivo. In un caso di questo genere può convenire effettuare eventuali successivi reinserimenti del testo nel corpus disabilitando l'opzione **uso memoria**: in questo modo l'operazione verrà effettuata appoggiando i dati intermedi su disco fisso, procedendo a velocità inferiore rispetto a quella iniziale del caso precedente, ma costante e sostanzialmente indipendente dalle dimensioni del testo.

Quanto detto si applica alla funzione **inserimento testi** e non al **controllo correttezza**, durante il quale non vengono generati dati da appoggiare in memoria centrale o su disco.

Opzioni 'gruppo di testi', 'compatta' e 'chiudi GATTO'

Volendo effettuare l'inserimento contemporaneo di più testi nel corpus, anziché ripetere l'operazione per ogni singolo testo, si può scrivere in un file ANSI l'elenco delle sigle dei testi coinvolti, una per riga senza segni di punteggiatura di alcun genere, e attivare l'opzione **gruppo di testi**. Così facendo non verrà chiesta la sigla del singolo testo da inserire ma il nome del file contenente la lista dei testi sui quali operare.

Attivando questa opzione verranno anche offerte le possibilità di impostare le operazioni di **compattazione del corpus** e di **uscita da GATTO** da eseguire automaticamente al termine dell'inserimento dei testi.

Nota: GATTO effettua una serie di azioni alla fine dell'operazione **inserimento testi**, indipendentemente dal fatto che questa sia stata effettuata su un testo singolo o su una lista di testi. Quando si hanno più testi da introdurre, per motivi di velocità conviene fare ricorso all'opzione **gruppo di testi**, onde evitare che lo stesso pacchetto di operazioni finali venga ripetuto per ognuno dei testi inseriti.

INSERIMENTO DI TESTI ASSOCIATI DI TIPO NOTE O TRAD

Se il testo che si vuole inserire nel corpus possiede un testo associato, anche per questo va effettuata l'operazione di inserimento. In tutti i casi l'inserimento del testo base deve essere precedente o contemporaneo a quello del testo associato.

Inserimento del testo base *seguito* dall'inserimento del testo associato

1. Introdurre nel corpus il testo base tramite la funzione **inserimento testi**,
2. introdurre nel corpus il testo associato mediante la funzione **inserim./sostituz. testi associati**, scegliendo il tipo di testo (Note o Trad); se vanno inseriti entrambi i tipi di testo associato effettuare due volte l'operazione, nell'ordine che si vuole.

Inserimento del testo base *contemporaneamente* all'inserimento del testo associato

Introdurre nel corpus il testo base con **inserimento testi** attivando l'opzione **ins. testi associati**.

Anche i testi associati prevedono una funzione di controllo correttezza, che può essere opportuno eseguire prima di inserire i testi associati nel corpus. In questo caso, poiché il controllo di correttezza dei testi associati può essere eseguito solo dopo che il testo base è stato inserito nel corpus, l'operazione di inserimento del testo base e dei testi associati andrà necessariamente eseguita in due tempi.

INSERIMENTO DI TESTI LEMMATIZZATI

L'inserimento di un testo discusso in questo capitolo assume che il testo stesso non sia stato lemmatizzato. In presenza di una lemmatizzazione precedentemente effettuata con GATTO questa operazione si configura in quella più generale di estrazione da un corpus e successivo reinserimento di un testo lemmatizzato, trattata nella parte V.

Si può però anticipare che l'inserimento di un testo lemmatizzato ha luogo con la stessa modalità di uno non lemmatizzato, salvo l'avvertenza di collocare preventivamente nella directory dei testi oltre al file `filgat.<sigla>.txt` contenente il testo anche il file `codlemmi.<sigla>` contenente la lemmatizzazione.

Attenzione

Una volta inserito un testo nel corpus non tentare di modificarne in alcun modo la sigla (nei nomi dei file, in bibliografia,...), in quanto si potrebbe danneggiare il corpus. Volendo eseguire questa operazione, vedere all'inizio del capitolo 25.

LEZIONE 13. COME INSERIRE UN TESTO IN UN CORPUS.

Si procederà ora ad inserire nel corpus **C:\prova** costruito nel corso della Lezione 11 i testi **es1** e **es2**, i cui file testo **filgat.es1.txt** e **filgat.es2.txt** si trovano nella cartella **esempi** del CD di GATTO o sono stati scaricati via Web insieme al manuale.

Preliminarmente i due file testo andranno copiati nella directory dei testi del corpus **Prova**, ovvero nella cartella **C:\gattesti**.

Inserimento del testo es1 nel corpus Prova

Per partire, avviare GATTO e portarsi nell'ambiente **Gestione base dati** mediante [**GEST. BASE DATI**].

Tramite [**CORPUS**] aprire la finestra APERTURA E GESTIONE GENERALE CORPUS; cliccare sulla riga del corpus **Prova**, aprirlo con [**APRI**], quindi tornare alla finestra GESTIONE BASE DATI con [**CHIUDI FINESTRA**].

Per prima cosa verrà effettuato il controllo di correttezza sintattica del contenuto del file **filgat.es1.txt**. A questo scopo usare [**TESTI.. + CONTROLLO CORRETTEZZA + TESTI**]; lasciare le opzioni impostate (**testo singolo + limita avvisi**) e cliccare su (**OK**). La lista dei *testi da controllare* dovrebbe offrire **es1** e **es2**: cliccare su **es1** e premere (**OK**). Il controllo di correttezza viene eseguito e, al termine, appare il messaggio **nessun errore localizzato nel testo es1**; accettare con (**OK**).

A questo punto si può procedere senz'altro con l'inserimento del testo **es1** nel corpus **Prova**.

Usare [**TESTI.. + INSERIMENTO TESTI**] e avviare l'operazione con (**OK**). Come prima scegliere **es1** e premere (**OK**). L'inserimento del testo nel corpus ha luogo, mostrando prima il procedere del controllo di correttezza tramite il numero di pagina e poi l'inserimento del testo tramite la lettera iniziale delle parole via via inserite negli archivi con criterio alfabetico (poiché **es1** è molto piccolo queste indicazioni potrebbero essere troppo rapide per essere viste chiaramente). L'operazione termina col messaggio **inserito nel corpus il testo es1**; accettare con (**OK**).

L'inserimento del testo nel corpus è completato. Per verificarlo, con [**CORPUS**] e [**INFORMAZIONI.. + STATISTICHE**] guardare il nuovo contenuto del corpus:

- testi 1
- occorrenze 126
- forme diverse 98

Chiudere questa finestra con <Esc> o cliccando su **x**. Volendo, uscire dall'ambiente e da GATTO nel solito modo.

Inserimento del testo es2 nel corpus Prova

Come nel caso precedente avviare GATTO e portarsi nell'ambiente **Gestione base dati** mediante [**GEST. BASE DATI**], poi tramite [**CORPUS**] aprire la finestra APERTURA E GESTIONE GENERALE DEI CORPUS, cliccare sulla riga del corpus **Prova**, aprirlo con [**APRI**], quindi tornare alla finestra GESTIONE BASE DATI con [**CHIUDI FINESTRA**].

Al solito, conviene effettuare il controllo di correttezza sintattica del contenuto del file **filgat.es2.txt** mediante [**TESTI.. + CONTROLLO CORRETTEZZA + TESTI**] e (**OK**). La lista dei *testi da controllare* dovrebbe offrire **es1** e **es2**: cliccare su **es2** e su (**OK**). A differenza del caso precedente viene rilevata la presenza di un errore in **filgat.es2.txt**. Il messaggio finale avverte che il file contenente la spiegazione dell'errore è il file **c:\prova.gat\out\err.es2.txt**. Questo file, di nome generico **err.<sigla>.txt**, è sempre collocato nella sottocartella **out** del corpus. È possibile stampare immediatamente il file o decidere di farlo in seguito; per averne una stampa immediata rispondere (**Si**).

Il messaggio contenuto nel file indica che vi è un carattere non ammesso all'interno di una forma, per l'esattezza il carattere "9" (codice ANSI 57). Il carattere errato si trova nella riga 4 della pagina 129 del testo: più esattamente è il 25° carattere di quella riga; un altro modo per localizzare l'errore è fornito dall'indicazione che esso si trova nella 10ª linea del file **filgat.es2.txt**.

Disponendo di questi dati, aprire il file **c:\gattesti\filgat.es2.txt** ad esempio con **Notepad** o **Wordpad**, poi portarsi alla pagina 129:

\$0129\$ *ço ch'io dei' scoder...*

e da lì raggiungere la 4ª riga della pagina:

che se de' scoder in Rom9ania de quello che no nde sé ni

è facile vedere che la parola *Rom9ania* contiene per l'appunto il carattere "9" evidentemente digitato per errore; tale carattere è il n. 25 nella riga. Si può anche vedere che la linea in questione è la 10ª del file. Rimuovere il carattere "9", salvare il file filgat e riprovare il controllo di correttezza che questa volta indicherà l'assenza di ulteriori errori.

Inserire il testo nel corpus tramite [TESTI.. + INSERIMENTO TESTI] e (OK). Scegliere **es2** e premere ancora (OK). Il messaggio finale comunica l'avvenuto inserimento; rispondere (OK). Volendo verificare le modifiche apportate alla struttura dei dati inclusi nel corpus, con [CORPUS] e [INFORMAZIONI.. + STATISTICHE] accedere alle statistiche di quest'ultimo, che risulteranno ora:

- testi 2
- occorrenze 1449
- forme diverse 502

LEZIONE 14. COME INSERIRE NEL CORPUS I TESTI ASSOCIATI, CONTENENTI NOTE, DIVERSE EDIZIONI O TRADUZIONI.

Nel caso che il testo da inserire in un corpus preveda la presenza di testi associati contenenti note o traduzioni, l'ordine delle operazioni da eseguire è il seguente:

- 1°. controllo di correttezza del testo base (facoltativo)
- 2°. inserimento del testo base
- 3°. controllo di correttezza del testo associato (facoltativo)
- 4°. inserimento del testo associato

I controlli di correttezza sono facoltativi perché vengono comunque eseguiti preliminarmente nella fase di inserimento di un testo nel corpus.

Come applicazione di quanto detto si procederà ora ad inserire nel corpus **Prova** il testo **es3**, cui sono associati sia il file di tipo Note che quello di tipo Trad.

Come primo passo è quindi necessario copiare i file **filgat.es3.txt**, **note.es3.txt**, **trad.es3.txt**, contenuti nella cartella **esempi** del CD di GATTO o scaricati via Web insieme al manuale, nella cartella **C:\gattesti**.

Per le operazioni di controllo correttezza e inserimento del testo base **es3** si rimanda alla Lezione precedente. Qui ci si soffermerà sulle fasi successive; per controllo, al termine dell'inserimento del solo testo base **filgat.es3.txt** il corpus **Prova** risulterà così composto:

- testi 3
- occorrenze 1921
- forme diverse 677

Inserimento del testo associato note.es3.txt nel corpus Prova

Una volta inserito il testo base, dalla finestra GESTIONE BASE DATI selezionare [TESTI.. + CONTROLLO CORRETTEZZA + TESTI ASSOCIATI NOTE], procedere con (OK), selezionare **es3** come *testo da controllare* e ancora (OK).

Viene segnalata la presenza di due errori, descritti nel file **c:\prova.gat\out\err.es3.txt**, stampando il quale si può vedere che gli errori segnalati sono:

- *duplicata la nota n. 2*
- *manca nel testo associato la nota n. 3.*

Aperto il file **c:\gattesti\note.es3.txt** con **Notepad** o **Wordpad** si può notare che in effetti contiene la nota numero 1, due note entrambe denotate dal numero 2 e la nota numero 4; manca la numero 3. Se si apre con un editore il file **filgat.es3.txt** si può vedere che lì sono invece previste le quattro note designate dai numeri tra 1 e 4; evidentemente in **c:\gattesti\note.es3.txt** la terza nota è stata indicata col numero 2 anziché col 3; per un riscontro è sufficiente controllare il testo corrispondente a tali note all'interno di **filgat.es3.txt**.

Per proseguire sostituire con un "3" il "2" posto all'inizio della terza nota nel file associato, chiudere quest'ultimo e riprovare il controllo di correttezza: questa volta verrà segnalata l'assenza di errori.

Per effettuare l'inserimento del file Note, selezionare [TESTI.. + INSERIM./SOSTITUZ. TESTI ASSOCIATI + TIPO NOTE] e (OK). Scegliere **es3** e premere ancora (OK). Il messaggio finale comunica l'avvenuto inserimento; rispondere (OK).

Inserimento del testo associato trad.es3.txt nel corpus Prova

Dalla finestra GESTIONE BASE DATI selezionare [TESTI.. + CONTROLLO CORRETTEZZA + TESTI ASSOCIATI TRAD], procedere con (OK), selezionare **es3** come *testo da controllare* e ancora (OK).

Anche in questo caso si ottiene una segnalazione di errore: *numero di brani diverso in testo base e associato: 13 contro 12.*

Questa situazione è alquanto critica poiché GATTO può segnalare il disallineamento tra i due testi causato da ripartizioni in numeri diversi di brani ma non è ovviamente in grado di localizzare più precisamente l'errore al loro interno. Se i due testi sono di dimensioni rilevanti e sono stati ripartiti in un numero elevato di brani l'individuazione degli errori può risultare complicata anche per l'utente. Per agevolare quest'ultimo GATTO offre una funzione che permette di visualizzare in parallelo i brani estratti dai due testi, facilitando la localizzazione dei punti dove la corrispondenza si interrompe.

Per attivare questa funzione, dalla finestra GESTIONE BASE DATI selezionare [TESTI.. + CONTROLLO CORRETTEZZA + ALLINEAMENTO TESTI BASE - TRAD], selezionare **es3** come *testo da controllare* e (OK).

Appare una coppia di finestre: quella più in alto è relativa al testo base, l'altra al testo associato. Entrambe (con un limite di 1000 caratteri) visualizzano il primo brano presente nel testo, ovvero quella parte di file che va dall'inizio fino al primo asterisco (ricordarsi che il carattere * funge da separatore tra brani successivi). Il riquadro in basso con fondo bianco evidenzia che il brano mostrato è il numero 1, mentre quello con fondo grigio alla sua sinistra ricorda che i brani presenti vanno dal numero 1 al numero 12. In effetti il messaggio di errore ha indicato che i due testi risultano suddivisi rispettivamente in 13 e 12 brani e nel riquadro grigio viene considerato il testo con numero minore di brani.

Trattandosi dei due brani iniziali, quelli mostrati ovviamente si corrispondono. Per localizzare il punto in cui la corrispondenza tra i due testi viene meno si possono visualizzare successivamente i brani numero 2, 3, ecc., ma si può più efficacemente utilizzare la tecnica detta di *bisezione*, andando a verificare l'allineamento nella zona centrale del testo, poi portandosi nel punto centrale della parte precedente se il disallineamento è già presente lì o nel punto centrale della parte successiva nel caso opposto, ripetendo poi questo procedimento che dimezza la zona indagata ad ogni passaggio, fino a localizzare il punto esatto in cui si trova l'errore.

Nel caso in oggetto converrà portarsi sul brano n. 6 (metà di 12), sostituendo "1" con "6" nel riquadro a fondo bianco in basso e poi cliccando su (**Mostra**): osservando i nuovi contenuti delle due finestre si vede che il secondo brano non è la traduzione del primo, dunque l'errore è precedente, cioè in uno dei primi 5 brani; conviene quindi portarsi in corrispondenza del nuovo punto centrale, cioè del brano numero 3 (sempre usando (**Mostra**)). Qui notiamo che la seconda finestra è vuota ma, se guardiamo i due file, vediamo che **trad.es3.txt** contiene in corrispondenza del 3° brano solo uno spazio, tra i due asterischi che lo delimitano, a significare che non esiste nel testo Trad la traduzione del 3° brano: è cioè un effetto voluto, quindi il 3° brano è a posto e l'errore non può essere che successivo, cioè nel 4° o nel 5°. Passando al brano numero 4 si vede che nella finestra inferiore si fa riferimento ai due fratelli *Eson* e *Pelias* che non compaiono in quella superiore: per localizzarli nel testo base bisogna passare al brano successivo, il numero 5. L'errore è stato trovato e consiste nella mancanza di un asterisco, nel testo associato, tra *...et Thosone aureo.* e *Eo tempore fuerunt..* (infatti queste due stringhe, entrambe presenti nel brano n. 4 del testo Trad, fanno riferimento rispettivamente ai brani n. 4 e n. 5 del testo base).

Uscire quindi dalla finestra con (**Chiudi**), procedere alla correzione in **trad.es3.txt** e riprovare il controllo di correttezza del testo Trad: questa volta verrà evidenziata l'assenza di errori.

Procedere infine all'inserimento del file Trad selezionando [TESTI.. + INSERIM./SOSTITUZ. TESTI ASSOCIATI + TIPO TRAD] e (OK). Scegliere **es3** e premere ancora (OK). Il messaggio finale comunica l'avvenuto inserimento; rispondere (OK).

PARTE III - LEMMATIZZAZIONE E INSERIMENTO DI IPERLEMMI

CAP. 9 - CONCETTO E MECCANISMI DI LEMMATIZZAZIONE. FUSIONE DI PIÙ LEMMATIZZAZIONI

I LEMMI IN GATTO

Nella parte introduttiva del manuale sono stati illustrati concetto e struttura dei lemmi in GATTO ed è stata fatta una veloce panoramica sulle procedure di lemmatizzazione. Per chiarezza alcuni di questi concetti verranno ora ripresi, inserendoli in una descrizione più sistematica delle procedure di lemmatizzazione previste all'interno del programma.

Innanzitutto per *lemma* si intende qui l'elemento cui viene riportato un insieme di forme che si distinguono fra loro soltanto per l'assetto grafico (varianti grafiche, con o senza valore di varianti fonetiche) e/o perché sono forme della flessione dello stesso verbo, sostantivo, aggettivo. In GATTO ogni lemma è memorizzato in forma di un gruppo di 4 stringhe:

- entrata lessicale
- categoria grammaticale
- disambiguatore
- commento.

Nel definire un nuovo lemma i primi due elementi vanno definiti obbligatoriamente, mentre l'introduzione degli altri è facoltativa.

L'utilità del *disambiguatore* nasce dal fatto che in GATTO due lemmi sono riconosciuti come distinti se differiscono per l'entrata lessicale e/o per la categoria grammaticale: nel caso di due lemmi diversi ma caratterizzati da uguaglianza di entrambi questi elementi è possibile introdurre nel campo *disambiguatore* di uno di essi o di entrambi stringhe arbitrarie finalizzate a distinguerli. Riassumendo, perché un nuovo lemma risulti diverso e distinto da quelli già presenti nel corpus, deve differire da ognuno di essi per l'entrata lessicale, la categoria grammaticale o il disambiguatore.

Il *commento* consente di inserire ulteriori elementi descrittivi del lemma, fermo restando che il contenuto di questo campo può essere visualizzato ma non contribuisce a definire il lemma né a distinguerlo dagli altri.

Nel definire un lemma non viene fatta alcuna distinzione tra lettere maiuscole e minuscole. I caratteri che possono essere inclusi in un lemma sono:

- le 26 lettere dell'alfabeto inglese, con tutte le accentazioni, dieresi, ecc. previste dalla tabella ANSI
- il simbolo ` (ANSI 96)
- l'apice '
- il carattere spazio
- i caratteri . () - = ? []
- i caratteri 0, 1, ..., 9 (se il corpus è impostato per accettare anche le cifre arabe all'interno dei lemmi)

Notare che il ricorso ai caratteri [] e ? presenta una limitazione: infatti questi caratteri, se digitati all'interno dei lemmi da cercare, vengono automaticamente considerati come *caratteri di controllo* e non come parte del lemma.

L'inserimento di un lemma in un corpus ha luogo tramite il procedimento di *lemmatizzazione*. L'insieme dei lemmi introdotti in un corpus ne costituisce il *lemmario*. L'insieme dei lemmi associati ad una certa forma del corpus ne costituisce il *dizionario di macchina*. L'insieme dei lemmi caratterizzati dalla stessa entrata lessicale di quello digitato in fase di lemmatizzazione forma il contenuto della *tabella degli omografi*. Tutti i 4 dati costituenti un lemma possono essere modificati in un secondo tempo, per correggere eventuali errori, mentre l'inserimento e la cancellazione di lemmi dal lemmario sono operazioni normalmente controllate in maniera automatica da GATTO; in particolare l'inserimento ha luogo al momento della prima lemmatizzazione effettuata nell'ambito del corpus la quale faccia uso del nuovo lemma, mentre la rimozione dal lemmario ha luogo quando viene rimossa dal corpus l'ultima lemmatizzazione che faccia uso del lemma in questione⁴.

⁴ A quest'ultima regola fa eccezione la formazione dei lemmi *muti*.

Un altro meccanismo di inserimento di lemmi nel lemmario è legato all'acquisizione al corpus di testi già lemmatizzati in precedenza (magari estratti da un altro corpus): in questo caso ogni testo porta con sé i propri lemmi e questi vengono automaticamente inglobati nel lemmario, qualora non già presenti.

PROCEDURA DI LEMMATIZZAZIONE IN GATTO

All'interno di GATTO utilizzando i due ambienti appositi è possibile procedere alla lemmatizzazione di un corpus, selezionando di volta in volta le occorrenze sulle quali agire; questa selezione ha luogo sulla base delle forme di cui le occorrenze sono espressione e dei testi in cui sono inserite. In GATTO *lemmatizzare* significa associare un certo lemma a una occorrenza specifica di una forma specifica in un testo specifico. Se il lemma utilizzato non era già presente nel lemmario del corpus vi viene aggiunto automaticamente. Nei riguardi della lemmatizzazione una forma polirematica opera come un tutt'uno, ovvero all'insieme delle parole componenti viene associato un lemma unico.

Una volta portata a termine, una lemmatizzazione ha una duplice conseguenza: realizza una connessione tra il lemma e la specifica occorrenza cui è stato associato ma in più crea una connessione a livello di corpus tra il lemma e la forma di cui l'occorrenza era espressione. Grazie a questa connessione una ricerca fatta a partire dal lemma permetterà di risalire a tutte le forme associate ad esso in uno o più testi del corpus (se però la ricerca è limitata ad un sottocorpus verranno prese in considerazione solo le associazioni forma - lemma derivanti da lemmatizzazioni effettuate all'interno del sottocorpus considerato), fornendone sia le occorrenze lemmatizzate col lemma in questione che, volendo, quelle non lemmatizzate o lemmatizzate con altri lemmi.

Gli ambienti di lemmatizzazione mettono a disposizione diverse tecniche di scelta del lemma da utilizzare:

- digitazione ex-novo (a prescindere dal fatto che il lemma esista o meno nel lemmario);
- selezione dal dizionario di macchina della forma;
- selezione dalla tabella degli omografi;
- prelievo dalla casella *lemma impostato*;
- copia da un'altra occorrenza già lemmatizzata col lemma voluto;
- generazione automatica nel caso di antropónimo o nome geografico.

Non è possibile associare più di un lemma ad una stessa occorrenza.

Nel corso della lemmatizzazione di una occorrenza è possibile associarle anche un *iperlemma*.

DIZIONARIO DI MACCHINA

Con tale termine si denota l'insieme dei lemmi associati ad una forma nell'ambito di un corpus. Questo dizionario è gestito in maniera del tutto automatica: ogni volta che lemmatizzando viene utilizzato un lemma fino a quel momento non associato, in quel corpus, alla forma considerata, il lemma viene aggiunto automaticamente al dizionario di macchina della forma e come tale verrà proposto all'utente, come possibile scelta, nel corso delle successive lemmatizzazioni riferite a quella forma, in qualunque testo del corpus.

Ogni voce che compone il dizionario di macchina, ovvero ogni associazione forma - lemma, nasce dunque al momento in cui il lemma viene associato alla forma per la prima volta nel corpus. La voce viene tolta dal dizionario di macchina, sempre automaticamente, al momento in cui l'ultimo utilizzo di tale associazione risulti rimosso dal corpus, per modifica di una lemmatizzazione o per cancellazione di un testo dal corpus.

L'utilità del dizionario di macchina deriva dal fatto che la lista dei lemmi proposti agevola la scelta da fare e che per lemmatizzare un'occorrenza è sufficiente cliccare sulla riga del dizionario che riporta il lemma voluto.

TABELLA DEGLI OMOGRAFI

Quando si intende lemmatizzare una certa forma con un particolare lemma sono possibili 3 situazioni alternative:

1. il lemma è già presente nel lemmario ed è già associato alla forma in oggetto,
2. il lemma è già presente nel lemmario ma non in associazione alla forma in oggetto,
3. il lemma non è presente nel lemmario.

Nel primo caso il lemma è mostrato automaticamente nel dizionario di macchina, a differenza degli altri due.

Se il lemma voluto non compare nel dizionario di macchina l'unico modo per usarlo è quello di digitarlo come nuovo lemma; a questo punto GATTO ricerca tra i lemmi del lemmario eventuali omografi di quello digitato. Se ve ne sono (caso 2), il dizionario di macchina viene momentaneamente sostituito da un'altra tabella, la *tabella degli omografi*, che mostra tutti i lemmi presenti nel lemmario ed aventi entrata lessicale identica a quella digitata; se il lemma voluto è incluso tra quelli mostrati per lemmatizzare basterà selezionarlo. Se invece il lemma desiderato non compare tra gli omografi suggeriti o se nessun omografo è presente (caso 3) si dovrà completare l'inserimento dei dati del nuovo lemma (categoria grammaticale ecc.), che verrà così aggiunto al lemmario.

IPERLEMMI

Gli *iperlemmi* costituiscono uno strumento che amplia la possibilità di localizzare occorrenze di gruppi di forme interconnesse dal punto di vista semantico, grammaticale o di altra natura, al di là di quanto consentito dal ricorso ai lemmi. Gli iperlemmi sono in sostanza categorie che possono essere attribuite ai lemmi: tramite gli iperlemmi più lemmi possono essere associati tra loro e in seguito ricercati simultaneamente, senza necessità di essere specificati singolarmente. Uno stesso lemma può essere associato a più iperlemmi e cercato sulla base di uno qualunque di essi: ad esempio il lemma **mela** può essere associato agli iperlemmi **frutta** e **cibo**. Dall'esempio fatto si evidenzia che, all'interno di un corpus, possono liberamente coesistere lemmi e iperlemmi caratterizzati dalla stessa entrata lessicale (ad esempio **cibo**).

Gli iperlemmi di cui si è parlato finora sono indicati in GATTO, per l'esattezza, come *iperlemmi di livello 1*. Il meccanismo con cui questi elementi consentono di raggruppare i lemmi può essere amplificato, raccogliendo tra loro *iperlemmi di livello 1* tramite associazione a uno o più *iperlemmi di livello 2*, in maniera del tutto analoga a quanto fatto con i lemmi. In generale, *iperlemmi di livello n* possono essere associati a *iperlemmi di livello n+1*, purché esista una catena di associazioni che, percorsa a ritroso a partire dall'iperlemma di livello n+1, porti fino a un lemma attraverso i livelli n, n-1, ..., 1; per coerenza di definizione da questo punto di vista un lemma potrebbe anche essere definito un *iperlemma di livello 0*.

Le forme (e relative occorrenze) potranno così essere cercate a partire da iperlemmi di qualunque livello, tramite la loro associazione a lemmi concatenati con gli iperlemmi indicati in avvio di ricerca.

Una ambiguità insita in questo meccanismo concerne la lemmatizzazione di occorrenze di una certa forma tramite un lemma al quale siano stati associati più iperlemmi di livello 1 (tra i quali, ma non necessariamente, uno potrebbe costituire una categoria prevalente): l'occorrenza, lemmatizzata col lemma in oggetto, è da mettere in relazione o meno con uno di questi iperlemmi, e se sì con quale?

Per risolvere questa ambiguità GATTO permette, in fase di lemmatizzazione di un'occorrenza, di associarle non solo un lemma ma, volendo, anche un iperlemma (di livello 1, già associato o meno al lemma, presente o meno nell'iperlemmario); sia l'iperlemma che il suo collegamento al lemma, qualora non fossero presenti nel corpus, verrebbero creati automaticamente.

L'insieme degli iperlemmi definiti in un corpus ne costituisce l'*iperlemmario*.

In GATTO ogni iperlemma è memorizzato in forma di un gruppo di 3 dati:

- entrata lessicale
- livello
- disambiguatore.

Nel definire un nuovo iperlemma i primi due elementi vanno indicati obbligatoriamente, mentre l'introduzione del terzo, il cui scopo è di discriminare due eventuali iperlemmi caratterizzati da uguale entrata lessicale e stesso livello, è facoltativa. Non esistono vincoli particolari circa i caratteri che possono far parte dell'entrata lessicale di un iperlemma.

A differenza di quanto avviene per i lemmi, la gestione degli iperlemmi è sostanzialmente manuale. Il loro inserimento nell'iperlemmario e la loro associazione a lemmi o ad altri iperlemmi, così come le operazioni di modifica o di cancellazione dall'iperlemmario, richiedono azioni esplicite da parte dell'utente; fanno eccezione tre casi:

1. inserimento automatico nell'iperlemmario di un nuovo iperlemma o di una nuova associazione lemma – iperlemma di livello 1 a seguito di una lemmatizzazione con uso di un (nuovo) iperlemma (ved. sopra),
2. inserimento automatico di iperlemmi già inclusi in un nuovo testo aggiunto nel corpus,

3. cancellazione automatica, in cascata a risalire, di tutti gli iperlemmi che risultassero privi di connessioni a livello inferiore, a seguito della cancellazione di un lemma o di un iperlemma *sottostante*.

Per questi motivi, la trattazione delle operazioni di gestione degli iperlemmi sono descritte nella parte V – Operazioni di gestione dei dati (capitolo 27), con esclusione dell'inserimento di nuovi iperlemmi, della loro associazione ai lemmi e alle specifiche occorrenze in fase di lemmatizzazione, inserite in questa parte per comodità dell'utente, essendo strettamente connesse alle operazioni di lemmatizzazione (capitolo 12).

LEMMA IMPOSTATO

È il lemma che è stato utilizzato nell'ultima operazione di lemmatizzazione o che comunque è stato copiato nel riquadro celeste nella testata delle finestre di lemmatizzazione. Di rapida assegnazione ai contesti, permette di accelerare le operazioni di lemmatizzazione quando queste fanno uso ripetuto di uno stesso lemma. Il *lemma impostato* viene anche riassegnato ogni volta che il dizionario di macchina viene aperto o viene riferito ad una forma diversa, sempre che il dizionario di macchina di quest'ultima non sia vuoto. Se il lemma corrispondente al *lemma impostato* viene rimosso dal dizionario di macchina a seguito della rimozione o sostituzione di una lemmatizzazione, anche il *lemma impostato* viene rimosso o sostituito. Se nel corso dell'ultima lemmatizzazione è stato associato all'occorrenza anche un iperlemma, il *lemma impostato* è in realtà costituito dalla coppia lemma – iperlemma usata.

FINESTRA A CONTESTO SINGOLO

Il contesto riportato nella finestra *a contesto singolo* è definito *a periodi*, ovvero è costituito, inizialmente, dal periodo contenente l'occorrenza da lemmatizzare più i periodi immediatamente precedente e successivo (se presenti all'interno della parte di testo riferita allo stesso riferimento organico). La dimensione del contesto può essere variata dall'utente a partire da un minimo di un periodo (quello che include l'occorrenza in oggetto) ed è sempre costituita da un numero dispari di periodi, centrati su quello cui appartiene l'occorrenza; in ogni caso il contesto visibile è limitato dalla schermata e dai limiti del riferimento organico in cui è incluso.

FUSIONE DI PIÙ LEMMATIZZAZIONI

Indipendentemente dalle modalità utilizzate per lemmatizzare o iperlemmatizzare, che saranno descritte in dettaglio nei prossimi capitoli, GATTO permette che più persone lavorino parallelamente sugli stessi testi, consentendo una fusione a posteriori delle operazioni di lemmatizzazione eseguite. Vediamo come.

Supponiamo che un corpus contenga tra gli altri N testi (magari, ma non necessariamente, già parzialmente lemmatizzati) che si vogliono lemmatizzare velocemente, utilizzando 3 persone contemporaneamente. Si comincia con l'estrarre gli N testi dal corpus (vedere ad esempio le Lezioni 49, 51, 54, 11), copiandoli in altri 2 corpora (creati per l'occasione o contenenti già altri testi).

Fatto questo saranno presenti 3 copie degli stessi N testi, installate su 3 corpora diversi a disposizione delle 3 persone che dovranno lemmatizzare.

A questo punto ciascuno dei lemmatizzatori potrà procedere, utilizzando il proprio corpus, a lemmatizzare la propria copia degli N testi in oggetto. L'unica avvertenza da tenere ben presente è che per ogni occorrenza presente in un testo alla fine verrà accettata, come ovvio, una sola lemmatizzazione: quindi converrà definire un criterio di ripartizione del lavoro di lemmatizzazione tale da evitare che più persone lemmatizzino una stessa occorrenza di uno stesso testo.

Quando i 3 lemmatizzatori avranno finito il proprio lavoro, uno dei 3 corpora utilizzati dovrà essere considerato come 'primario': in caso di conflitto di lemmatizzazioni, cioè nel caso che una stessa occorrenza risulti lemmatizzata da più persone, GATTO utilizzerà automaticamente la lemmatizzazione presente nel corpus primario.

Per ogni testo contenuto in uno dei corpora 'secondari' il processo di fusione della lemmatizzazione si articola in 3 fasi, ovvero estrazione della lemmatizzazione dal corpus secondario, spostamento dei file così generati, riversaggio del loro contenuto nel corpus primario. Se, come nel caso dell'esempio, vi è più di un corpus secondario, l'operazione andrà ripetuta per ogni corpus secondario, preferibilmente fermo restando quello primario.

I fase: estrazione della lemmatizzazione da uno dei corpora secondari

Si opera su uno dei corpora secondari.

Dalla finestra di avvio di GATTO entrare nell'ambiente **Gestione base dati** con [GEST. BASE DATI], quindi scegliere [CORPUS]; cliccare sulla riga del corpus secondario interessato, che risulta così evidenziato e aprirlo con [APRI].

Nella finestra GESTIONE BASE DATI cliccare su [TESTI.. + ESTRAZIONE FILE TESTO + PER FUSIONE LEMMATIZZAZIONI]; mantenere l'impostazione **testo singolo** e cliccare su **(OK)**; nella lista di sigle presentata cliccare su quella relativa al testo voluto e procedere con **(OK)**; viene eseguita l'estrazione del testo e al termine un messaggio avverte che il testo estratto si trova nella sottocartella **out** del corpus, nel file **...fondilemmi.XXX.txt**, dove XXX è la sigla del testo considerato.

II fase: trasferimento dei file

Dalla sottocartella **out** del corpus secondario prelevare i file **fondilemmi.XXX** e **fondilemmi.XXX.txt** e copiarli nella directory dei testi del corpus primario.

III fase: fusione delle lemmatizzazioni

Si opera sul corpus primario.

Dalla finestra di avvio di GATTO entrare nell'ambiente **Gestione base dati** con [GEST. BASE DATI], quindi scegliere [CORPUS]; cliccare sulla riga del corpus primario, che risulta così evidenziato e aprirlo con [APRI].

Nella finestra GESTIONE BASE DATI cliccare su [TESTI.. + FUSIONE LEMMATIZZAZIONI]; mantenere l'impostazione **testo singolo** e cliccare su **(OK)**; nella lista di sigle presentata cliccare su quella relativa al testo voluto e procedere con **(OK)**; viene eseguita la fusione delle lemmatizzazioni.

Al termine, la copia del testo XXX inclusa nel corpus primario conterrà l'unione delle lemmatizzazioni che erano state effettuate nei 2 corpora coinvolti: come detto prima, se una certa occorrenza di XXX era stata lemmatizzata nel corpus primario, qualunque lemmatizzazione o iperlemmatizzazione effettuata sulla corrispondente occorrenza di XXX del testo secondario verrà totalmente ignorata, come se non fosse mai stata eseguita.

Per completare la fusione sul testo XXX basterà ripetere le 3 fasi suesposte utilizzando l'altro corpus secondario ipotizzato nell'esempio, fermo restando lo stesso corpus primario (si possono anche invertire i ruoli dei corpora, basta fare attenzione al fatto che la lemmatizzazione del corpus di arrivo, che svolge la funzione di 'primario', prevale sempre sull'altra).

Al termine, tutte le copie dei file **fondilemmi.XXX** e **fondilemmi.XXX.txt** possono essere cancellate.

La fusione delle lemmatizzazioni può essere eseguita testo per testo o per più testi simultaneamente (vedere Lezione 54).

CONTENUTO DEI CAPITOLI E DELLE LEZIONI SUCCESSIVE

I prossimi due capitoli, basandosi su una serie di esempi, forniranno una rassegna completa delle funzionalità dei moduli di lemmatizzazione inclusi in GATTO.

In particolare:

- **Capitolo 10** descrive la procedura di lemmatizzazione basata sulla ricerca delle occorrenze da lemmatizzare a partire dalla selezione delle forme (ed eventualmente dei testi) all'interno del corpus. Questo capitolo include 4 lezioni:
 - Lezione 15** rassegna dei metodi di lemmatizzazione, per il caso di lemmatizzazione con la finestra a contesti multipli;
 - Lezione 16** lemmatizzazione con la finestra a contesto singolo;
 - Lezione 17** rassegna delle opzioni di lemmatizzazione e comandi di spostamento tra le occorrenze da lemmatizzare;
 - Lezione 18** funzioni accessorie alla lemmatizzazione e comandi per acquisizione di informazioni.

- **Capitolo 11** descrive la seconda procedura di lemmatizzazione disponibile, specificamente progettata per agevolare la lemmatizzazione, eventualmente esaustiva, di singoli testi del corpus. Anche questo capitolo include 4 lezioni:
 - Lezione 19** rassegna dei metodi di lemmatizzazione, per il caso di lemmatizzazione standard a contesti multipli;
 - Lezione 20** lemmatizzazione con la finestra a contesto singolo, in modalità standard o sequenziale;
 - Lezione 21** opzioni di lemmatizzazione e rassegna dei comandi di spostamento tra le occorrenze di una stessa forma e tra forme diverse;
 - Lezione 22** funzioni accessorie alla lemmatizzazione e comandi per acquisizione di informazioni.

Tra i contenuti dei due pacchetti di 4 Lezioni vi è un certo parallelismo, inserito al fine di agevolare la comprensione delle differenze ma anche delle corrispondenze funzionali che vi sono tra i due meccanismi di lemmatizzazione.

CAP. 10 - LEMMATIZZAZIONE SUL CORPUS

PROCEDURA DI LEMMATIZZAZIONE

Per avviare la lemmatizzazione è necessario prima di tutto entrare nella finestra introduttiva dell'ambiente **Lemmatizzazione sul corpus**, tramite [LEMMATIZZAZIONE + SUL CORPUS] nella finestra iniziale di GATTO, provvedendo poi ad aprire il corpus desiderato.

La lemmatizzazione prevede come prima fase l'individuazione delle forme le cui occorrenze, tutte o in parte, si intendono lemmatizzare; l'operazione viene estesa per default all'intero corpus ma, volendo, può essere limitata ad un sottoinsieme dei testi costituenti (sottocorpus). Le forme coinvolte possono essere selezionate sulla base del loro assetto grafico.

In pratica, viene effettuata una operazione di ricerca in tutto e per tutto simile a quelle descritte nella parte IV del manuale, che tratta appunto dell'ambiente **Ricerche**, salvo alcune limitazioni relative alle tipologie di ricerche effettuabili e alle opzioni disponibili. Per questo motivo la fase di ricerca verrà qui descritta in maniera limitata, rinviando invece alla parte successiva del manuale per una trattazione completa. C'è solo da aggiungere che la ricerca delle forme da lemmatizzare può essere impostata non solo sulla base dell'assetto grafico ma anche degli eventuali lemmi già associati, dal momento che le operazioni di lemmatizzazione consentono anche di modificare o rimuovere le lemmatizzazioni preesistenti.

Una volta che le occorrenze da lemmatizzare sono state localizzate, si possono copiare tutte insieme, con i relativi contesti, nella finestra di lemmatizzazione, oppure se ne può copiare solo una parte. A questo punto si può procedere alla effettiva lemmatizzazione dei contesti copiati, con lemmatizzazioni successive per singole occorrenze o per gruppi; è anche possibile lemmatizzare con un'unica operazione tutti i contesti presenti.

Al termine, si può tornare alla finestra che contiene i risultati della ricerca precedente, prelevare un altro pacchetto di contesti e procedere a lemmatizzare questi ultimi, oppure si può ripartire effettuando una nuova ricerca o, infine, uscire dall'ambiente **Lemmatizzazione sul corpus**.

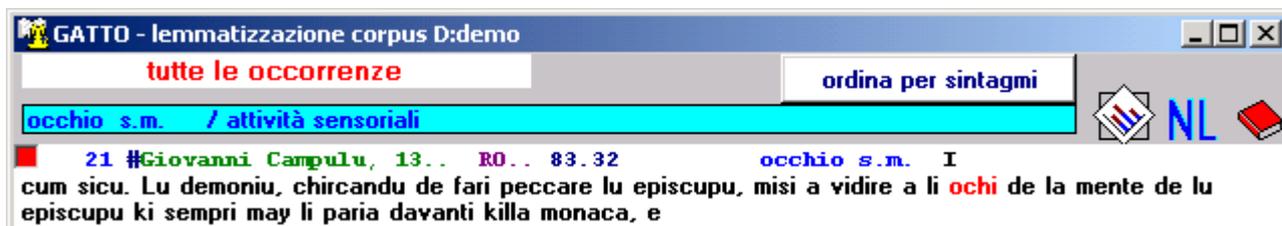
LEMMATIZZAZIONE DEL CONTESTO CORRENTE O DEI CONTESTI SELEZIONATI

Quando si entra nella finestra di lemmatizzazione (a contesti multipli), ovvero quella che visualizza tra i contesti ottenuti come risultato della ricerca quelli scelti per essere lemmatizzati, il primo contesto è marcato con un piccolo riquadro rosso, posto sulla sinistra, che lo individua come *contesto corrente*, cioè quello cui si applicheranno le prossime azioni di lemmatizzazione; si può procedere a lemmatizzare il *contesto corrente*, passando poi al contesto successivo o ad un altro, oppure si può definire un qualunque altro contesto come *contesto corrente* e procedere a lemmatizzare quest'ultimo, e così via. Naturalmente ad ogni lemmatizzazione potrà essere scelto un lemma diverso. Esistono meccanismi che permettono di accelerare le operazioni qualora si intenda applicare lo stesso lemma a più contesti; in tutti i casi l'operazione avverrà però per singoli contesti.

Se i contesti in gioco sono numerosi o se gli stessi lemmi sono applicabili a più contesti conviene utilizzare una seconda tecnica, che consiste nel selezionare tutti i contesti che si intende lemmatizzare con un

certo lemma e poi, dopo aver attivato il flag **lemmatizza selezionati**, nell'assegnare il lemma voluto a tutti i *contesti selezionati* con un'operazione unica, indipendentemente dal loro numero.

STRUTTURE DELLE TESTATE DELLA FINESTRA DI LEMMATIZZAZIONE E DEL CONTESTO



Prima riga di testata della finestra

Contiene, nell'ordine:

1. logo e nome del programma,
2. nome della finestra (LEMMATIZZAZIONE CORPUS),
3. disco e nome del corpus.

Seconda riga di testata della finestra

Contiene la dicitura "**tutte le occorrenze**", "**solo occorrenze non lemmatizzate**" o "**solo occorrenze lemmatizzate**", in rosso su bianco, per precisare se le occorrenze che vengono mostrate sono tutte quelle ottenute con la ricerca, solo quelle non ancora lemmatizzate o solo quelle già lemmatizzate.

Dopo questa dicitura vi è un pulsante (scritta blu su sfondo bianco) che consente di impostare alternativamente come ordine di presentazione dei contesti quello iniziale, ovvero per testo e, a parità di testo, per forme, oppure quello *per sintagmi* (per la descrizione di questo tipo di ordinamento, vedere la Lezione 29).

Terza riga di testata della finestra

Contiene, nell'ordine:

1. entrata lessicale (es.: occhio) e categoria grammaticale (es.: s.m.) del **lemma impostato**, ovvero del lemma utilizzato nel corso dell'ultima operazione di lemmatizzazione, in blu su celeste; se in tale operazione è stato coinvolto anche un iperlemma questo compare nella stessa riga (es.: attività sensoriali), separato dal lemma dal carattere /,
2. icona che rappresenta un istogramma: se l'icona è storta vuol dire che il quadro **statistiche** non è visualizzato; se è dritta vuol dire che lo è; cliccandola si inverte la situazione,
3. icona "NL" o "↓" indicante il significato del tasto <Invio> (nel primo caso assegnazione di un nuovo lemma al contesto corrente, nel secondo lemmatizzazione del contesto corrente mediante il *lemma impostato* e passaggio al contesto successivo),
4. icona **libro aperto** o **libro chiuso**: mostra lo stato del dizionario di macchina, rispettivamente aperto o chiuso, della forma di cui l'occorrenza del contesto corrente è espressione; cliccandola si inverte la situazione.

Riga di testata del contesto

Può contenere nell'ordine:

1. riquadro rosso: dove presente indica il *contesto corrente*, ovvero quello oggetto delle eventuali operazioni di lemmatizzazione;
2. numero del contesto, tra quelli visualizzati; in rosso nei contesti selezionati per la lemmatizzazione, in blu negli altri;
3. simbolo "#" indicante, dove compare, la presenza di una nota nei dati bibliografici, in blu scuro;
4. titolo abbreviato del testo, in verde;
5. simbolo "A" indicante, dove compare, la presenza di un brano associato al contesto (nota o traduzione), in marrone;
6. stringa "RO..", indicante il riferimento organico, in viola;
7. numero di verso per le opere in versi, in verde;

8. riferimento topografico: possono comparire il volume in blu, la pagina, la colonna nella pagina e la riga (nella colonna se presente oppure nella pagina) in blu scuro;
9. entrata lessicale del lemma associato all'occorrenza (se già lemmatizzata), in blu;
10. categoria grammaticale del lemma associato all'occorrenza (se già lemmatizzata), in blu;
11. simbolo "X" indicante che il lemma associato all'occorrenza (se già lemmatizzata) ha un contenuto non nullo nel campo *disambiguatore*, in nero;
12. simbolo "I" indicante che anche un iperlemma è associato all'occorrenza (se già lemmatizzata), in nero;

Cliccando sul numero di contesto (2) il contesto viene alternativamente selezionato o deselezionato per la lemmatizzazione; cliccando sui campi 3 o 4 si apre la finestra con i dati bibliografici e statistici del testo da cui il contesto è tratto; cliccando sul campo 5 si visualizza il brano associato; cliccando sui campi 6, 7 o 8 viene mostrato il riferimento organico completo; cliccando sui campi dal 9 al 12 vengono mostrati tutti i dati del lemma associato e dell'eventuale iperlemma.

LEZIONE 15. COME EFFETTUARE LA LEMMATIZZAZIONE

Combinando fra loro le tecniche di lemmatizzazione disponibili ('nuovo lemma', da dizionario di macchina, con uso o meno della tabella degli omografi, con inserimento o meno degli iperlemmi, utilizzando il 'lemma impostato',...) si possono ottenere decine di possibili percorsi di lemmatizzazione. Non è ovviamente ragionevole, né didatticamente opportuno, passare in rassegna la casistica completa, dal momento che un risultato molto probabile sarebbe quello di ingenerare una confusione non indifferente nel lettore, il che non è esattamente l'obiettivo che si prefigge il manuale.

In questa Lezione, così come nelle successive, verrà fatto riferimento al corpus **Demo** in quanto già parzialmente lemmatizzato e dotato di un dizionario di macchina, a differenza del corpus **Prova** che contiene solo testi non lemmatizzati. Verranno ora mostrati diversi esempi di lemmatizzazione, ciascuno effettuato tramite un percorso diverso dagli altri, per fornire una panoramica abbastanza ampia.

Per partire, avviare GATTO e portarsi nell'ambiente **Lemmatizzazione sul corpus** mediante [LEMMATIZZAZIONE + SUL CORPUS].

Tramite [CORPUS] aprire la finestra di selezione del corpus; cliccare sulla riga del corpus **Demo**, aprirlo con [APRI], quindi tornare alla finestra LEMMATIZZAZIONE CORPUS con [CHIUDI FINESTRA].

Esempio 1: lemmatizzazione del contesto corrente utilizzando il dizionario di macchina

Come primo esempio si lemmatizzerà un'occorrenza di **cavaliere** con **cavaliere (s.m.)**. Si dovranno inizialmente localizzare nel corpus tutte le occorrenze della forma **cavaliere**; i contesti relativi andranno trasferiti nella finestra di lemmatizzazione dopo di che, utilizzando il lemma **cavaliere (s.m.)** già presente nel dizionario di macchina di **cavaliere**, si procederà ad associarlo alla prima occorrenza non lemmatizzata di tale forma.

Partendo dalla finestra LEMMATIZZAZIONE CORPUS, avviare la ricerca per forme con [RICERCHE + PER FORME]; nel primo riquadro bianco della colonna **forme da cercare** digitare **cavaliere** e trasferire le occorrenze trovate nella finestra di lemmatizzazione tramite [CONTESTI].

La finestra visualizzata mostra le prime occorrenze localizzate, tutte tratte da "Novellino". Il primo contesto è marcato come *contesto corrente* dal riquadro rosso posto a sinistra del numero di contesto: la prossima operazione di lemmatizzazione agirà automaticamente su questo contesto.

Prima di continuare, aprire il dizionario di macchina di **cavaliere** cliccando sull'icona "libro chiuso" in alto a destra; il dizionario contiene l'unico lemma **cavaliere (s.m.)**, che è proprio quello che si intende utilizzare.

Per lemmatizzare l'occorrenza è sufficiente fare un singolo clic in un punto qualunque della riga di **cavaliere (s.m.)** nel riquadro del dizionario di macchina: nel rigo di testata del contesto apparirà, in blu, la dicitura **cavaliere s.m.**, che evidenzia l'effettuata lemmatizzazione del contesto tramite il lemma di entrata lessicale **cavaliere** e categoria grammaticale **s.m.**. Nel contempo il riquadro rosso si è posizionato sul contesto successivo, che ora è il nuovo *contesto corrente*.

Per uscire dalla finestra, predisponendosi per una nuova lemmatizzazione, usare il bottone (N. ricerca/Esci) in basso a destra. Volendo a questo punto uscire del tutto dall'ambiente **Lemmatizzazione sul corpus** usare [ESCI].

Esempio 2: lemmatizzazione dei contesti selezionati utilizzando il dizionario di macchina

Il meccanismo precedente, agendo per singoli contesti, consente di effettuare lemmatizzazioni di precisione, ma è piuttosto lento (vedere comunque, a questo riguardo, la Lezione 18) se il numero di contesti da lemmatizzare è elevato. Il ricorso alla tecnica di *selezione dei contesti* permette di accelerare le operazioni.

Supponiamo di voler lemmatizzare, sempre facendo uso del dizionario di macchina, tutte le occorrenze non ancora lemmatizzate della forma **camera**.

Come prima, partendo dalla finestra LEMMATIZZAZIONE CORPUS, portarsi nella finestra LEMMATIZZAZIONE: RICERCA PER FORME con [RICERCHE + PER FORME]; cancellare le impostazioni precedenti tramite [CANCELLA TUTTI I DATI], quindi digitare la forma **camera** e procedere con [AVVIO RICERCA]. La griglia inclusa nella finestra LEMMATIZZAZIONE: FORME LOCALIZZATE evidenzia che nel corpus **Demo** sono presenti 6 occorrenze di **camera**, ma non specifica nulla circa il loro stato di lemmatizzazione. Per saperne di più, cliccare su (**Mostra lemmi**): questa volta ci viene precisato che, di tali 6 occorrenze, 5 sono non lemmatizzate e 1 lo è con **camera (s.f.)**.

Volendo operare solo sulle occorrenze non ancora lemmatizzate, selezionare la riga della griglia relativa alle 5 occorrenze non lemmatizzate cliccando in un punto qualunque di essa: il simbolo **X** nella prima colonna conferma l'avvenuta selezione. Portare la riga così selezionata nella finestra ACCUMULATORE con [COPIA IN ACC.], selezionare quindi, allo stesso modo, la riga così ottenuta e trasferire i 5 contesti nella finestra di lemmatizzazione con [MOSTRA CONTESTI]. Come si vede il percorso seguito questa volta è leggermente diverso dal precedente: avendo voluto operare una selezione ulteriore tra i contesti (esclusione di quelli già lemmatizzati) è stato necessario passare per due finestre aggiuntive.

Aprire il dizionario di macchina cliccando sull'icona "**libro chiuso**": il dizionario contiene l'unico lemma **camera (s.f.)**. Scorrendo i 5 contesti è evidente che tale lemma si applica perfettamente a tutti quindi, anziché effettuare 5 lemmatizzazioni distinte, procederemo con un'operazione unica; a questo scopo cliccare sul numero di ognuno dei 5 contesti per selezionarlo (l'avvenuta selezione è evidenziata dal fatto che il numero di contesto passa da blu a rosso; cliccandolo di nuovo il colore cambia alternativamente). Se si sbaglia mira, cliccando troppo a sinistra o troppo a destra, si possono ottenere risultati non previsti, ovvero apertura della finestra dei dati bibliografici o trasferimento ad altro contesto del riquadro rosso di *contesto corrente*: ciò non crea problemi (la finestra può essere richiusa con <Esc> e il posizionamento del contesto corrente non ha relazione con la lemmatizzazione dei contesti selezionati).

Cliccare sull'option box lemm. selez. in basso a sinistra per attivare la lemmatizzazione dei contesti selezionati, quindi cliccare sulla riga di **camera (s.f.)** del dizionario di macchina, come nell'esempio precedente: come si può vedere, tutti i 5 contesti selezionati vengono lemmatizzati con tale lemma.

Notare che, al termine dell'operazione, l'*option box lemm. selez.* viene automaticamente disabilitato, in favore di **lemm. singola** (ovvero lemmatizzazione del *contesto corrente*): si tratta di una protezione, volta a non mantenere attiva oltre il tempo necessario un'impostazione che, se non voluta, può introdurre errori di lemmatizzazione non sempre facili da individuare e rimuovere.

Uscire dalla finestra e predisporre per una nuova lemmatizzazione con (**N. ricerca/Esci**).

Esempio 3: lemmatizzazione del contesto corrente utilizzando un lemma diverso da quelli offerti dal dizionario di macchina (e non presente nel lemmario)

Non sempre il dizionario di macchina contiene il lemma desiderato: in questi casi le indicazioni contenute in esso vanno rifiutate in favore della digitazione diretta di un nuovo lemma.

Si voglia, ad esempio, lemmatizzare un'occorrenza di **acciò** con **a ciò che (cong.)**.

Nella finestra LEMMATIZZAZIONE: RICERCA PER FORME (eventualmente rimuovere tutte le impostazioni con [CANCELLA TUTTI I DATI]) digitare **acciò**, cliccare sul bottone (**Mostra lemmi**) e procedere con [AVVIO RICERCA]; selezionare la riga relativa alle 11 occorrenze non lemmatizzate di **acciò** (prima riga della griglia) e procedere con [COPIA IN ACC.]; selezionare ancora la riga relativa ad **acciò** e scegliere [MOSTRA CONTESTI]. Aprire il dizionario di macchina.

Come si vede, l'unico lemma offerto è **acciocché (cong.)**. Se si preferisce la scrittura separata **a ciò che** basta procedere nella lemmatizzazione introducendo questo nuovo lemma e ignorando il suggerimento contenuto nel dizionario di macchina. A questo scopo cliccare sul bottone (**N. lem/omo/ipe**); nella finestrina **nuovo lemma** digitare **a ciò che** e cliccare su (**OK**); poiché il nuovo lemma non è presente nel lemmario del corpus, all'utente viene chiesto di introdurre anche gli altri elementi, quanto meno la categoria grammaticale: digitare **cong.** nel riquadro **categoria grammaticale** e poi dare (**OK**). Si ottengono due risultati,

immediatamente visibili: l'occorrenza è stata lemmatizzata col nuovo lemma e quest'ultimo è entrato a far parte del dizionario di macchina di **acciò**.

Uscire dalla finestra con **(N. ricerca/Esci)**.

Nota

In tutti i casi in cui si inseriscono i dati secondari di un nuovo lemma, la stessa finestra utilizzata in questo esempio per introdurre la categoria grammaticale (e volendo anche disambiguatore e commento), consente di associare all'occorrenza un iperlemma (già esistente o nuovo) digitandolo nelle ultime due caselle in basso (in tal caso l'ultima, relativa al disambiguatore dell'iperlemma, è disponibile ma non necessaria).

Esempio 4: lemmatizzazione del contesto corrente utilizzando un lemma diverso da quelli offerti dal dizionario di macchina (ma già presente nel lemmario)

Questo esempio è molto simile al precedente, salvo che in questo caso il nuovo lemma scelto, pur non essendo incluso nel dizionario di macchina della forma da lemmatizzare, è già presente nel lemmario del corpus, evidentemente in associazione con altre forme.

Si voglia, ad esempio, lemmatizzare un'occorrenza di **paia** con **parere (v.)**.

Partire dalla finestra LEMMATIZZAZIONE: RICERCA PER FORME, rimuovere eventualmente le impostazioni precedenti con **[CANCELLA TUTTI I DATI]** (questo suggerimento non verrà più ripetuto), digitare **paia** e procedere con **[CONTESTI]**. Aprire il dizionario di macchina.

Come si può vedere, l'unico lemma associato a **paia** è **paglia (s.f.)**, mentre l'occorrenza inclusa nel primo contesto è palesemente una forma verbale.

Come prima, cliccare su **(N. lem/omo/ipe)** e nella nuova finestra digitare **parere** seguito da **(OK)**.

Questa volta la parola digitata è omografa di 2 lemmi presenti nel lemmario, **parere (s.m.)** e **parere (v.)**, mostrati all'interno della *tabella degli omografi*. Ci interessa il secondo, quindi cliccare sulla riga di **parere (v.)**. Si ottengono 3 risultati: l'occorrenza è lemmatizzata con **parere (v.)**, quest'ultimo lemma è aggiunto al dizionario di macchina della forma e il contesto successivo diviene *contesto corrente*.

Esempio 5: lemmatizzazione del contesto corrente con un nuovo lemma

Lemmatizzazione di un contesto di **rege** con **re (s.m.)**, lemma non presente nel lemmario, senza fare ricorso al dizionario di macchina.

Come in precedenza, impostare la ricerca della forma **rege** e trasferirne direttamente i contesti nella finestra di lemmatizzazione con **[CONTESTI]**; questa volta non aprire il dizionario di macchina (aprendolo si constaterrebbe comunque che è vuoto, coerentemente col fatto che tutte le occorrenze di **rege** risultano non lemmatizzate).

Per lemmatizzare il primo contesto cliccare su **(N. lem/omo/ipe)**, digitare **re**, cliccare su **(OK)**, digitare **s.m.** nella casella apposita e terminare con **(OK)**. La lemmatizzazione è stata effettuata. Aprendo il dizionario di macchina si può vedere che non è più vuoto in quanto il nuovo lemma vi è stato incluso.

Esempio 6: lemmatizzazione di tutti i contesti non lemmatizzati

Nell'esempio 2 è stato mostrato come lemmatizzare più contesti con un'operazione unica, ricorrendo alla selezione dei contesti interessati. Quando *tutti* i contesti presentati vanno lemmatizzati con lo stesso lemma non è necessario selezionarli ad uno ad uno, ma è possibile selezionarli tutti insieme; se i contesti da lemmatizzare sono *quasi* tutti, può convenire selezionarli tutti con questa tecnica e poi deselegionare, sempre cliccando sul numero di contesto, quelli da escludere dalla lemmatizzazione.

Supponiamo di voler lemmatizzare tutte le occorrenze non lemmatizzate di **allegro** con **allegro (agg.)**.

Impostare nel solito modo la ricerca della forma **allegro** e portarsi nella finestra di lemmatizzazione con **[CONTESTI]**. Delle 6 occorrenze presenti solo l'ultima è lemmatizzata.

Cliccare su **(Selez. tutte)** per selezionare tutte le occorrenze, deselegionare l'ultima cliccando sul numero 6 (questa operazione non è in realtà necessaria perché rilemmatizzando un'occorrenza con lo stesso lemma già assegnatole in precedenza non succede niente); cliccare su **lemm. selez.**

Volendo fare uso del dizionario di macchina cliccare sull'icona relativa e poi sulla riga di **allegro (agg.)**; in alternativa cliccare su **(N. lem/omo/ipe)**, digitare **allegro**, cliccare su **(OK)**, e poi sulla riga di **allegro (agg.)** nella *tabella degli omografi*. In entrambi i casi tutte le occorrenze non ancora lemmatizzate vengono associate al lemma. L'*option box* **lemm. selez.** viene disabilitato; i contesti marcati come *selezionati* rimangono tali (anche per agevolare un eventuale intervento correttivo in caso di lemmatizzazione errata,

magari con un lemma diverso da quello voluto). Per deselezionare tutte le occorrenze cliccare su (**Deselez. tutte**).

Esempio 7: lemmatizzazione con nuovo omografo

È stato già visto più volte, negli esempi precedenti, l'utilizzo della *tabella degli omografi*, la quale viene presentata automaticamente in conseguenza della digitazione di un lemma appunto omografo di altri presenti nel lemmario. Nei casi visti è sempre stato accettato uno degli omografi proposti. Si vedrà ora come procedere nel caso in cui nessuno di essi sia quello desiderato.

Per lemmatizzare alcune delle occorrenze della forma **saggio** è stato fatto uso del lemma **saggio (agg.)**. In altri casi il lemma corretto è **saggio (s.m.)**, mai usato finora: si utilizzerà appunto quest'ultimo per lemmatizzare un contesto di **saggio**.

Procedere come al solito, fino ad ottenere nella finestra di lemmatizzazione le 8 occorrenze della forma **saggio**. Per verificare che non ve ne sono altre, cliccare sull'icona che rappresenta un istogramma inclinato di 45°, in alto a destra: apparirà in basso un riquadro indicante che sono state localizzate appunto 8 occorrenze della forma **basso**, delle quali 3 lemmatizzate. Cliccando di nuovo sull'icona il riquadro scompare.

Come si può vedere, il 3° contesto andrebbe lemmatizzato col lemma **saggio (s.m.)**; per dichiararlo contesto corrente si può, alternativamente:

1. utilizzare due volte il tasto freccia in basso <↓>;
2. cliccare in un punto qualunque del testo del 3° contesto;
3. cliccare, nel 3° contesto, nel punto riservato al riquadro rosso;
4. utilizzare il bottone (**Vai a occ. n.**), quindi digitare **3** e (**OK**).

Una volta che il 3° contesto è stato dichiarato *contesto corrente*, cliccare su (**N. lem/omo/ipe**) e digitare **saggio**; come era prevedibile, si apre la tabella degli omografi, suggerendo la soluzione **saggio (agg.)**. Per rifiutarla, cliccare di nuovo su (**N. lem/omo/ipe**) (per l'appunto questo bottone indica "Nuovo lemma", oppure "Nuovo omografo" o infine "Nuovo iperlemma") e inserire la categoria grammaticale del nuovo lemma, ovvero **s.m.**; terminare con (**OK**). È facile vedere che l'operazione è andata a buon fine.

Esempio 8: lemmatizzazione con utilizzo di un iperlemma associato al lemma scelto

Come discusso in precedenza, nei casi in cui al lemma che si intende utilizzare è associato un iperlemma di livello 1, è possibile associare anche quest'ultimo direttamente all'occorrenza in corso di lemmatizzazione.

La forma **charta** è lemmatizzabile col lemma **carta (s.f.)**, al quale è associato in **Demo** l'iperlemma **scrittura**. Si procederà a lemmatizzare un'occorrenza di **charta** associando a tale occorrenza sia il lemma che l'iperlemma.

Cercare e portare nella finestra di lemmatizzazione le occorrenze della forma **charta**. A parte la prima occorrenza, le altre visibili non sono lemmatizzate; portarsi sul contesto n. 2 e aprire il dizionario di macchina.

Lemmatizzare l'occorrenza con **carta (s.f.)** cliccando sulla riga corrispondente nel dizionario di macchina: poiché al lemma risulta associato un iperlemma di livello 1 (**scrittura**), questo viene proposto come disponibile per essere associato all'occorrenza: cliccare anche sulla riga di **scrittura**. Il dizionario di macchina riassume l'aspetto abituale, la lemmatizzazione è stata effettuata e, se si osserva la riga di testata del 2° contesto, si noterà che accanto a **carta s.f.** (in blu), compare una **"I"** in nero, ad indicare che all'occorrenza è associato anche un iperlemma. Per sapere quale, cliccare su **carta**, su **s.f.** o su **I** per far apparire la finestra **lemma associato**, la quale mostra i dati associati all'occorrenza, compreso l'iperlemma (chiudere la finestra con <Esc>).

L'esempio è terminato ma per comodità si consiglia di non uscire dalla finestra di lemmatizzazione, che verrà utilizzata anche negli esempi successivi.

Esempio 9: lemmatizzazione con utilizzo di un nuovo iperlemma da associare al lemma scelto

Nell'esempio precedente è stato utilizzato un iperlemma già presente in **Demo**. Vediamo ora come effettuare l'operazione precedente associando al lemma un iperlemma nuovo.

Sempre con riferimento ai dati dell'esempio precedente, la forma sarà **charta**, il lemma **carta (s.f.)**, l'iperlemma **materiali**.

Nella finestra di lemmatizzazione che contiene le occorrenze di **charta**, portarsi sul contesto n. 3 e lemmatizzarlo cliccando sulla riga relativa a **carta (s.f.)** nel dizionario di macchina; come prima, viene proposto anche l'iperlemma **scrittura**; per associare all'occorrenza un diverso iperlemma, cliccare su (**N. lem/omo/ipe**), digitare **materiali** e chiudere con (**OK**).

La lemmatizzazione si è conclusa e cliccando sui dati del lemma o sulla **I**, nella testata dell'occorrenza, è possibile vedere che ad essa è stato associato l'iperlemma appena digitato.

L'ultima finestra che si è aperta è quella che consente di inserire i dati secondari di un nuovo lemma: tutte le volte che si digita un lemma che non è presente nel lemmario, si apre questa finestra tramite la quale si possono associare tutti i dati (categoria grammaticale, disambiguatore, commento) al nuovo lemma.

Come ulteriore conseguenza dell'operazione, il nuovo iperlemma è associato più in generale al lemma **carta (s.f.)**. Per verifica, cliccare nel dizionario di macchina in un punto qualunque col tasto destro, poi cliccare col tasto sinistro sulla riga del lemma, in una delle colonne esclusa "lemma", ovvero sotto "**cat. gr.**", "**disambiguatore**" o "**commento**": ora gli iperlemmi associati a **carta (s.f.)** sono due: **materiali** e **scrittura** (chiudere la nuova finestra con <Esc>).

Se si fosse cliccato col tasto destro sul dizionario di macchina e poi, col tasto sinistro, sulla riga di **carta (s.f.)** ma nella colonna "**lemma**", si sarebbe aperta una finestra elencante non gli iperlemmi ma le forme associate al lemma (nel caso in questione **carta**, **charta** e **charte**).

Esempio 10: lemmatizzazione con utilizzo di un nuovo iperlemma da associare al lemma scelto, nel caso in cui a quest'ultimo non sia al momento associato nessun iperlemma

Il meccanismo suggerito nell'esempio precedente per associare un nuovo iperlemma ad un'occorrenza non funziona se il lemma usato è privo di iperlemmi di livello 1: in tal caso, infatti, il clic sulla riga nel dizionario di macchina o nella tabella degli omografi assegna il lemma all'occorrenza e chiude la lemmatizzazione, senza offrire all'utente la possibilità di proseguire e introdurre l'iperlemma.

Per risolvere il problema bisogna trattare il lemma come un *nuovo lemma* anche se in realtà è già presente nel lemmario, ignorando i suggerimenti contenuti sia nel dizionario di macchina che nella tabella degli omografi, in modo da costringere il programma ad aprire la finestra di introduzione dei dati secondari del lemma, incluso l'iperlemma. Il tutto sarà più chiaro con un esempio.

Si ricerchino e si portino nella finestra di lemmatizzazione le occorrenze della forma **navi**. Dal riquadro **Statistiche** (clic su icona "**istogramma inclinato**", in alto a destra) si può vedere che ci sono 12 occorrenze, delle quali 2 lemmatizzate col lemma **nave (s.f.)**.

Supponiamo di voler lemmatizzare la prima occorrenza con quel lemma, associandole in più l'iperlemma **mezzi di trasporto** (del quale non si sa se esiste nel corpus né se è associato al lemma **nave (s.f.)**).

Un primo modo per sapere se l'iperlemma è associato al lemma consiste nell'aprire il dizionario di macchina e, come fatto nell'esempio precedente, cliccare al suo interno col tasto destro e poi, con quello sinistro, nei riquadri "**cat. gr.**", "**disambiguatore**" o "**commento**" nella riga del lemma: nessun iperlemma risulta associato al lemma (uscire con <Esc> dalla nuova finestra).

In alternativa si può semplicemente provare a lemmatizzare il primo contesto col lemma, e vedere se vengono o meno presentati degli iperlemmi. A questo scopo controllare che il primo contesto sia quello corrente, poi cliccare col tasto sinistro sulla riga di **nave (s.f.)** nel dizionario di macchina: la lemmatizzazione è conclusa, quindi il lemma era privo di iperlemmi e non è stato possibile associare l'iperlemma voluto al contesto.

A questo punto, poiché si voleva invece associare alla prima occorrenza di **navi** sia il lemma che il nuovo iperlemma, bisogna riportare il riquadro rosso al primo contesto e ripetere la lemmatizzazione (il fatto che a questo punto il contesto risulti già lemmatizzato è irrilevante: il lemmatizzatore di GATTO permette sempre di rilemmatizzare un'occorrenza, rimuovendo automaticamente il lemma precedente).

Questa volta però, per non ritrovarsi nella stessa situazione di prima, avviare la lemmatizzazione ignorando il dizionario di macchina e cliccando su (**N. lem/omo/ipe**), poi digitare **nave** e (**OK**); la tabella degli omografi suggerisce **nave s.f.**; ignorarla e cliccare di nuovo su (**N. lem/omo/ipe**). Nella finestra di introduzione dei dati secondari del lemma digitare **s.f.** sotto **categoria grammaticale** e **mezzi di trasporto** sotto **iperlemma**; terminare con (**OK**).

Se, con le tecniche già descritte, si vanno a guardare i dati relativi al lemma associato al primo contesto di **navi** si ritrova il solito lemma, con in più l'iperlemma; se dal dizionario di macchina si visualizzano gli iperlemmi associati al lemma **nave (s.f.)** si ritrova **mezzi di trasporto**. D'ora in poi nel corso di ogni lemmatizzazione effettuata con questo lemma verrà offerta all'utente la possibilità di associare all'occorrenza anche l'iperlemma citato.

È importante avere ben chiaro che l'operazione eseguita nel corso di questo esempio non ha affatto creato un duplicato nel lemmario: ci si potrebbe cioè attendere la simultanea presenza di due lemmi, entrambi con entrata lessicale **nave** e categoria grammaticale **s.f.**, uno associato all'iperlemma **mezzi di**

trasporto e l'altro no. In realtà GATTO si è accorto che il lemma introdotto come *nuovo* era già presente nel lemmario e si è limitato a fare uso di questo, semplicemente agganciandogli il nuovo iperlemma.

Esempio 11: lemmatizzazione senza iperlemma, con rifiuto degli iperlemmi suggeriti

Come è stato detto più volte, a partire dal momento in cui ad un lemma è associato un iperlemma di livello 1, quest'ultimo viene sempre proposto quando si lemmatizza facendo uso del lemma; ovviamente rimane la possibilità di lemmatizzare facendo uso del solo lemma, senza associare all'occorrenza anche l'iperlemma. Questo esempio mostra come fare.

Si portino nella finestra di lemmatizzazione le occorrenze della forma **charta**, già utilizzata in precedenza, alla quale è associato il lemma **carta (s.f.)**, a sua volta collegato agli iperlemmi **materiali** e **scrittura**. Questa volta il contesto n. 4 verrà lemmatizzato col solo lemma, senza iperlemmi.

Prima di tutto attivare come contesto corrente il n. 4, aprire il dizionario di macchina e cliccare sulla riga di **carta s.f.** per lemmatizzare l'occorrenza. Per concludere la lemmatizzazione senza inserirvi nessuno dei due iperlemmi suggeriti cliccare in un punto qualunque della finestra **iperlemmi disponibili** col tasto destr del mouse. Se invece si fosse usato il tasto <Esc>, l'intera operazione di lemmatizzazione sarebbe stata annullata.

Per concludere la trattazione degli iperlemmi va fatto notare che non è possibile rimuovere l'iperlemma associato ad un'occorrenza, lasciando invece il lemma: per ottenere questo risultato è necessario ripetere la lemmatizzazione dell'occorrenza, questa volta utilizzando il solo lemma.

Esempio 12: lemmatizzazione mediante il lemma impostato

Il *lemma impostato* è il lemma che può comparire (in blu su sfondo celeste) nella terza riga della finestra di lemmatizzazione. Vengono di volta in volta inseriti in questo campo sia il lemma usato nell'ultima operazione di lemmatizzazione sia quello che compare per primo in un dizionario di macchina, al momento della sua apertura.

È possibile ridefinire manualmente il lemma impostato sulla base del lemma assegnato a un contesto. Per farlo, cliccare col tasto destro in un punto qualunque del testo di quest'ultimo. Per rimuovere il lemma impostato cliccare col tasto destro su un contesto non lemmatizzato.

In **Demo** la forma **dentro** risulta associata a diversi lemmi; tra questi vi è **dentro (avv.)**. Scopo di questo esempio è di trasferire tale lemma nel riquadro *lemma impostato* e poi utilizzarlo per lemmatizzare alcuni altri contesti.

Con le solite tecniche cercare nel corpus e trasferire nella finestra di lemmatizzazione le occorrenze della forma **dentro**; aprire il dizionario di macchina. Compaiono 4 lemmi, il primo dei quali viene automaticamente copiato nel riquadro *lemma impostato*.

Osservando i contesti mostrati, si vede che il 2, il 7 e l'8 potrebbero essere lemmatizzati con **dentro (avv.)**, usato tra l'altro nel contesto n. 5. A questo punto cliccare col tasto destr in un punto qualunque del testo del contesto n. 5: come risultato il lemma usato per tale contesto viene definito come attuale *lemma impostato*. Selezionare ora i contesti 2, 7 e 8 cliccando sui loro numeri e attivare l'*option box lemm. selez.* Infine cliccare col tasto sinistro in un punto qualunque del testo del contesto corrente: quest'ultima operazione lemmatizza utilizzando il lemma impostato. In altre parole, se fosse stato attivo l'*option box lemm. singola* sarebbe stato lemmatizzato il (solo) contesto corrente; poiché invece era attiva l'opzione alternativa quella stessa operazione ha applicato la lemmatizzazione ai contesti selezionati (notare che il contesto corrente non è stato lemmatizzato in quanto non era stato selezionato, ma niente impedisce di selezionarlo per lemmatizzarlo assieme agli altri).

Un altro modo per lemmatizzare facendo uso del lemma impostato consiste nel cliccare, all'interno del riquadro del dizionario di macchina, nella zona vuota (se c'è) sottostante all'ultimo lemma riportato.

Sempre circa il lemma impostato c'è da notare che l'operazione di caricamento di quel campo sulla base del lemma associato ad un contesto (tasto destro sul contesto), trasporta in quel campo anche l'eventuale iperlemma presente nel contesto di partenza: tale iperlemma verrà associato al lemma in tutte le lemmatizzazioni effettuate in seguito sulla base del lemma impostato.

Esempio 13: lemmatizzazione con antroponimi

Nel caso di antroponimi, soprattutto in presenza di forme polirematiche, la lemmatizzazione può essere resa più agevole utilizzando un apposito bottone che genera lemmi polirematici aventi lo stesso assetto grafico dell'antroponimo cui si riferiscono.

Vediamo la cosa con un esempio: si voglia lemmatizzare **Salvi del Forese** con il lemma antroponimo omografo.

Nella finestra LEMMATIZZAZIONE CORPUS avviare la ricerca per forme con [RICERCHE + PER FORME]; digitare **salv*** e procedere con [AVVIO RICERCA]. Tra le forme localizzate vi è per l'appunto un'occorrenza di **Salvi del Forese**; selezionarla e cliccare su [COPIA IN ACC.]; selezionare di nuovo la riga e proseguire con [MOSTRA CONTESTI]. Ovviamente appare l'unico contesto individuato. Per lemmatizzarlo, cliccare su (N. Ant): appare la finestra di introduzione dati secondari nella quale sono proposti sia il lemma (**salvi del forese**), sia la categoria grammaticale (**n.p.**), sia il disambiguatore (**antr.::**). È possibile accettare o modificare i dati proposti, così come è possibile integrarli usando le altre caselle presenti nella finestra. Al termine completare la lemmatizzazione con (OK) (o annullarla con (Annulla)).

Esempio 14: lemmatizzazione con nomi geografici

Con una logica in tutto e per tutto simile a quella degli antroponimi è possibile operare sui nomi geografici.

A titolo di esempio si lemmatizzerà **San Donato a Torre** usando un lemma polirematico avente lo stesso assetto grafico.

Nella finestra LEMMATIZZAZIONE CORPUS avviare la ricerca per forme con [RICERCHE + PER FORME]; digitare **san *** e procedere con [AVVIO RICERCA]. L'unica forma localizzata è **san donato a torre**; selezionarla e cliccare su [COPIA IN ACC.]; selezionare di nuovo la riga e proseguire con [MOSTRA CONTESTI]. Appare l'unico contesto individuato. Per lemmatizzarlo, cliccare su (N. Top): appare la finestra di introduzione dati secondari nella quale sono proposti sia il lemma (**san donato a torre**), sia la categoria grammaticale (**n.p.**) sia il disambiguatore (**n.g.::**). È possibile accettare o modificare i dati proposti, così come è possibile integrarli usando le altre caselle presenti nella finestra. Al termine completare la lemmatizzazione con (OK) (o annullarla con (Annulla)).

LEZIONE 16. USO DELLA FINESTRA A CONTESTO SINGOLO

Nel corso della Lezione precedente sono stati mostrati diversi esempi di lemmatizzazione, tutti effettuati utilizzando la finestra a contesti multipli, la quale visualizza un certo numero di contesti assegnando a ciascuno due righe dello schermo.

Qualora le caratteristiche del testo in generale o di uno specifico contesto in particolare richiedessero la lettura di una parte di testo più estesa è possibile passare alla finestra *a contesto singolo* che, come suggerisce la dicitura, mostra un solo contesto per volta ma con la possibilità di ampliarne le dimensioni.

Per vedere un esempio di questa visualizzazione impostare la lemmatizzazione della forma **casa**. Portarsi sull'occorrenza n. 2, trasferendovi il riquadro rosso. Per passare alla finestra a contesto singolo cliccare su (Occ. singola) in basso. A questo punto lo stesso contesto viene mostrato da solo; la sua riga di testata è invariata mentre il corpo del contesto viene rappresentato riproducendo la struttura del testo così come contenuta nel testo; la lunghezza del contesto è pari a 3 periodi e la parola da lemmatizzare compare in quello centrale ("...**Tornò a casa al suo marito e mostrolli crucciose sembianti.** ...") Volendo estendere il contesto usare (Allarga): ad ogni clic il contesto viene esteso di due periodi, con l'aggiunta di uno in testa e di uno in coda. Arrivati al 5° clic, si può notare che il contesto non si estende più all'indietro, cioè che non vengono inseriti ulteriori periodi all'inizio: questo è dovuto al fatto che è stato raggiunto l'inizio del testo associato al riferimento organico cui appartiene il contesto. Per ridurre le dimensioni del contesto usare (Restringi), con effetti opposti: il valore limite inferiore per la lunghezza del contesto è un periodo. In tutti i casi l'estensione del contesto è comunque limitata alla parte di testo associata al riferimento organico immediatamente precedente e comunque la parte visualizzata non oltrepassa la schermata.

Volendo passare al contesto successivo utilizzare il tasto <↓> come nel caso dei contesti multipli: la dimensione del contesto in periodi rimane quella precedentemente impostata.

Tutti gli altri comandi sono identici al caso della visualizzazione a contesti multipli. Anche il meccanismo di lemmatizzazione dei contesti selezionati funziona regolarmente, anche se i contesti cui si applica non sono visibili (del resto la stessa situazione può aver luogo nella finestra a contesti multipli, la quale mostra solo alcuni contesti, non tutti quelli lemmatizzabili con l'operazione); l'unica differenza è che l'impostazione degli *option box lemm. singola* o *lemm. selez.* va effettuata dalla finestra a contesti multipli: l'impostazione scelta rimane valida anche dopo il passaggio alla finestra a contesto singolo.

Per tornare alla finestra a contesti multipli usare (**Occ. multiple**): in questo caso la finestra si apre mostrando il contesto attuale come primo contesto. Per far riapparire anche i contesti precedenti risalire col tasto <↑> o usando gli altri comandi disponibili per cambiare il contesto corrente.

LEZIONE 17. SELEZIONE DEL TIPO DI OCCORRENZE E COMANDI DI SPOSTAMENTO TRA CONTESTI

RESTRIZIONI NELL'ACCESSO ALLE OCCORRENZE SULLA BASE DEL LORO STATO DI LEMMATIZZAZIONE

Normalmente tutti i contesti ottenuti come risultato di una ricerca e trasferiti nella finestra di lemmatizzazione sono disponibili per essere lemmatizzati: esiste però la possibilità di limitare le operazioni di lemmatizzazione ai soli contesti non ancora lemmatizzati, oppure a quelli già lemmatizzati.

Per agire su questa opzione, una volta arrivati alla finestra di lemmatizzazione, utilizzare il bottone (**Tipo occorr.**).

A titolo di esempio, impostare la lemmatizzazione della forma **castello**. Nella finestra di lemmatizzazione compariranno 3 contesti, di cui il primo e il terzo risultano lemmatizzati con **castello (s.m.)**, mentre il secondo non è lemmatizzato. Nella seconda riga di testata compare la dicitura **tutte le occorrenze**.

Cliccando su (**Tipo occorr.**) la dicitura viene cambiata in **solo occorrenze non lemmatizzate** e compare solo il secondo contesto; notare che il numero di contesto viene mantenuto. Ripetendo l'operazione si passa alla dicitura **solo occorrenze lemmatizzate** e corrispondentemente solo i contesti 1 e 3 sono presentati. Cliccando ancora sul solito bottone si ritorna alla situazione di partenza.

COMANDI UTILIZZABILI PER IL PASSAGGIO AD ALTRI CONTESTI

Avendo ultimato la panoramica sulle operazioni di lemmatizzazione vere e proprie è opportuno riassumere e completare il quadro dei comandi disponibili per passare da un contesto ad un altro.

clik sul contesto

Se il nuovo contesto è visibile è possibile trasformarlo in *contesto corrente* cliccando in un suo punto qualunque, esclusa la riga di testata.

<↓> e <↑>

Passaggio al contesto successivo o precedente.

<Pag↓> e <Pag↑> (finestra a contesti multipli)

Passaggio alla schermata successiva o precedente.

<↖>

Passaggio al primo contesto.

(Vai a occ. n.)

Accesso diretto a una occorrenza definita in base al numero d'ordine.

LEZIONE 18. FUNZIONI ACCESSORIE, ULTERIORI INFORMAZIONI SULLE RICERCHE DI PARTENZA E RASSEGNA DEI COMANDI PER VISUALIZZARE INFORMAZIONI

Le Lezioni precedenti hanno descritto l'insieme delle operazioni di lemmatizzazione previste nell'ambiente **Lemmatizzazione sul corpus**. Prima di chiudere l'argomento verranno brevemente passati in rassegna alcuni comandi e operazioni effettuabili, in parte già citati nel corso delle precedenti Lezioni che, insieme ad alcune ulteriori considerazioni di dettaglio, completano il quadro informativo.

FUNZIONI ACCESSORIE

Sostituzione di una lemmatizzazione preesistente

Questa situazione si verifica quando ci si accinge a rilemmatizzare un'occorrenza già lemmatizzata con un lemma diverso, per cambiare quest'ultimo e/o intervenire sull'iperlemma. Non c'è nessuna azione speciale da compiere: è sufficiente procedere con la normale procedura di lemmatizzazione e la lemmatizzazione precedente verrà rimossa automaticamente e sostituita.

Cancellazione di lemmatizzazioni preesistenti

Portarsi sul contesto in oggetto trasformandolo quindi nel *contesto corrente*, poi cliccare su (**Canc. lemma**). In alternativa selezionare i contesti dai quali si vogliono rimuovere le lemmatizzazioni, cliccare su **lemm. selez.** e quindi su (**Canc. lemma**). Fare attenzione ai contesti scelti poiché questa operazione non chiede conferma e non è annullabile.

Come selezionare o deselezionare contemporaneamente tutte le occorrenze mostrate

Utilizzare i bottoni (**Selez. tutte**) e (**Deselez. tutte**) posti in basso al centro.

Ridefinizione del lemma impostato a partire da quello associato a un contesto

È sufficiente cliccare all'interno del testo del contesto considerato col tasto destra del mouse: il lemma usato per lemmatizzare il contesto (completo di eventuale iperlemma) verrà copiato nel riquadro del lemma impostato.

Rimozione del lemma impostato

Fermo restando che si tratta di un'operazione superflua, in quanto la presenza del lemma impostato, se non utilizzato, non ha effetti, un modo per rimuoverlo è quello di seguire le indicazioni del paragrafo precedente, avendo cura di scegliere un contesto non lemmatizzato (il clic col tasto destro importa nel lemma impostato l'*assenza di lemma* del contesto di partenza).

Tasto <Enter> per attivare l'inserimento di un nuovo lemma

Normalmente per assegnare un nuovo lemma ad un'occorrenza si fa uso del bottone (**N. lem/omo/ipe**), attivabile cliccandolo o tramite <Alt> + <n>. Nella terza riga di testata della finestra esiste sulla destra un'icona che può mostrare il simbolo "NL" o "↓". Se il simbolo è "NL" il tasto <Enter> acquista il significato di "Nuovo lemma". Per invertire l'icona cliccarvi sopra.

Lemmatizzazione veloce di occorrenze consecutive utilizzando lo stesso lemma

Volendo lemmatizzare diverse occorrenze consecutive della stessa forma facendo uso dello stesso lemma, senza utilizzare il meccanismo del selezionamento preventivo, importare nel riquadro *lemma impostato* il lemma voluto (lemmatizzando una delle occorrenze con tale lemma oppure trasferendolo da un'occorrenza alla quale sia stato già associato - vedere sopra "Ridefinizione del lemma impostato a partire da quello associato a un contesto") e poi lemmatizzare le occorrenze cliccandovi sopra in successione (col tasto sinistro).

Se (cliccandovi sopra) si fa in modo che l'icona posta sulla destra nella terza riga di testata evidenzii il simbolo "↓", anziché "NL", la stessa operazione potrà realizzarsi digitando <Enter>, anziché cliccando col mouse, per ogni contesto successivo che si intende lemmatizzare sempre a partire dal lemma impostato.

ULTERIORI INFORMAZIONI SULLE RICERCHE DA CUI PARTIRE E SULLE FINESTRE DEI RISULTATI

Lemmatizzazione a partire da una ricerca di più forme

In tutti gli esempi fatti le ricerche sono state dirette a localizzare di volta in volta i contesti di una forma specifica. Più in generale facendo uso dei caratteri jolly è possibile trasferire contemporaneamente nell'ACCUMULATORE e da lì nella finestra di lemmatizzazione occorrenze di forme diverse. Ad esempio la ricerca di **cas*** localizzerebbe le occorrenze di tutte le forme che iniziano per **cas**, come **case**, **castello**, ecc.

Lemmatizzazione a partire da una ricerca diversa rispetto a quella per forme

Tutti gli esempi fin qui presentati hanno preso l'avvio da una ricerca per forme all'interno del corpus. In realtà la finestra LEMMATIZZAZIONE: MODALITÀ RICERCA prevede altri 4 tipi di ricerca, in alternativa. Questi possono essere utilizzati per usi particolari:

- **ricerca per forme mai lemmatizzate:** consente di risalire alle occorrenze delle sole forme che risultano non essere state mai lemmatizzate nel corpus. Ad esempio, la ricerca di **a***, seguita da [AVVIO RICERCA] fornirà in **Demo** circa **1860** forme diverse impostando [RICERCHE + PER FORME] ma solo **260** circa partendo da [RICERCHE PER FORME MAI LEMMATIZZATE] (usare [N. ELEMENTI] nella finestra LEMMATIZZAZIONE: FORME LOCALIZZATE,
- **ricerca per lemmi, categorie grammaticali, disambiguatori:** al contrario queste ricerche hanno lo scopo di isolare preventivamente i contesti già lemmatizzati, per esempio al fine di modificarne o rimuoverne le lemmatizzazioni.

Limitazione a uno o più sottocorpus nell'impostazione della ricerca e conseguente selezione dei contesti da lemmatizzare

A differenza di quanto illustrato finora, è possibile limitare la ricerca dalla quale prende l'avvio ogni operazione di lemmatizzazione a uno o più sottocorpus, per esempio volendo lemmatizzare specificamente un certo dialetto o linguaggio o i testi di un certo autore.

L'operazione si svolge in due tempi: definizione del (o dei) sottocorpus e successiva limitazione della ricerca ad uno o più di essi.

La definizione dei sottocorpus viene effettuata a partire dalla finestra LEMMATIZZAZIONE: MODALITÀ RICERCA, tramite il menu [SOTTOCORPUS]: per i dettagli dell'operazione vedere il cap. 20 (Lezione 40).

La limitazione di una ricerca a dei sottocorpus è fatta contestualmente con l'impostazione della ricerca stessa, quindi ad esempio nella finestra LEMMATIZZAZIONE: RICERCA PER FORME; anche in questo caso, per i dettagli, vedere il cap. 20 (Lezione 41).

Utilizzo di opzioni nell'impostazione della ricerca

L'impostazione di una ricerca può essere modificata facendo ricorso ad una serie di opzioni disponibili nella finestra di impostazione; tali opzioni differiscono tra loro in funzione del tipo di ricerca impostata.

Informazioni complete a questo riguardo possono essere trovate nella parte IV del manuale, essenzialmente nei capp. 14, 15 e 16.

Ritorno all'accumulatore al termine di una lemmatizzazione

Una volta impostata una ricerca e arrivati alla finestra LEMMATIZZAZIONE: ACCUMULATORE (usando in successione i menu [AVVIO RICERCA] e [COPIA IN ACC.]) è possibile procedere al trasferimento nella finestra di lemmatizzazione di tutte le forme riportate nella griglia oppure solo di alcune di esse. Terminata la lemmatizzazione di queste, si può tornare a LEMMATIZZAZIONE: ACCUMULATORE cliccando su (**Torna a acc.**), selezionare altre forme, trasferire queste ultime nella finestra di lemmatizzazione, e così via.

Modifica dei dati contenuti nell'accumulatore al termine di una lemmatizzazione

La finestra LEMMATIZZAZIONE: ACCUMULATORE riporta il numero di occorrenze delle diverse coppie forma-lemma presenti: questi dati possono variare in funzione delle operazioni di lemmatizzazione effettuate.

Ad esempio, si imposti la ricerca per forme di **cade**, cliccando sul bottone (**mostra lemmi**); poi procedere con [AVVIO RICERCA], selezionare entrambe le righe ottenute e cliccare su [COPIA IN ACC.]. In LEMMATIZZAZIONE: ACCUMULATORE è evidenziato che esistono 1 occorrenza di **cade** non lemmatizzata e 1 lemmatizzata con **cadere (v.)**. Si selezionino entrambe e si continui con [MOSTRA CONTESTI]. Si lemmatizzi anche il primo contesto (quello non lemmatizzato) con **cadere (v.)**, quindi si torni a LEMMATIZZAZIONE: ACCUMULATORE cliccando su (**Torna a acc.**). Si potrà vedere che i due numeri di occorrenze sono cambiati e risultano 0 e 2, in conseguenza dell'operazione effettuata. Se si tentasse ora di selezionare solo la riga di **cade** non lemmatizzata si otterrebbe come risposta il messaggio "**nessuna occorrenza localizzata**". Una successiva ricerca di **cade** non mostrerebbe più questa riga.

Questo meccanismo di aggiornamento in tempo reale dell'accumulatore opera solo sulle righe di dati già presenti in esso all'avvio della lemmatizzazione: in altre parole se, lemmatizzando, ad una delle forme citate nell'accumulatore venisse associato un nuovo lemma, la nuova coppia forma-lemma non verrebbe visualizzata immediatamente in una nuova riga della griglia.

Presenza di barrette nere orizzontali sopra la testata di un contesto, in corrispondenza del titolo abbreviato o a destra (nella zona dei dati del lemma associato)

La prima delle due barrette citate, quella cioè soprastante il titolo abbreviato, indica che il contesto che segue è tratto da un testo diverso rispetto al contesto precedente.

La seconda barretta, posta sulla destra, indica che il contesto successivo è relativo a una parola diversa rispetto al precedente.

Il primo contesto mostrato riporta sempre entrambe le barrette.

COMANDI PER VISUALIZZARE INFORMAZIONI

In conclusione, ecco una lista di informazioni che possono essere ottenute e i comandi necessari per visualizzarle.

Dati bibliografici e statistici del testo da cui un contesto è estratto

Cliccare, nella riga di testata del contesto, sul titolo abbreviato (verde) o sul simbolo "#" (blu scuro), se presente.

Brano associato al contesto (se presente)

Cliccare sul carattere "A" (marrone), se presente, nella riga di testata del contesto.

Dicitura completa del riferimento organico

Cliccare, nella riga di testata del contesto, in corrispondenza del riferimento organico o del riferimento topografico.

Dati completi del lemma (ed eventualmente dell'iperlemma) usato per lemmatizzare un'occorrenza

Cliccare, nella riga di testata del contesto, in corrispondenza dei dati del lemma o del simbolo "I".

Statistiche relative al numero di occorrenze lemmatizzate, non lemmatizzate e totali mostrate

Cliccare sull'icona "istogramma storto" in alto a destra. Cliccare di nuovo sull'icona per chiudere il riquadro **statistiche**.

Elenco delle forme attualmente associate, nel corpus, ad un lemma presente nel dizionario di macchina

Cliccare col tasto destra in un punto qualunque del dizionario di macchina e successivamente col tasto sinistro in un punto interno alla colonna **lemma** della riga del lemma desiderato.

Al termine, premere <Esc> per ripristinare il dizionario di macchina.

Elenco degli iperlemmi attualmente associati, nel corpus, ad un lemma presente nel dizionario di macchina

Cliccare col tasto destra in un punto qualunque del dizionario di macchina e successivamente col tasto sinistro in un punto esterno alla colonna **lemma** della riga del lemma desiderato.

Al termine, premere <Esc> per ripristinare il dizionario di macchina.

CAP. 11 - LEMMATIZZAZIONE PER SINGOLO TESTO

PROCEDURA DI LEMMATIZZAZIONE

Per avviare la lemmatizzazione è necessario prima di tutto entrare nella finestra introduttiva di tale *ambiente*, tramite [LEMMATIZZAZIONE + PER SINGOLO TESTO] nella finestra iniziale di GATTO, provvedendo poi ad aprire il corpus desiderato.

Questa modalità di lemmatizzazione opera a livello di singolo testo quindi preliminarmente andrà selezionato quest'ultimo tramite la sua sigla (menu [TESTO]). Il passo successivo dipende dalla modalità di lemmatizzazione scelta, *standard* o *sequenziale*. Nell'ipotesi di lemmatizzazione *standard* si dovrà selezionare la forma da lemmatizzare con [LEMMATIZZAZIONE + SELEZIONE FORMA]: a quel punto verranno rese accessibili alla lemmatizzazione tutte le occorrenze della forma scelta presenti nel testo indicato. L'utente dovrà posizionarsi sull'occorrenza desiderata ed associarle il lemma opportuno; potrà trattarsi di un

lemma già presente o meno nel lemmario del corpus: nel primo caso basterà selezionare il lemma tra quelli disponibili mentre nel secondo il lemma dovrà essere digitato ex novo, precisando anche la categoria grammaticale ed eventualmente il disambiguatore e il commento; in questo secondo caso al termine della operazione di lemmatizzazione il lemma verrà inserito nel lemmario del corpus e sarà disponibile per essere selezionato nell'ambito delle lemmatizzazioni immediatamente successive.

Terminata la lemmatizzazione di un'occorrenza si può procedere lemmatizzandone un'altra sempre relativa alla stessa forma oppure cambiando quest'ultima tramite passaggio alle forme alfabeticamente adiacenti o per digitazione diretta, sempre rimanendo nello stesso testo. Tornando alla finestra LEMMATIZZAZIONE PER SINGOLO TESTO (finestra iniziale dell'ambiente) si può cambiare testo e ripartire scegliendo nuovamente la forma con cui riprendere il lavoro di lemmatizzazione.

Se l'occorrenza che viene lemmatizzata lo era già, con altro lemma, la precedente lemmatizzazione viene rimossa automaticamente. È anche possibile eliminare semplicemente le lemmatizzazioni effettuate in precedenza.

Terminata una lemmatizzazione, GATTO si posiziona automaticamente sull'occorrenza successiva della stessa forma nello stesso testo, se presente, oppure comunica con un messaggio che non vi sono altre occorrenze.

LEMMATIZZAZIONE IN MODALITÀ STANDARD O SEQUENZIALE

La procedura descritta nel paragrafo precedente si basa sulla modalità *standard* di lemmatizzazione. Questa tecnica permette di passare in rassegna agevolmente in successione alfabetica tutte le forme presenti in un testo e per ciascuna tutte le occorrenze che vi compaiono. Si tratta di un meccanismo adatto anche nel caso si voglia lemmatizzare solo una specifica forma, magari limitatamente ad alcune occorrenze.

GATTO offre una seconda modalità di lemmatizzazione, uguale alla precedente per quanto attiene ai meccanismi di lemmatizzazione, ma diversa per l'ordine di presentazione delle occorrenze da lemmatizzare: si tratta della modalità sequenziale che, partendo dalla prima parola o da una parola qualunque del testo, consente di procedere alla lemmatizzazione delle parole che via via vi si incontrano, scorrendolo in avanti o indietro.

È possibile combinare le due modalità passando in ogni momento dall'una all'altra: si può ad esempio procedere con la lemmatizzazione sequenziale fino ad arrivare ad una forma della quale si ritenga opportuno lemmatizzare altre occorrenze, operazione effettuabile passando alla modalità standard; una volta ultimata questa operazione è sufficiente riportarsi all'occorrenza in corrispondenza della quale è stata abbandonata la modalità sequenziale, ripristinare quest'ultima e procedere lemmatizzando le successive parole del testo.

STRUTTURE DELLE TESTATE DELLA FINESTRA DI LEMMATIZZAZIONE E DEL CONTESTO



La testata della finestra di lemmatizzazione contiene una serie di informazioni disposte su tre righe e riferite a tutti i contesti sottostanti. Per quanto riguarda questi ultimi, ognuno di loro è formato da una riga di testata, con ulteriori informazioni specifiche del contesto e da due righe che visualizzano il contesto vero e proprio.

Prima riga di testata della finestra

Contiene, nell'ordine:

1. logo e nome del programma,
2. nome della finestra (LEMMATIZZAZIONE CORPUS),
3. disco e nome del corpus,
4. "testo" e sigla del testo in corso di lemmatizzazione.

Seconda riga di testata della finestra

Contiene, nell'ordine:

1. dicitura "**norm**" o "**seq**", in giallo su nero, per indicare la modalità di lemmatizzazione in uso, rispettivamente standard o sequenziale,
2. dicitura "**forme: tutte**", "**forme: non lem**" o "**forme: lemm**", in azzurro su nero, per indicare se le forme che vengono rese accessibili alla lemmatizzazione sono tutte quelle incluse nel testo, solo quelle delle quali nessuna occorrenza nel testo è stata lemmatizzata o quelle di cui almeno una occorrenza nel testo è stata già lemmatizzata,
3. dicitura "**occ: tutte**", "**occ: non lem**" o "**occ: lemm**", in verde su nero, per indicare se le occorrenze che vengono mostrate sono tutte quelle della forma in oggetto, solo quelle non ancora lemmatizzate o solo quelle già lemmatizzate,
4. sigla del testo e parte iniziale del titolo abbreviato, in bianco su blu: cliccandola si apre la finestra con i dati bibliografici e statistici del testo,
5. forma in corso di lemmatizzazione, in verde su bianco.

Terza riga di testata della finestra

Contiene, nell'ordine:

1. entrata lessicale e categoria grammaticale del **lemma impostato**, ovvero del lemma utilizzato nel corso dell'ultima operazione di lemmatizzazione, in blu su celeste; se in tale operazione è stato coinvolto anche un iperlemma questo compare nella stessa riga, separato dal lemma dal carattere "/",
2. icona che rappresenta un istogramma: se l'icona è storta vuol dire che il quadro **statistiche** non è visualizzato; se è dritta vuol dire che lo è; cliccandola si inverte la situazione,
3. icona "**NL**" o "**↓**" indicante il significato del tasto <Invio> (nel primo caso assegnazione di un nuovo lemma all'occorrenza corrente, nel secondo lemmatizzazione del contesto corrente mediante il *lemma impostato* e passaggio al contesto successivo),
4. icona **libro aperto** o **libro chiuso**: mostra lo stato del dizionario di macchina, rispettivamente aperto o chiuso; cliccandola si inverte la situazione.

Riga di testata del contesto

Può contenere, nell'ordine:

1. riquadro rosso: dove presente indica il *contesto corrente*, ovvero quello oggetto delle eventuali operazioni di lemmatizzazione,
2. numero dell'occorrenza della forma nel testo, in verde per l'occorrenza corrente, in blu per le altre,
3. simbolo "**A**" indicante la presenza di un brano associato al contesto (nota o traduzione), in marrone,
4. entrata lessicale del lemma associato all'occorrenza (se già lemmatizzata), in rosso,
5. categoria grammaticale del lemma associato all'occorrenza (se già lemmatizzata), in verde,
6. simbolo "**X**" indicante che il lemma associato all'occorrenza (se già lemmatizzata) ha un contenuto non nullo nel campo disambiguatore, in nero,
7. simbolo "**I**" indicante che anche un iperlemma è associato all'occorrenza (se già lemmatizzata), in nero,
8. riferimento organico, in blu,
9. riferimento topografico, in viola.

Cliccando l'elemento 3 viene mostrato il brano associato; cliccando gli elementi 4, 5, 6 o 7 viene mostrato il lemma associato, con eventuale iperlemma; cliccando gli elementi 8 o 9 si vede il riferimento organico per esteso.

LEZIONE 19. COME LEMMATIZZARE UN TESTO

Per partire, avviare GATTO e portarsi nell'ambiente **Lemmatizzazione per singolo testo** mediante [LEMMATIZZAZIONE - PER SINGOLO TESTO].

Tramite [CORPUS] aprire la finestra di selezione del corpus; cliccare sulla riga del corpus **Demo**, aprirlo con [APRI], quindi tornare alla finestra LEMMATIZZAZIONE PER SINGOLO TESTO con [CHIUDI FINESTRA].

Bisogna ora scegliere il testo da lemmatizzare, con [TESTO]: selezionare il testo **vad** e dare (OK). Se si sbaglia testo basta ripetere l'operazione e il nuovo testo sostituirà il precedente.

Verranno ora effettuate delle lemmatizzazioni utilizzando i vari metodi disponibili, ovvero:

1. tramite dizionario di macchina,
2. tramite tabella degli omografi,
3. tramite il *lemma impostato*,
4. per digitazione diretta.

Esempio 1: lemmatizzazione di un'occorrenza utilizzando il dizionario di macchina

Come primo esempio si lemmatizzerà un'occorrenza della forma **chuy** (si tratta di un testo veneziano). Cliccare su [LEMMATIZZAZIONE.. + SELEZIONE FORMA] e digitare **chuy** seguito da (OK). Cliccando sull'icona **libro chiuso** nella terza riga di testata, a destra, si aprirà il dizionario di macchina di **chuy**: l'unico lemma utilizzato all'interno dei testi del corpus per lemmatizzare **chuy** è **cui (rel.)**. Il testo **vad** contiene due occorrenze di **chuy**, la prima lemmatizzata con **cui (rel.)**, la seconda non lemmatizzata.

Supponiamo di voler lemmatizzare anche la seconda occorrenza con **cui (rel.)**.

Il quadratino rosso posto a sinistra del numero d'ordine del primo contesto indica che un'eventuale operazione di lemmatizzazione si applicherà a quel contesto; si dovrà quindi preliminarmente trasferire il quadratino sul secondo contesto, cliccando in un punto qualunque delle sue due righe di testo o a sinistra del numero d'ordine 2 (evitare, per il momento, di cliccare su altri punti della riga di testata del contesto). Come risultato, il quadratino si trasferisce sul secondo contesto, il cui numero d'ordine diventa di colore verde (mentre l'1 del precedente diventa blu): il secondo contesto è ora il nuovo *contesto corrente*, abilitato a essere oggetto di operazioni di lemmatizzazione.

A questo punto l'operazione di lemmatizzazione è semplicissima: cliccare una volta col tasto sinistro in un punto qualunque della riga relativa a **cui (rel.)** nel dizionario di macchina: il lemma scelto appare, in rosso, nella testata del contesto, ad indicare che l'occorrenza è stata lemmatizzata col lemma desiderato. Con (Menu) (in basso a destra) tornare alla finestra LEMMATIZZAZIONE PER SINGOLO TESTO.

Nota

Se lemmatizzando col dizionario di macchina e quindi cliccando su uno dei lemmi inclusi al suo interno venisse presentata la lista degli iperlemmi associati a tale lemma (se esistenti) per una ulteriore scelta tra essi, se non si è interessati alla cosa premere <Esc> per rifiutare la scelta e completare regolarmente la lemmatizzazione.

Esempio 2: altra lemmatizzazione di un'occorrenza utilizzando il dizionario di macchina

Selezionare come testo **od** e come forma da lemmatizzare **morto**, tramite [LEMMATIZZAZIONE.. + SELEZIONE FORMA]; portarsi sul contesto n. 7; aprire il dizionario di macchina cliccando su **libro chiuso**; questa volta vengono proposti 3 lemmi diversi, il primo dei quali è presente anche nel riquadro **lemma impostato** in alto; il lemma voluto non è però quello suggerito come lemma impostato, ovvero **morto (agg.)**, ma bensì **morire (v.)**. Cliccando in un punto qualunque della riga relativa a questo lemma si ottiene il duplice effetto di lemmatizzare l'occorrenza con **morire (v.)** e di sostituire con questo lemma il lemma impostato precedente (l'utilità di questo cambiamento verrà vista in seguito). Chiudere la finestra con (Menu).

Esempio 3: lemmatizzazione di un'occorrenza utilizzando la tabella degli omografi

Selezionare come testo **vad** e come forma da lemmatizzare **mitade** e aprire il dizionario di macchina. In questo caso la forma **mitade** risulta non essere stata mai lemmatizzata nel corpus **Demo**; si dovrà quindi procedere digitando il lemma direttamente. Cliccare su (N. lem/omo/ipe) e digitare **metà** seguito da (OK). Un lemma con entrata lessicale **metà** esiste già nel lemmario del corpus e viene quindi proposto come omografo disponibile. Si può vedere che il colore di fondo della tabella *omografi disponibili*, giallo, è diverso da quello del *dizionario di macchina*, celeste; ciò è fatto per evitare di confondere le due tabelle.

Volendo semplicemente interrompere l'operazione a questo punto senza effettuare la lemmatizzazione e ripristinando la tabella dei lemmi premere <Esc>; se invece, come è il caso, si tratta proprio del lemma cercato, per completare la lemmatizzazione basta cliccare sulla riga di **metà (s.f.)**.

Notare che a questo punto il dizionario di macchina di **mitade** non è più vuoto ma contiene il lemma **metà (s.f.)**: questo dato sarà d'ora in poi presentato lemmatizzando **mitade** in questo o in qualunque altro testo del corpus **Demo**.

Chiudere la finestra con (Menu).

Nota

Se lemmatizzando con la tabella degli omografi e quindi cliccando su uno dei lemmi inclusi al suo interno venisse presentata la lista degli iperlemmi associati a tale lemma (se esistenti) per una ulteriore scelta tra essi, se non si è interessati alla cosa premere <Esc> per rifiutare la scelta e completare regolarmente la lemmatizzazione.

Esempio 4: lemmatizzazione di un'occorrenza utilizzando il lemma impostato

Selezionare ora la forma **parte** e aprire, al solito, il dizionario di macchina. Il lemma impostato risulta essere **parte (s.f.)**. Si supponga di voler lemmatizzare la terza occorrenza di **parte** proprio col lemma impostato.

Preliminarmente bisogna, come fatto in precedenza, definire il terzo contesto come *contesto corrente*, cliccando su un punto qualunque delle due righe di testo o a sinistra del suo numero d'ordine "3"; un altro modo per muoversi tra i contesti è quello di scorrerli usando i tasti <↓> e <↑>.

Una volta trasformato il terzo contesto in contesto corrente, cliccare in un punto delle due righe di testo: l'occorrenza verrà lemmatizzata col lemma impostato.

Riassumendo le cose viste:

- cliccando su un contesto diverso dal contesto corrente, lo si trasforma nel contesto corrente,
- cliccando sul contesto corrente lo si lemmatizza col lemma impostato, se presente.

Come dopo ogni lemmatizzazione, indipendentemente dalla modalità utilizzata per effettuarla, il successivo contesto della forma, se esiste, diviene automaticamente il nuovo contesto corrente.

Chiudere la finestra con **(Menu)**.

Esempio 5: lemmatizzazione di un'occorrenza per digitazione diretta del lemma

Selezionare la forma **mey** e aprire il dizionario di macchina, che risulta vuoto in quanto **mey** non è mai stata lemmatizzata in **Demo**.

Il lemma da associare all'unica occorrenza di **mey** è, evidentemente, **mio (poss.)**. Come nell'esempio 3 cliccare su **(N. lem/omo/ipe)**, digitare **mio** e quindi cliccare su **(OK)**. A differenza dell'esempio 3 questa volta il lemma non solo non è stato associato in precedenza alla forma **mey** ma addirittura non compare nel lemmario di **Demo**: GATTO chiede quindi i dati ulteriori che definiscono il nuovo lemma per inserirlo nel lemmario. L'unico dato obbligatorio è la categoria grammaticale, nel caso in questione "**poss.**"; digitare l'informazione, trascurare gli ulteriori campi e cliccare su **(OK)** (un altro modo per lasciare vuoti i campi successivi e chiudere la finestra è battere <Invio> per ognuno dei campi sottostanti).

A questo punto si sono ottenuti i risultati seguenti:

- l'occorrenza di **mey** è stata lemmatizzata con **mio (poss.)**,
- il lemma **mio (poss.)** è stato inserito nel dizionario di macchina di **mey**,
- il lemma **mio (poss.)** è stato inserito nel riquadro *lemma impostato*,
- il lemma **mio (poss.)** è stato inserito nel lemmario del corpus **Demo** e associato alla forma **mey**.

Chiudere la finestra con **(Menu)**.

Esempio 6: lemmatizzazione di un'occorrenza per digitazione diretta del lemma e rifiuto di omografi presenti nel dizionario di macchina della forma

Nella finestra LEMMATIZZAZIONE PER SINGOLO TESTO con [TESTO] selezionare il testo **in** e poi indicare come parola da lemmatizzare **omnipotente**. Aprire il dizionario di macchina.

Usando il tasto <↓> portarsi al contesto numero 10, il quale si trova nella schermata successiva. In questo contesto **Onnipotente** è usato come sostantivo maschile. Il dizionario di macchina, se aperto, non mostra il lemma voluto.

Procedere con **(N. lem/omo/ipe)**, digitare **omnipotente** e cliccare su **(OK)**; come era da aspettarsi si apre la *tabella degli omografi*, prospettando come possibile lemma **omnipotente (agg.)**, che non è quello desiderato. Per forzare GATTO ad inserire un nuovo lemma (cliccando su **omnipotente (agg.)** il lemma non voluto viene accettato, pigiando <Esc> l'intera operazione viene annullata) cliccare una seconda volta su **(N. lem/omo/ipe)** e questa volta verranno richiesti gli ulteriori dati del nuovo elemento da introdurre nel lemmario. Digitare la categoria grammaticale **s.m.** e premere **(OK)**: la lemmatizzazione con relativo inserimento del nuovo lemma nel lemmario va così a buon fine.

Esempio 7: lemmatizzazione di un'occorrenza per digitazione diretta del lemma e rifiuto di omografi non presenti nel dizionario di macchina della forma

Questa situazione si verifica quando nel corpus esistono lemmi omografi di quello che si digita, non applicati però alla forma in corso di lemmatizzazione.

Con [TESTO] selezionare il testo **d3** e poi indicare come parola da lemmatizzare **defunto**. Aprire il dizionario di macchina, che risulta vuoto.

Procedere a lemmatizzare l'unico contesto presente con (**N. lem/omo/ipe**), digitare **defunto** e cliccare su (**OK**); si apre la *tabella degli omografi*, prospettando come possibile lemma **defunto (s.m.)**, che non è quello desiderato, in quanto nel caso dell'occorrenza considerata il lemma corretto è **defunto (agg.)**. Per forzare GATTO ad inserire il nuovo lemma cliccare una seconda volta su (**N. lem/omo/ipe**) e questa volta verranno richiesti gli ulteriori dati del nuovo elemento da introdurre nel lemmario. Digitare la categoria grammaticale **agg.** e premere (**OK**): la lemmatizzazione con relativo inserimento del nuovo lemma nel lemmario va così a buon fine. il nuovo lemma viene inserito nel dizionario di macchina della forma **defunto**.

LEZIONE 20. COME LEMMATIZZARE CON LA FINESTRA A CONTESTO SINGOLO O IN MODALITÀ SEQUENZIALE

USO DELLA FINESTRA A CONTESTO SINGOLO

Nel corso della Lezione precedente sono stati mostrati diversi esempi di lemmatizzazione, tutti effettuati in modalità standard utilizzando la finestra a contesti multipli, la quale visualizza un certo numero di contesti assegnando a ciascuno due righe dello schermo.

Qualora le caratteristiche del testo in generale o di uno specifico contesto in particolare richiedessero la lettura di una parte di testo più estesa è possibile passare alla finestra a contesto singolo che, come suggerisce la dicitura, mostra un solo contesto per volta ma con la possibilità di ampliarne le dimensioni.

Per vedere un esempio di questa visualizzazione impostare la lemmatizzazione della forma **peraviso** nel testo **btb**. Sono presenti due occorrenze. Si tratta di un brano in versi e la stringa " / " indica la separazione tra due versi successivi. Per passare ora alla finestra a contesto singolo cliccare su (**Occ. singola**) in basso. A questo punto lo stesso contesto viene mostrato da solo; la sua riga di testata è invariata mentre il corpo del contesto viene rappresentato riproducendo la struttura del testo in versi; la lunghezza del contesto è pari a 3 periodi e la parola da lemmatizzare compare in quello centrale. Volendo estendere il contesto usare (**Allarga**): ad ogni clic il contesto viene allungato di due periodi, con l'aggiunta di uno in testa e di uno in coda, fino a raggiungere i limiti del testo associato al riferimento organico precedente. Per ridurre le dimensioni del contesto usare (**Restringi**), con effetti opposti: il valore limite inferiore per la lunghezza del contesto è un periodo. In tutti i casi l'estensione del contesto è comunque limitata alla parte di testo associata al riferimento organico immediatamente precedente.

Volendo passare al contesto successivo utilizzare il tasto <↓> come nel caso dei contesti multipli: la dimensione del contesto in periodi rimane quella precedentemente impostata. Effettuare il passaggio al contesto numero 2.

Tutti gli altri comandi sono identici al caso della visualizzazione a contesti multipli. Per tornare a quest'ultima usare (**Occ. multiple**): in questo caso la finestra a contesti multipli contiene inizialmente il solo contesto numero 2 in quanto viene aperta a partire dal contesto visualizzato nella finestra *a contesto singolo*. Per far riapparire anche il primo contesto risalire di una posizione con <↑>.

LEMMATIZZAZIONE IN MODALITÀ SEQUENZIALE

La lemmatizzazione sequenziale è a prima vista identica a quella standard a contesto singolo: viene mostrato un contesto di 3 periodi, estensibile o riducibile con i tasti (**Allarga**) e (**Restringi**) fino al limite del riferimento organico. Ciò che cambia è il meccanismo di selezione di forme e occorrenze, definito in questo caso dall'ordine di comparizione delle parole nel testo.

Dalla finestra LEMMATIZZAZIONE PER SINGOLO TESTO si apre il testo **vad** e poi si attiva la lemmatizzazione sequenziale tramite [LEMMATIZZAZIONE.. + MODALITÀ SEQUENZIALE]; si apre anche il dizionario di macchina. Il fatto che si sta operando in modalità sequenziale è segnalato dalla scritta **seq** in giallo su nero all'inizio della seconda riga di testata (la sigla per la modalità standard è **norm**). L'occorrenza evidenziata in rosso è *Allo*, prima parola del testo, la quale non è lemmatizzata; anzi alla forma **allo** non è associato nessun lemma, come si evince dal dizionario di macchina vuoto; il numero d'ordine **1** in verde nella

riga di testata non si riferisce al numero della parola nel testo ma, come sempre, indica che si sta vedendo la prima occorrenza della forma **allo** nel testo.

Volendo lemmatizzare *allo* si può procedere come nei casi precedenti: cliccare su (**N. lem/omo/ipe**), digitare **a** e cliccare su (**OK**), quindi cliccare sull'omografo **a (prep.)**, che verrà aggiunto al dizionario di macchina di **allo**. In questo caso al termine della lemmatizzazione il programma rimane posizionato sul contesto lemmatizzato, anziché passare al successivo.

Per passare alla parola successiva usare <→>: verrà evidenziata la parola *nome*, altrove già lemmatizzata con **nome (s.m.)**; si tratta in particolare della prima occorrenza della forma **nome**, come evidenzia l'1 riportato in alto a sinistra.

Si può navigare nel testo spostandosi con i tasti <←> e <→>. Quando la nuova occorrenza fuoriesce dal periodo in cui si trovava la precedente cambia la parte di testo mostrata e si ha uno scorrimento di un periodo, in modo da mantenere sempre la parola evidenziata entro il periodo collocato in posizione centrale tra quelli visualizzati. Per meglio vedere questo meccanismo abbandonare il testo **vad** (caratterizzato da periodi iniziali piuttosto lunghi e quindi scomodi per questa prova) e aprire il testo **ddt**. Anche in questo caso attivare la lemmatizzazione sequenziale e riaprire il dizionario di macchina. La parola evidenziata è *va*, inclusa nel periodo "[...] *va cercando*." Inizialmente i periodi mostrati sono 2, in quanto manca per forza la frase precedente a quella che contiene la parola evidenziata, prima parola del testo. La situazione non cambia spostandosi su *cercando* con <→>; il successivo spostamento a destra porterà sulla prima parola del secondo periodo del testo (*Al*): si vedrà allora apparire un terzo periodo in coda ai precedenti, portando con ciò la visualizzazione a regime, ovvero:

- 1° periodo: [...] *va cercando*. frase precedente a quella che include la parola evidenziata,
- 2° periodo: *Al nome...Giesù gloriando*. frase che include la parola evidenziata,
- 3° periodo: *Or che...va giubilando*. frase successiva a quella che include la parola evidenziata.

La situazione si mantiene inalterata spostandosi sulle parole successive fino a *gloriando*. Passando a *Or* ha luogo uno scorrimento dei periodi, con esclusione del primo ed inserimento di uno nuovo, il quarto del testo.

Durante questa prova, scorrendo le parole incluse nel secondo periodo capita alcune volte di veder apparire il simbolo "A" in marrone accanto al numero dell'occorrenza: ciò accade ad esempio in corrispondenza della parola *ave*. In tal caso, cliccando su "A", si aprirà una finestra che evidenzia una nota inclusa nell'apparato critico e relativa alla parola in questione, ovvero *ave (u.f.)] ave (ed.)*. Per chiudere la nota cliccare sul simbolo **x** o usare <Esc>.

Un tipo di informazione che può risultare utile lemmatizzando a contesto singolo o in modalità sequenziale, ovvero in tutti i casi in cui è visibile un solo contesto per volta, è dato dal numero di occorrenze della forma evidenziata presenti nel testo, lemmatizzate e non. Per attivarla cliccare su (**Statistiche**) o sull'icona "istogramma storto". Portandosi ad esempio sulla parola *cor* (6° rigo del testo **ddt**) si vede che si è in corrispondenza della 1ª occorrenza nel testo, in accordo con quanto riportato nella riga di testata, su 2 occorrenze presenti, nessuna delle quali lemmatizzata.

Supponiamo ora di voler lemmatizzare sia l'occorrenza di **cor** evidenziata che la successiva, senza abbandonare la modalità di lemmatizzazione sequenziale. Per fare questo il modo più semplice è cominciare col cliccare su un punto qualunque del contesto per lemmatizzare l'occorrenza corrente sulla base del lemma impostato, ovvero **cuore (s.m.)**; il riquadro **statistiche** mostra ora che una delle due occorrenze di **cor** nel testo è lemmatizzata. Usare ora il tasto <↓> per passare alla seconda occorrenza della stessa forma, come nella lemmatizzazione standard; la nuova situazione viene evidenziata sia dal **2** nella riga di testata che dal riquadro **statistiche**. La nuova occorrenza di **cor** è inclusa nella frase *Fra tuctca la gente lo mio cor si muta...* Lemmatizzare anche questa occorrenza col lemma impostato, cliccando in un punto qualunque del contesto. A questo punto col tasto <↑> si può tornare alla precedente occorrenza di **cor** e da lì riprendere la lemmatizzazione sequenziale interrotta, ripartendo dalla parola *meo* (*Or che fai cor meo, che non prendi...*), tramite i tasti <←> e <→>.

LEZIONE 21. OPZIONI DI LEMMATIZZAZIONE (TIPO DI FORME, TIPO DI OCCORRENZE) E COMANDI DI SPOSTAMENTO TRA OCCORRENZE DELLA STESSA FORMA E TRA FORME DIVERSE

RESTRIZIONI NELL'ACCESSO ALLE FORME DI UN TESTO SULLA BASE DEL LORO STATO DI LEMMATIZZAZIONE

Gli esempi considerati finora non fissavano alcun limite di accesso a forme e occorrenze incluse nel testo scelto, permettendo di muoversi liberamente al suo interno in fase di lemmatizzazione.

Talvolta è più pratico, in funzione di specifiche esigenze, poter limitare l'accesso alle sole forme mai lemmatizzate nell'ambito del testo in questione o che al contrario presentino al suo interno almeno un'occorrenza lemmatizzata.

Per capire meglio il funzionamento di questo meccanismo di controllo aprire in lemmatizzazione il testo **in** e portarsi sulla forma **amurj**. Come si può vedere la prima occorrenza di questa forma è lemmatizzata, mentre la seconda non lo è.

Uno dei modi per passare ad un'altra forma inclusa nel testo si basa sull'uso dei tasti <=> e <=>. Questi comandi erano stati già incontrati in modalità sequenziale e lì avevano la funzione di far accedere alle parole adiacenti nel testo; in lemmatizzazione standard gli stessi comandi permettono invece di passare alle forme alfabeticamente adiacenti incluse nel testo. Utilizzando in questo caso il comando <=> più volte si passa nell'ordine alle forme **amustrassi**, **amustratu**, **anania**, **ananti**. Se si attiva il quadro **statistiche** con (**Statistiche**) o con l'icona "istogramma storto" si nota che **ananti** presenta 4 occorrenze lemmatizzate su 14; retrocedendo alfabeticamente con <=> a **anania** si vede che questa forma non presenta invece occorrenze lemmatizzate, a differenza di **amustratu** (una su una), **amustrassi** (una su una) e **amurj** (una su due).

La scansione di queste forme ha avuto luogo quindi a prescindere dal loro stato di lemmatizzazione. Va ricordato che una forma si definisce *lemmatizzata* in un testo quando almeno una delle occorrenze incluse in esso lo è, *non lemmatizzata* nel caso opposto. Trovandosi ora sulla forma **amurj** cliccare su (**Tipo forme**) in basso: i contesti scompaiono; cliccare una seconda volta. I contesti ricompaiono con la differenza che nella seconda riga di testata la dicitura (azzurra su sfondo nero) **forme: tutte** è stata sostituita da **forme: lemm**. Questo significa che la scansione del testo verrà limitata alle sole forme già lemmatizzate. Per prova, usare come prima ripetutamente <=> per raggiungere la forma **ananti**: si noterà che questa volta vengono mostrate nell'ordine **amustrassi**, **amustratu** e **ananti** ma non **anania**. Impostando "tipo forme" a "lemmatizzate" si è detto al programma di limitare appunto la visualizzazione alle forme già lemmatizzate, mostrandone comunque tutte le occorrenze, lemmatizzate o meno. Sempre restando su **ananti** cliccare due volte (non il doppio clic!) su (**Tipo forme**), in modo da far apparire nella seconda riga di testata la dicitura (azzurra su sfondo nero) **forme:non lem**. In questo modo abbiamo fatto la scelta opposta rispetto alla precedente, chiedendo di mostrare solo le forme prive di occorrenze lemmatizzate nel testo considerato. Come si può notare i contesti di **ananti** e l'indicazione **ananti** in testata sono scomparsi e la maggior parte dei bottoni in basso ha le scritte in nero, per indicare che sono stati disabilitati: il motivo risiede nel fatto che la forma corrente (**ananti**) non soddisfa più il vincolo posto, in quanto è lemmatizzata, mentre è stato richiesto di visualizzare solo le forme non lemmatizzate (cliccando ancora su (**Tipo forme**) così da riportarsi su **forme: lemm**, **ananti** ricomparirebbe); in questa situazione la finestra di lemmatizzazione viene congelata e rimangono attivi solo i comandi che consentono di portarsi ad altre forme o di tornare alla finestra LEMMATIZZAZIONE PER SINGOLO TESTO. Spostandosi ora alla forma alfabeticamente precedente mediante <=> si ripresenta la forma **anania**, per l'appunto non lemmatizzata; ripetendo l'operazione si arriva ad **amoni**, anch'essa non lemmatizzata.

Gli spostamenti alle forme *adiacenti alfabeticamente* vanno perciò intesi nel senso più generale di spostamenti alle forme *alfabeticamente più prossime, compatibilmente con l'eventuale vincolo imposto sullo stato di lemmatizzazione*.

Un altro modo per passare a lemmatizzare una diversa forma del testo fa uso del bottone (**Forma**). Cliccandolo viene richiesto di digitare la forma desiderata. Trovandosi ancora su **amoni**, con il "tipo forme" impostato su "forme non lemmatizzate", cliccando su (**Forma**), digitando **ananti** e dando (**OK**) si otterrà in risposta il messaggio **forma non visualizzabile nella modalità forme: non lem**: questo è corretto, in quanto **ananti** è lemmatizzata e quindi esclusa dall'impostazione attuale del tipo di forme. Cliccando allora su (**Tipo forme**) per impostare il tipo di forme "lemmatizzate" si vedrà sparire la forma **amoni** e i suoi contesti; per contro usando ora (**Forma**) per accedere a **ananti** l'operazione andrà a buon fine e verranno visualizzate le prime 8 occorrenze (su 14, verificare in **statistiche**) di questa forma.

RESTRIZIONI NELL'ACCESSO ALLE OCCORRENZE DI UNA FORMA DI UN TESTO SULLA BASE DEL LORO STATO DI LEMMATIZZAZIONE

GATTO fornisce una seconda opzione di limitazione dei contesti, questa volta basata sullo stato di lemmatizzazione dei singoli contesti e non della forma di appartenenza.

Con procedure analoghe a quelle illustrate nel paragrafo precedente si può utilizzare il bottone (**Tipo occorr.**); l'impostazione ottenuta sarà ancora visibile nella seconda riga di testata (diciture **occ:tutte**, **occ: non lem**, **occ: lemm** in verde su sfondo nero).

Per analizzare il comportamento di questa seconda opzione, sempre restando nel testo **in** portarsi sulla forma **andatu**, controllando che il tipo di forme sia impostato a **forme: tutte**. Attivare anche il riquadro **statistiche**. Si vede che **andatu** ha 6 occorrenze, delle quali due lemmatizzate (la 1 e la 5) e quattro no (la 2, la 3, la 4 e la 6).

Cliccando su (**Tipo occorr.**) si limita la visualizzazione alle sole occorrenze non lemmatizzate (confrontare la scritta in testata): notare che i numeri d'ordine delle occorrenze non sono compresi tra 1 e 4 ma sono quelli riferiti al gruppo completo di 6 occorrenze. Cliccando un'altra volta su (**Tipo occorr.**) la situazione si inverte e vengono mostrate solo le occorrenze numero 1 e 5, lemmatizzate. Un terzo clic sul bottone ripristina la situazione di partenza.

Cliccare di nuovo su (**Tipo occorr.**) per limitare di nuovo la visualizzazione alle occorrenze non lemmatizzate. Agendo due volte sul tasto <-> portarsi in successione sulle forme **andau** e **andaulo**. Quest'ultima viene regolarmente rappresentata (vedere dicitura verde su fondo bianco nella seconda riga di testata, a destra), però senza alcun contesto. La spiegazione del fatto è ricavabile dal riquadro statistiche da cui si evince che la forma consta di una sola occorrenza, lemmatizzata e quindi esclusa dalla visualizzazione essendo il "tipo di occorrenze" impostato su "non lemmatizzate".

Ci si potrebbe chiedere per quale motivo la forma **andaulo** sia stata visualizzata lo stesso. Il motivo è che l'opzione "tipo di occorrenze" non agisce nella selezione delle forme da mostrare, selezione delegata all'altra opzione "tipo di forme" che al momento è impostata su "tutte" e che è l'unica abilitata ad escludere completamente delle forme dalla visualizzazione.

È importante quindi aver chiaro che le due opzioni agiscono in maniera combinata, per ottenere effetti del tipo "visualizzare tutte le occorrenze non ancora lemmatizzate delle forme già lemmatizzate parzialmente"; la combinazione delle due opzioni offre perciò uno strumento piuttosto potente ma che va usato con attenzione per non trovarsi di fronte a comportamenti a prima vista inspiegabili.

Per concludere, va notato che entrambe le opzioni si applicano anche al caso di lemmatizzazione sequenziale, per cui il passaggio alle parole precedenti e successive si effettua rispettando comunque i vincoli imposti dalle due opzioni; se questi comportano una distanza in numero di parole superiore ad un certo valore prefissato, ma modificabile agendo sulle opzioni generali di GATTO (vedere Appendice B), la nuova parola non verrà raggiunta e al suo posto verrà presentato un messaggio di avvertimento.

COMANDI UTILIZZABILI PER IL PASSAGGIO AD ALTRE OCCORRENZE DELLA STESSA FORMA

Avendo ultimato la panoramica sulle operazioni di lemmatizzazione vere e proprie è opportuno riassumere e completare il quadro dei comandi disponibili per passare da un'occorrenza ad un'altra, sempre restando nell'ambito della stessa forma. I passaggi ad un determinato contesto sono effettuati compatibilmente con l'eventuale vincolo imposto sul suo stato di lemmatizzazione (opzione "tipo di occorrenza").

clik sul contesto

Se il nuovo contesto è visibile è possibile trasformarlo in contesto corrente cliccando in un suo punto qualunque, esclusa la riga di testata.

<↓> e <↑>

Passaggio al contesto successivo o precedente.

<Pag↓> e <Pag↑> (finestra a contesti multipli)

Passaggio alla schermata successiva o precedente.

<↖>

Passaggio al primo contesto.

(Vai a occ. n.)

Accesso diretto a una occorrenza definita in base al numero d'ordine.

COMANDI UTILIZZABILI PER IL PASSAGGIO AD ALTRE FORME

<→> e <↔> (modalità standard)

Passaggio alla forma alfabeticamente successiva o precedente compatibile con l'eventuale vincolo imposto al suo stato di lemmatizzazione (opzione "tipo di forma").

<→> e <↔> (modalità sequenziale)

Passaggio alla parola di testo successiva o precedente compatibile con la combinazione dei due vincoli "tipo di forme" e "tipo di occorrenze", con limite di distanza per la ricerca (vedere l'Appendice B - *Impostazione delle opzioni generali*).

(Forma)

Passaggio alla forma digitata compatibilmente con l'eventuale vincolo imposto al suo stato di lemmatizzazione (opzione "tipo di forma").

LEZIONE 22. FUNZIONI ACCESSORIE E RASSEGNA DEI COMANDI PER VISUALIZZARE INFORMAZIONI

Le Lezioni precedenti hanno descritto l'insieme delle operazioni di lemmatizzazione previste in **Lemmatizzazione per singolo testo**. Rimangono da passare rapidamente in rassegna alcune funzioni accessorie che completano il quadro delle azioni effettuabili.

Saranno anche elencati i comandi destinati ad ottenere informazioni di varia natura, alcuni dei quali sono stati già incontrati nel corso delle Lezioni precedenti.

FUNZIONI ACCESSORIE

Sostituzione di una lemmatizzazione preesistente

Questa situazione si verifica quando ci si accinge a rilemmatizzare un'occorrenza già lemmatizzata con un lemma diverso. Non c'è nessuna azione speciale da compiere: è sufficiente procedere con la normale procedura di lemmatizzazione e la lemmatizzazione precedente verrà rimossa automaticamente e sostituita.

Cancellazione di una lemmatizzazione preesistente

Portarsi sul contesto in oggetto trasformandolo quindi nel *contesto corrente*, poi cliccare su **(Canc. lemma)**. Fare attenzione al contesto scelto poiché l'operazione non richiede conferma e non è annullabile.

Accesso all'ultima lemmatizzazione effettuata in un testo

Questa funzione agevola la continuazione di operazioni di lemmatizzazione sistematica interrotte al termine di una sessione in quanto consente di riprendere l'attività laddove si è interrotta.

Come esempio, si consideri l'ultima lemmatizzazione effettuata nel corso di queste Lezioni: si tratta del lemma **a (prep.)** attribuito all'unica occorrenza della forma **allo** nel testo **vad** (vedere Lemmatizzazione in modalità sequenziale - Lezione 20). Aprire nuovamente il testo **vad** dalla finestra LEMMATIZZAZIONE PER SINGOLO TESTO e selezionare come forma da lemmatizzare, ad esempio, **che**. Navigare un po' tra le occorrenze di **che** e portarsi, volendolo, su altre forme dello stesso testo senza però effettuare nessuna lemmatizzazione. A questo punto uscire dal testo con **(Menu)**. Ora utilizzare il menu [LEMMATIZZAZIONE.. + DALL'ULTIMA LEMMATIZZAZIONE]: a prescindere da tutte le operazioni successivamente effettuate su quel testo o su altri, ci si troverà posizionati proprio in corrispondenza dell'occorrenza di **allo**, ultima occorrenza lemmatizzata in ordine di tempo nel testo **vad**.

Ridefinizione del lemma impostato a partire da un contesto

Il *lemma impostato* viene automaticamente riassegnato ogni volta che è stata effettuata una lemmatizzazione o ci si è portati col dizionario di macchina aperto su una nuova forma.

Esiste un altro modo, questa volta controllato dall'utente, per assegnare il *lemma impostato*.

Aprire il testo **in**, selezionare la forma **ananti** e aprire il dizionario di macchina. Come risultato il lemma impostato risulterà essere **annanti (avv.)**.

Supponiamo di volerlo sostituire con **annantiché (cong.)** usato per lemmatizzare il contesto numero 4. Prima trasformare il contesto numero 4 nel contesto corrente con <↓> o cliccandovi sopra, quindi cliccare sul contesto col tasto destra del mouse: il lemma usato per lemmatizzare il contesto verrà copiato nel riquadro del lemma impostato e da questo momento sarà utilizzabile per successive lemmatizzazioni. Si può usare questa tecnica per agevolare la lemmatizzazione di alcune occorrenze di una forma sulla base di lemmatizzazioni precedentemente applicate ad altre.

Tasto <Enter> per attivare l'inserimento di un nuovo lemma

Normalmente per assegnare un nuovo lemma ad un'occorrenza si fa uso del bottone (**N. lem/omo/ipe**), attivabile cliccandolo o tramite <Alt> + <n>. Nella terza riga di testata della finestra esiste sulla destra un'icona che può mostrare il simbolo "NL" o "↓". Se il simbolo è "NL" il tasto <Enter> acquista il significato di "Nuovo lemma". Per invertire l'icona cliccarvi sopra.

Lemmatizzazione veloce di occorrenze consecutive utilizzando lo stesso lemma

Volendo lemmatizzare diverse occorrenze consecutive della stessa forma facendo uso dello stesso lemma, trasferire nel riquadro *lemma impostato* il lemma voluto (lemmatizzando la prima delle occorrenze con tale lemma o trasferendolo da un'occorrenza alla quale sia stato già associato - vedere sopra "Ridefinizione del lemma impostato a partire da un contesto") e poi lemmatizzare le occorrenze cliccandovi sopra in successione: poiché un clic sul contesto corrente lo lemmatizza a partire dal lemma impostato e definisce automaticamente il contesto successivo come nuovo contesto corrente l'operazione è piuttosto veloce.

Se, cliccandovi sopra, si fa in modo che l'icona posta sulla destra nella terza riga di testata evidenzi il simbolo "↓", anziché "NL", la stessa operazione potrà realizzarsi definendo al solito come contesto corrente il primo contesto da lemmatizzare e poi digitando <Enter> per ogni contesto successivo che si intende lemmatizzare sempre a partire dal lemma impostato.

COMANDI PER VISUALIZZARE INFORMAZIONI

In conclusione, ecco una lista di informazioni che possono essere ottenute operando col modulo di lemmatizzazione e i comandi necessari per visualizzarle.

Dati bibliografici e statistici del testo considerato

Cliccare, al centro della seconda riga di testata, il titolo abbreviato riportato parzialmente scritto in bianco su blu.

Brano associato al contesto (quando presente)

Cliccare sul carattere "A" scritto in marrone nella riga di testata del contesto, immediatamente a destra del numero d'ordine.

Dati completi del lemma (ed eventualmente dell'iperlemma) usato per lemmatizzare un'occorrenza

Cliccare, nella riga di testata del contesto, in corrispondenza dei dati del lemma con cui risulta lemmatizzato (se lo è).

Riferimento organico completo

Cliccare, nella riga di testata del contesto, in corrispondenza del riferimento organico o del riferimento topografico.

Statistiche relative al numero di occorrenze lemmatizzate, non lemmatizzate e totali della forma corrente nel testo considerato

Cliccare su (**Statistiche**) o sull'icona "istogramma storto" nella terza riga di testata. Cliccare di nuovo sul bottone o sull'icona volendo richiudere il riquadro **statistiche**.

Elenco delle forme attualmente associate, nel corpus, ad un lemma presente nel dizionario di macchina

Cliccare col tasto destro in un punto qualunque del dizionario di macchina e successivamente col tasto sinistro in un punto interno alla colonna **lemma** della riga del lemma desiderato.

Al termine, premere <Esc> per ripristinare il dizionario di macchina.

Elenco degli iperlemmi attualmente associati, nel corpus, ad un lemma presente nel dizionario di macchina

Cliccare col tasto destro in un punto qualunque del dizionario di macchina e successivamente col tasto sinistro in un punto esterno alla colonna **lemma** della riga del lemma desiderato.

Al termine, premere <Esc> per ripristinare il dizionario di macchina.

CAP. 12- INSERIMENTO DI IPERLEMMI

La natura degli iperlemmi è stata illustrata in precedenza (Introduzione e capitolo 9), per cui si rimanda a quelle note per quanto attiene la loro descrizione.

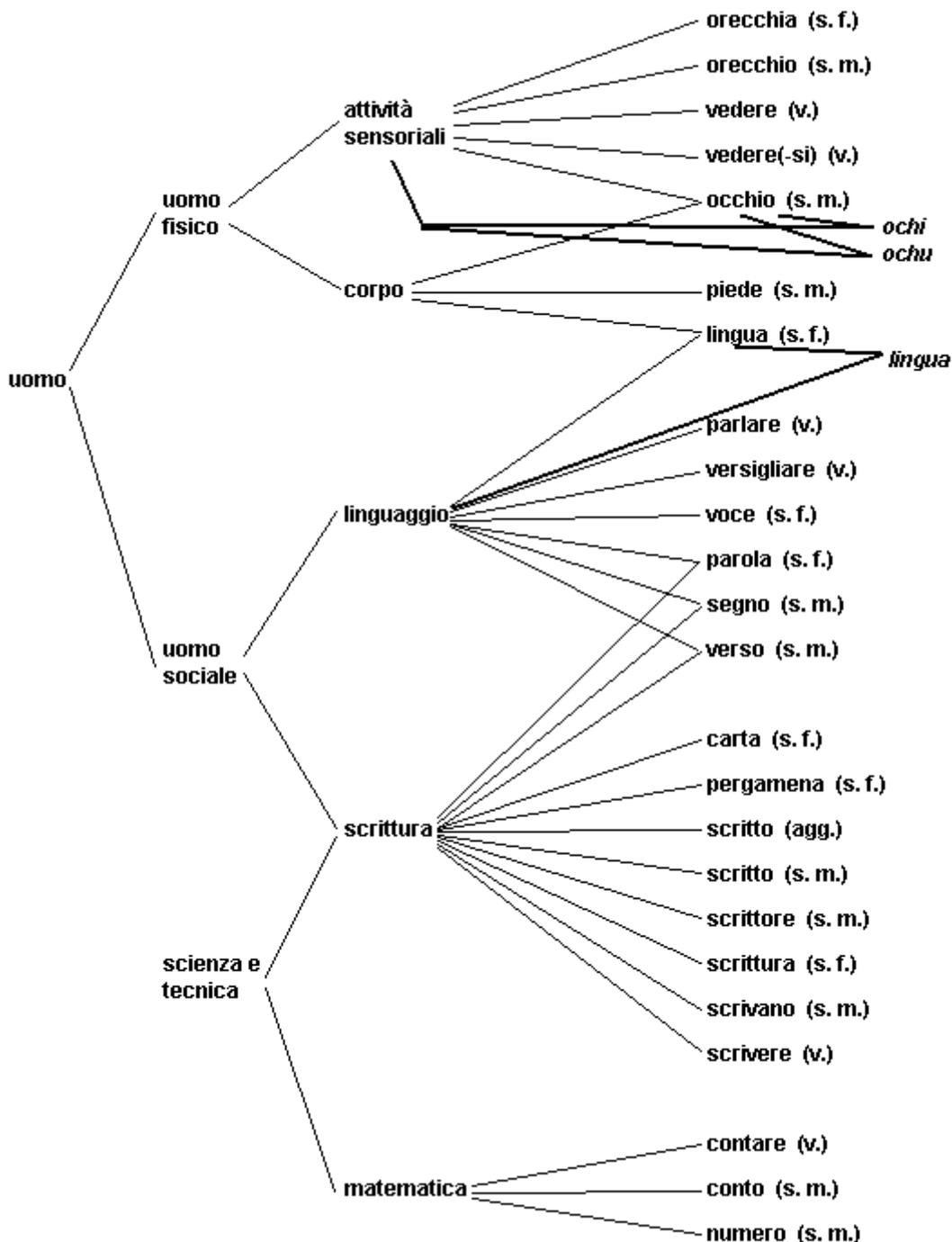
Va ricordato che gli iperlemmi possono operare in due modi distinti:

1. come elementi che raggruppano, secondo liberi criteri, lemmi o altri iperlemmi collocati in posizione *sottostante* nella struttura piramidale in cui gli iperlemmi stessi sono organizzati, e nella quale i lemmi possono essere pensati come costituenti il livello di base,
2. come puntatori a specifiche occorrenze lemmatizzate, alle quali risulta quindi associata una coppia lemma più iperlemma.

Come detto in precedenza, gli iperlemmi del primo gruppo possono appartenere ad un *livello* qualunque, mentre quelli del secondo sono sempre di *livello 1*.

Alle due funzioni degli iperlemmi corrispondono due diversi meccanismi di inserimento nel corpus. Il primo è inglobato nell'ambiente **Gestione base dati**, il secondo in **Lemmatizzazione** in quanto più strettamente connesso con questa operazione. Tra gli iperlemmi di livello 1 inseriti sulla base dei due diversi meccanismi non viene fatta alcuna successiva distinzione.

Gli iperlemmi sono stati introdotti pensando non tanto ad una applicazione specifica quanto ad una serie di ambiti nei quali la strutturazione gerarchica dell'informazione linguistica offerta dagli iperlemmi può fornire un utile supporto, anche in considerazione della loro duplice possibilità di aggancio per un verso ai lemmi (e di conseguenza alle forme collegate) e per l'altro ad occorrenze specifiche.



Per esemplificare le possibilità di strutturazione gerarchica dell'informazione linguistica offerte dagli iperlemmi si veda lo schema soprastante, nel quale si rappresenta il saggio di classificazione onomasiologica del lessico realizzato per il corpus **Demo**⁵ sulla base, e con minimi adattamenti, del *Begriffssystem* di Hallig-von Wartburg (Rudolf Hallig, Walther von Wartburg, *Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas – Système raisonné des concepts pour servir de base à la lexicographie. Essai d'un schéma de classement*, seconda ed. riveduta ed ampliata, Berlino, Akademie-Verlag, 1963 [prima ed. tedesca in «Abhandlungen der Deutschen Akademie der Wissenschaften zu Berlin. Klasse für Sprachen, Literatur und Kunst», 1952]).

In questo schema:

- **uomo** è l'unico iperlemma di livello 3,

⁵ In versione originale, ovvero prima delle alterazioni eventualmente introdotte svolgendo gli esempi riportati nei capitoli precedenti e relativi alla lemmatizzazione.

- **uomo fisico** e **uomo sociale** sono iperlemmi di livello 2, connessi superiormente a **uomo** e inferiormente a **attività sensoriali, corpo, linguaggio, scrittura** (tutti iperlemmi di livello 1),
- la base della struttura connettiva piramidale è costituita dai lemmi **orecchia (s.f.)**,..., **numero (s.m.)**.
Si può vedere il caso di più iperlemmi che puntano agli stessi lemmi, ad esempio **linguaggio** e **scrittura** verso **parola (s.f.)**.

Quasi tutti gli iperlemmi riportati nello schema operano secondo il primo dei due modi citati in precedenza, ovvero come *puntatori a iperlemmi o lemmi*. Vi sono però due casi di iperlemmi di livello 1 che sono anche associati direttamente a specifiche occorrenze di forme. Si tratta dell'iperlemma **attività sensoriali** connesso con occorrenze di **ochi** e **ochu**, occorrenze lemmatizzate con **occhio (s.m.)** e dell'iperlemma **linguaggio** connesso con occorrenze di **lingua** lemmatizzate con **lingua (s.f.)**. Questo meccanismo consente di evidenziare le diverse accezioni della forma. Queste ultime connessioni sono riportate con un tratto più marcato per distinguerle dalle precedenti.

Le connessioni illustrate nello schema verranno sfruttate nelle esercitazioni di *ricerche tramite iperlemmi* (Lezione 38); le connessioni fra differenti tipi di elementi (iperlemmi di livello superiore o inferiore, lemmi, occorrenze) verranno ripercorse sfruttando le diverse modalità di ricerca disponibili. A questo riguardo va tenuto presente che le ricerche per iperlemmi, una volta arrivate a livello dei lemmi, possono condurre direttamente a tutte le occorrenze lemmatizzate con questi ultimi a prescindere dalle forme di appartenenza, oppure possono essere indirizzate in modo più mirato a specifiche forme (e relative occorrenze) associate a quei lemmi od anche, volendo, ad occorrenze di tali forme magari non lemmatizzate o lemmatizzate con lemmi diversi.

Le prossime due Lezioni illustreranno i meccanismi utilizzabili per introdurre nuovi iperlemmi in un corpus, utilizzando sia il corpus **Prova** costruito nel corso delle Lezioni relative all'inserimento di un testo che il corpus **Demo**. La struttura di iperlemmi di cui **Demo** dispone già agevola i nuovi inserimenti, mentre **Prova** è utile per capire come procedere partendo dalla completa assenza di iperlemmi nel corpus; in questo caso andranno preventivamente introdotte alcune lemmatizzazioni per disporre della base della piramide su cui andranno a collocarsi successivamente gli iperlemmi.

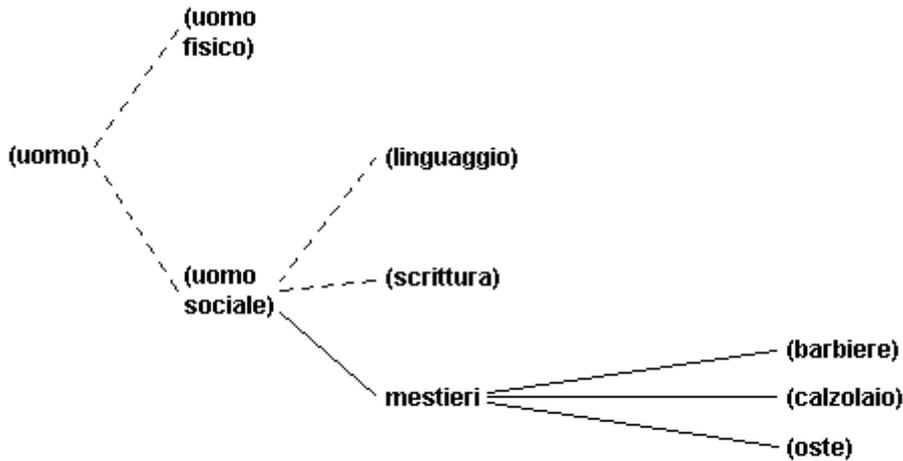
La Lezione 23 tratterà l'inserimento di iperlemmi collegati a lemmi o ad altri iperlemmi, la Lezione 24 mostrerà come procedere alle associazioni dirette iperlemma - occorrenza (questo secondo argomento è stato già illustrato nelle Lezioni 15 e 19 relative alla lemmatizzazione e viene qui riproposto per motivi di completezza e organicità di informazione).

LEZIONE 23. COME INSERIRE IPERLEMMI E COLLEGARLI AI LEMMI.

L'inserimento di iperlemmi verrà illustrato con due esempi. Nel primo la struttura di iperlemmi già presente nel corpus **Demo** e illustrata nello schema precedente verrà ampliata, introducendovi un nuovo elemento; nel secondo verranno inseriti alcuni lemmi e iperlemmi nel corpus **Prova** costruito in precedenza con i testi **es1**, **es2** e **es3** e contenente per ora solo forme.

Esempio 1: inserimento di un nuovo iperlemma nel corpus Demo

Scopo dell'esercitazione è di introdurre un nuovo iperlemma nel corpus **Demo**. Alla fine la struttura relativa risulterà integrata secondo lo schema seguente, dove le parentesi e i tratteggi indicano elementi e collegamenti preesistenti (vedere lo schema precedente).



Barbiere, calzolaio e oste sono lemmi, gli altri iperlemmi.

Avviare GATTO e portarsi nell'ambiente **Gestione base dati** mediante [GEST. BASE DATI].

Tramite [CORPUS] andare a selezionare il corpus **Demo**, aprirlo con [APRI], quindi tornare alla finestra GESTIONE BASE DATI con [CHIUDI FINESTRA].

Aprire la finestra di gestione degli iperlemmi con [IPERLEMMI]. Questa finestra presenta una struttura piuttosto articolata, peraltro non evidenziata subito all'apertura della finestra: per visualizzarla, cliccare su [MODIFICA].

Complessivamente la schermata di GATTO risulta divisa in 5 zone, 2 nella parte alta, 3 in basso.

La parte superiore è occupata per la maggior parte da una griglia che può mostrare, alternativamente, il contenuto dell'iperlemmario o del lemmario del corpus; per ogni iperlemma o lemma sono visualizzati tutti gli elementi, compresi quindi il disambiguatore e il livello nel caso degli iperlemmi.

A destra della griglia vi è un riquadro, in questo caso riportante l'intestazione **iperlemma da modificare**, che include 3 caselle e che può contenere a seconda dell'operazione impostata nessuna, una o due frecce blu su sfondo bianco (avendo cliccato su [MODIFICA] è visibile una sola freccia a sinistra della casella **iperlemma**). Le tre caselle presenti nel riquadro mostrano o permettono di assegnare o modificare i dati dell'iperlemma di volta in volta preso in considerazione (*iperlemma corrente*).

Nella parte inferiore della finestra, a sinistra, vi è una griglia destinata a contenere gli iperlemmi di livello immediatamente superiore a quello dell'*iperlemma corrente* e *collegati* con quest'ultimo.

Sulla destra vi è una griglia analoga, destinata a contenere gli iperlemmi di livello immediatamente inferiore a quello del *lemma corrente* e *collegati* con quest'ultimo. Se l'*iperlemma corrente* è di livello 1, le sue connessioni inferiori saranno però dirette a lemmi anziché iperlemmi: la griglia a destra mostrerà quindi iperlemmi o lemmi a seconda della situazione (in partenza, poiché il livello impostato inizialmente per l'*iperlemma corrente* è 1, vedi il riquadro in alto a destra, la griglia è impostata sui lemmi).

Sempre in basso, al centro, vi è un ultimo riquadro composito che include vari elementi: una casella di inserimento, un riquadro che indica l'operazione in corso di esecuzione e i relativi bottoni di comando. All'interno della casella a sfondo bianco si possono digitare le lettere iniziali degli elementi presentati nella griglia superiore, così da limitarne il numero; ad esempio, digitandovi la lettera **s**, la griglia mostrerà solo gli iperlemmi che iniziano per **s**, in questo caso **scienza e tecnica** e **scrittura**; per ripristinare il contenuto della griglia cancellare la lettera **s**.

Completata la descrizione degli elementi presenti nella finestra, si può procedere all'inserimento del nuovo iperlemma. Chiudere la finestra di modifica con (**Annulla**) e selezionare [AGGIUNGI].

L'iperlemma che si vuole inserire (vedere schema precedente) è **mestieri**, di livello 1.

Il tipo di operazione, *aggiunta di un nuovo iperlemma*, è quello già impostato nel riquadro centrale a sfondo giallo. Iperlemma, eventuale disambiguatore e livello andranno introdotti usando il riquadro **nuovo iperlemma** in alto a destra. Il livello voluto è 1 e risulta già impostato; se l'inserimento avesse riguardato un iperlemma di livello diverso questo andava impostato nel riquadro cliccando sulla freccia a destra della scritta **livello 1** e scegliendo tra quelli disponibili. Nel corpus **Demo** il massimo livello è attualmente 4, come si può vedere cliccando sulla freccia: ciò è dovuto al fatto che l'iperlemma di massimo livello introdotto è **uomo (liv. 3)**, per cui al massimo si può inserire un iperlemma di livello 4, da connettere inferiormente

proprio a **uomo**; se questa operazione venisse effettivamente svolta subito dopo verrebbe dichiarato disponibile anche il livello 5, e così via.

Tornando all'inserimento in corso, nella casella sottostante a **livello 1** e identificata dall'etichetta **iperlemma** digitare **mestieri**; lasciare in bianco in quanto non necessaria la casella **disambiguatore**. Come detto altre volte un iperlemma può essere inserito solo se connesso inferiormente con iperlemmi di livello inferiore o, nel caso di iperlemmi di livello 1, con lemmi. Se a questo punto si prova a concludere l'operazione di inserimento dell'iperlemma cliccando sul bottone (**Registra**) si otterrà il messaggio di errore **l'iperlemma è privo di associazioni con elementi del livello inferiore**. In effetti l'operazione è incompleta in quanto manca la programmata associazione di **mestieri** ai lemmi **barbiere**, **calzolaio** e **oste**, o a un qualunque altro lemma. Premere (**OK**) per chiudere il messaggio.

I lemmi da associare a **mestieri** vanno prelevati dal lemmario, da visualizzare nella griglia superiore; a questo scopo utilizzare [**MOSTRA LEMMI**]. Per visualizzare i lemmi voluti, scorrere il lemmario utilizzando la barra di scorrimento o utilizzare il riquadro bianco al centro: digitando in quest'ultimo la stringa **bar** i lemmi visualizzati si ridurranno a 8, il quarto dei quali è **barbiere (s.m.)**. Selezionare **barbiere** cliccando sul riquadro grigio all'inizio della riga, che apparirà evidenziata; connettere questo lemma al nuovo iperlemma **mestieri** mediante (**Collega**): l'avvenuta connessione tra lemma e iperlemma è confermata dall'inserimento del lemma selezionato nella griglia in basso a destra. Ripetere l'operazione per i lemmi **calzolaio (s.m.)** e **oste (s.m.)**.

Lo schema di inserimento di **mestieri** prevedeva anche di connetterlo superiormente all'iperlemma di livello 2 **uomo sociale**. La selezione dell'elemento da connettere con **mestieri** riguarda ora non più i lemmi ma gli iperlemmi, quindi utilizzare [**MOSTRA IPERLEMMI**] per reimpostare la griglia superiore sull'iperlemmario; in fondo alla lista appare l'iperlemma di livello 2 **uomo sociale**. Selezionarlo cliccando sul corrispondente riquadro grigio a sinistra quindi, sempre tramite (**Collega**), copiarlo nella griglia in basso a sinistra iperl. collegati di liv. 2.

A questo punto è bene aprire un inciso per chiarire che la scelta della griglia (sinistra o destra) nella quale inserire gli elementi selezionati (da collegare al nuovo iperlemma) è fatta in maniera automatica dal programma sulla base dei corrispondenti livelli. Supponiamo ad esempio che si stia aggiungendo un nuovo iperlemma di livello 2; si immagini di selezionare contemporaneamente, nell'iperlemmario, un iperlemma di livello 1, uno di livello 3 e uno di livello 4; l'uso di (**Collega**) produrrebbe il seguente effetto: l'iperlemma di livello 3 comparirebbe nella griglia di sinistra, quello di livello 1 in quella di destra, quello di livello 4 verrebbe semplicemente ignorato, in quanto il suo livello 4 impedisce di connetterlo direttamente col nuovo iperlemma di livello 2.

Tornando all'inserimento di **mestieri**, a questo punto i dati contenuti nella parte bassa della finestra sono:

- **nuovo iperlemma:** livello 1, iperlemma **mestieri**, disambiguatore vuoto
- griglia **iperl. collegati di liv. 2: uomo sociale**
- griglia **lemmi collegati: barbiere, calzolaio, oste.**

corrispondenti alla struttura che si voleva introdurre in **Demo**.

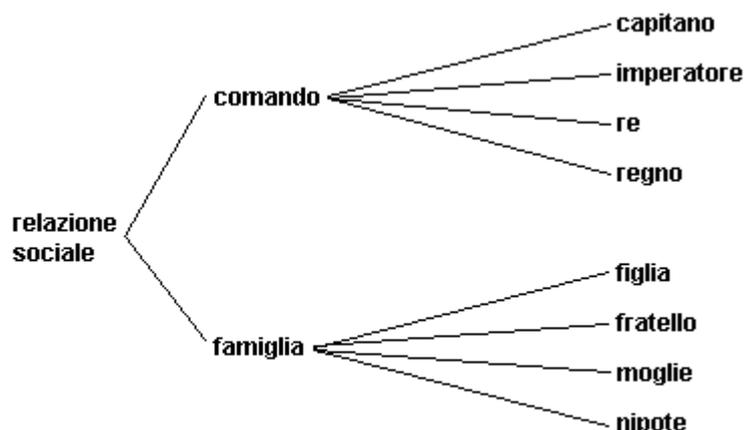
Non rimane che completare l'operazione procedendo alla registrazione del nuovo iperlemma, mediante il bottone (**Registra**): il nuovo iperlemma **mestieri**, di livello 1, comparirà nella griglia superiore in quanto inserito nell'iperlemmario.

Usare [**CHIUDI FINESTRA**] per tornare alla finestra **GESTIONE BASE DATI**.

Esempio 2: inserimento di alcuni iperlemmi nel corpus Prova

Il corpus **Prova** non contiene iperlemmi e neanche lemmi. Sarà quindi necessario partire dall'inserimento di questi ultimi, per costruire una base alla quale sovrapporre la struttura degli iperlemmi. In questa Lezione si farà uso della **Lemmatizzazione per singolo testo**, ma sarebbe stato possibile utilizzare la **Lemmatizzazione sul corpus**, senza differenze nel risultato.

Più precisamente obiettivo dell'esercitazione sarà quello di inserire in **Prova** la struttura di lemmi e iperlemmi indicata nello schema che segue.



Come primo passo si inseriranno i lemmi, seguendo le procedure illustrate nel capitolo 9.

Se si è nella finestra GESTIONE BASE DATI portarsi alla finestra iniziale di GATTO mediante [ALTRO AMBIENTE].

Dalla finestra iniziale di GATTO portarsi nell'ambiente **Lemmatizzazione per singolo testo** mediante [LEMMATIZZAZIONE + PER SINGOLO TESTO]. Aprire se necessario il corpus **Prova**, quindi con [TESTO] selezionare il testo **es3** e dare (OK). Cliccare su [LEMMATIZZAZIONE.. + SELEZIONE FORMA] e digitare **capitanij** seguito da (OK).

Per lemmatizzare l'unica occorrenza di **capitanij** usare (N. lem/omo/ipe), digitare **capitano** seguito da (OK), poi introdurre la categoria grammaticale **s.m.** e (OK).

Utilizzando (Forma) per passare ad altra forma e (Menu) per cambiare testo effettuare in maniera analoga le altre lemmatizzazioni suggerite in tabella (per le forme per le quali compaiono più occorrenze è sufficiente lemmatizzarne una a piacere).

testo	forma	lemma	cat. gr.
es3	fratri	fratello	s.m.
es3	imperatorum	imperatore	s.m.
es3	nepotes	nipote	s.m.
es3	reges	re	s.m.
es3	regnum	regno	s.m.
es2	filia	figlia	s.f.
es2	moier	moglie	s.f.

A questo punto uscire dall'ambiente **Lemmatizzazione per singolo testo** con [ALTRO AMBIENTE], tornare in **Gestione base dati** e cliccare su [IPERLEMMI]; selezionare [AGGIUNGI] (le altre operazioni sono disabilitate essendo l'iperlemmario vuoto).

Nella casella **iperlemma** in alto a destra digitare **comando**; attivare la visualizzazione del lemmario nella griglia superiore con [MOSTRA LEMMI]; tenendo premuto il tasto <Ctrl> per la selezione multipla selezionare nella griglia superiore, cliccando i riquadri grigi a inizio riga, i lemmi **capitano**, **imperatore**, **re** e **regno**; rilasciare il tasto <Ctrl>; collegare i lemmi al nuovo iperlemma con (Collega).

Come si può notare, la griglia **lemmi collegati** include la colonna **occ.**, la quale contiene sempre **0**: questo dato indica il numero di occorrenze attualmente collegate direttamente all'iperlemma.

Completare l'operazione con (Registra).

Ripetere l'operazione per il nuovo iperlemma **famiglia**, sempre di livello **1**, collegato con i lemmi **figlia**, **fratello**, **moglie** e **nipote**.

L'ultimo passaggio consiste nell'inserire l'iperlemma di livello **2 relazione sociale**. A questo scopo aprire di nuovo la finestra di inserimento iperlemmi con (Aggiungi), quindi utilizzando la freccia in basso posta nella casella sottostante la dicitura **nuovo iperlemma** (in alto a destra), selezionare il livello **2**; nel sottostante riquadro **iperlemma** digitare **relazione sociale**. Poiché gli elementi di livello inferiore cui questo iperlemma va connesso non sono più lemmi ma iperlemmi con [MOSTRA IPERLEMMI] impostare la griglia superiore alla visualizzazione degli iperlemmi: compariranno così **comando** e **famiglia**, entrambi di livello

1; selezionarli entrambi e collegarli al nuovo iperlemma con (**Collega**); inserire l'iperlemma nell'iperlemmario con (**Registra**).

Per verificare il lavoro fatto cliccare su [NAVIGA], selezionare l'iperlemma **famiglia** e cliccare sulla freccia blu che porta dalla griglia superiore alla casella laterale **iperlemma**, nella quale comparirà l'iperlemma selezionato; contemporaneamente nelle griglie sottostanti verranno visualizzati gli elementi connessi di livello superiore e inferiore: **relazione sociale** a sinistra e **figlia, fratello, moglie e nipote** a destra.

Terminare l'operazione con (**Annulla**) e chiudere la finestra con [CHIUDI FINESTRA].

LEZIONE 24. COME ASSOCIARE DIRETTAMENTE GLI IPERLEMMI ALLE SINGOLE OCCORRENZE.

Come accennato questa Lezione, inclusa nel presente capitolo per motivi di completezza e organicità di informazione, ribadisce nozioni già contenute nelle Lezioni 15 e 19 a proposito della lemmatizzazione. Chi avesse già seguito gli esempi lì riportati può saltare tranquillamente questa Lezione senza perdere informazioni.

Operando col corpus **Demo** portarsi nell'ambiente **Lemmatizzazione per singolo testo**, selezionare il testo **in** e impostare **libri** come forma da lemmatizzare (con meccanismi del tutto simili si possono associare direttamente gli iperlemmi alle singole occorrenze utilizzando l'ambiente **Lemmatizzazione sul corpus**).

Delle 6 occorrenze presenti le prime due risultano lemmatizzate col lemma **libro (s.m.)**. L'assenza di una **I** nera nella riga di testata delle due occorrenze lemmatizzate evidenzia che a nessuna di esse è associato un iperlemma.

Supponiamo ora di voler lemmatizzare altre due occorrenze con la combinazione lemma **libro (s.m.)** e iperlemma **scrittura**.

Aperto il dizionario di macchina compare ovviamente il lemma **libro (s.m.)**: se si lemmatizzasse una nuova occorrenza con tale lemma, cliccando ad esempio sul dizionario di macchina o prelevandolo dal *lemma impostato*, non sarebbe possibile associare all'occorrenza anche l'iperlemma qualora questo non risultasse già collegato col lemma usato; sarebbe quindi utile sapere in partenza se al lemma **libro (s.m.)** è o no già associato l'iperlemma **scrittura**. Come già visto nei capitoli che trattano della lemmatizzazione, cliccando in un punto qualunque del dizionario di macchina col tasto destra del mouse lo si predispone a fornire notizie aggiuntive sui lemmi inclusi, precisamente forme o iperlemmi collegati; a quel punto per conoscere gli eventuali iperlemmi collegati con **libro (s.m.)** cliccare col tasto sinistra in un punto della riga corrispondente, esternamente alla colonna intestata **lemma**. Come si può vedere non esistono iperlemmi associati a **libro (s.m.)**. Utilizzare <Esc> per ripristinare il dizionario di macchina.

Ottenuta questa informazione preliminare si può passare alla lemmatizzazione, che non potrà quindi essere fatta a partire dal contenuto del dizionario di macchina.

Posizionarsi sul contesto numero 3, cliccare su (**N. lem/omo/ipe**) e digitare **libro** seguito da (**OK**); ovviamente GATTO propone il lemma omografo **libro (s.m.)** presente nel lemmario; rifiutare l'indicazione cliccando ancora su (**N. lem/omo/ipe**) e digitare **s.m.** come categoria grammaticale e **scrittura** come iperlemma, infine dare (**OK**). La lemmatizzazione è conclusa correttamente, come mostrano i dati in testata al terzo contesto; cliccando su tale testata si apre una finestra che mostra l'effettiva associazione di lemma e iperlemma a quel contesto.

Anche il *lemma impostato* è stato modificato e contiene ora la coppia **libro (s.m.)** più **scrittura**.

A questo punto lemmatizzando col lemma impostato si ottiene la contemporanea associazione sia del lemma che dell'iperlemma ai successivi contesti.

Lemmatizzando invece dal dizionario di macchina si ha libertà di scelta se utilizzare solo il lemma **libro (s.m.)** o anche l'iperlemma **scrittura**. Vediamo in pratica queste possibilità sulle due occorrenze successive: verranno lemmatizzate entrambe con **libro (s.m.)**, ma solo alla prima verrà associato anche l'iperlemma **scrittura**.

Se non vi si è già, portarsi sul contesto numero 4. Cliccare sul dizionario di macchina per lemmatizzare con **libro (s.m.)**; GATTO propone di collegare all'occorrenza l'unico iperlemma associato al lemma scelto, ovvero **scrittura**: per accettare l'indicazione cliccare sulla riga di **scrittura**. La lemmatizzazione è conclusa, con associazione della coppia lemma più iperlemma all'occorrenza.

Si è a questo punto sul contesto numero 5. Come prima cliccare sulla riga di **libro (s.m.)** nel dizionario di macchina; come prima viene proposto l'iperlemma **scrittura**; cliccare col tasto destra in un punto qualunque della finestra che mostra l'iperlemma, per rifiutarlo: viene ripristinato il dizionario di macchina,

mentre la lemmatizzazione è completata facendo uso del solo lemma. Anche il lemma impostato viene modificato di conseguenza e non include più l'iperlemma.

Come si è visto, un clic sinistro sulla riga dell'iperlemma lo accetta, un clic destro in un punto qualunque della finestra degli iperlemmi lo rifiuta; esiste una terza possibilità, l'uso del tasto <Esc>, per far abortire la lemmatizzazione in corso.

La tecnica di associazione di un iperlemma di livello 1 ad un lemma e il suo contemporaneo collegamento ad un'occorrenza di una forma non sono influenzate dalla preesistenza o meno, negli archivi del corpus, dell'iperlemma e della sua associazione al lemma: se questi elementi sono già presenti nella base di dati ne viene fatto uso automaticamente e senza generare duplicati, altrimenti vengono costruiti. Se esistesse un iperlemma omografo di quello indicato, ma di livello maggiore di 1, ne verrebbe aggiunto uno nuovo di livello 1.

CAP. 13- LEMMI E COPPIE FORMA-LEMMA MUTI. ESPORTAZIONE E IMPORTAZIONE DI LEMMARIO E IPERLEMMARIO

Nel capitolo 9 sono stati descritti i meccanismi e gli effetti delle operazioni di lemmatizzazione in GATTO.

In particolare va ricordato che, una volta portata a termine, una lemmatizzazione ha una duplice conseguenza: realizza una connessione tra il lemma e la specifica occorrenza cui è stato associato ma in più crea una connessione a livello di corpus tra il lemma e la forma di cui l'occorrenza era espressione. Grazie a questa connessione una ricerca fatta a partire dal lemma permetterà di risalire a tutte le forme associate ad esso in uno o più testi del corpus, fornendone sia le occorrenze lemmatizzate col lemma in questione che, volendo, quelle non lemmatizzate o lemmatizzate con altri lemmi.

Un altro aspetto della gestione dei lemmi in GATTO è che ogni coppia forma-lemma rimane negli archivi finché nel corpus è presente almeno un'occorrenza della forma lemmatizzata col lemma; se questa condizione viene meno (per cancellazione di testi dal corpus o per modifica delle lemmatizzazioni in oggetto), la coppia forma-lemma viene cancellata automaticamente dagli archivi. Analogamente, un lemma viene mantenuto nel lemmario del corpus fino a che, sempre per le possibili cause sopracitate, non accada che l'ultima occorrenza lemmatizzata con tale lemma (o, il che è equivalente, finché l'ultima coppia forma-lemma associata a quel lemma) cessi di essere presente nel corpus.

Questo meccanismo di gestione semiautomatica dei lemmi e dei loro collegamenti con le forme è risultata essere la più funzionale rispetto agli scopi di GATTO, per lo meno nella maggior parte dei casi.

Esiste tuttavia almeno una situazione nella quale questo meccanismo finisce per creare difficoltà a chi lemmatizza, complicandone il lavoro. Si tratta del caso, tutt'altro che raro, della presenza di due o più corpus distinti nella stessa lingua, per i quali può essere utile utilizzare il dizionario di macchina di uno per lemmatizzare gli altri. In GATTO, poichè il dizionario di macchina di un corpus si sviluppa dinamicamente sulla sola base delle lemmatizzazioni già effettuate in quello stesso corpus, in assenza di altri strumenti l'unico modo per applicare un dizionario di macchina ad un altro corpus è:

- 1 lemmatizzare il primo corpus
- 2 creare un duplicato del primo corpus, su cui operare i passi che seguono
- 3 aggiungere i testi del secondo corpus
- 4 procedere con la lemmatizzazione del secondo pacchetto di testi
- 5 cancellare i testi del primo pacchetto, così che il corpus duplicato finisca per contenere solo i testi, lemmatizzati, del secondo corpus

Questo sistema presenta lo svantaggio di presentare, in fase di lemmatizzazione del secondo pacchetto di testi, le occorrenze dei due corpus mescolate tra loro, rendendo decisamente difficoltoso il lavoro di chi lemmatizza.

Per risolvere il problema sono stati introdotti in GATTO due nuovi elementi: *lemmi muti* e *coppie forma-lemma mute*. Si tratta di lemmi e collegamenti lemma-forma presenti e disponibili negli archivi anche se non collegati ad alcuna occorrenza lemmatizzata nel corpus corrente; in questo modo elementi non derivati dalla lemmatizzazione già effettuata sul corpus (ma da quella di un altro corpus) possono essere assunti come dizionario di macchina, quindi come base per le operazioni di lemmatizzazione successive.

Per produrre lemmi e coppie forma-lemma muti in un corpus a partire da un altro corpus già lemmatizzato, è sufficiente esportare il lemmario (ed eventualmente l'iperlemmario) di quest'ultimo in un

database apposito mediante l'operazione di *esportazione lemmario* e quindi reimportarlo nel nuovo corpus con l'operazione di *importazione lemmario*.

Quando, in fase di lemmatizzazione del nuovo corpus, viene utilizzato un elemento *muti*, ad esso vengono associate una o più occorrenze lemmatizzate; da quel momento il lemma o la coppia cessano di essere *muti* e divengono elementi normali, soggetti tra l'altro a cancellazione automatica qualora le nuove occorrenze associate ad essi venissero rimosse dal corpus.

Da quanto detto segue che la sequenza delle operazioni per applicare il dizionario di macchina di un corpus ("origine") ad un altro ("destinazione") è la seguente:

- 1 lemmatizzare il corpus "origine"
- 2 esportarne il lemmario
- 3 importare il lemmario estratto da "origine" nel corpus "destinazione" (in questa fase si sceglie se importare o meno anche l'eventuale iperlemmario)
- 4 procedere alla lemmatizzazione del corpus "destinazione", facendo uso anche dei lemmi muti e delle coppie forma-lemma mute introdotti al suo interno dalla precedente operazione di importazione
- 5 al termine, volendolo, rimuovere gli elementi muti rimasti inutilizzati

Il meccanismo dell'importazione del lemmario si applica a tutti i lemmi, inclusi nel database esportato dal corpus di partenza, per i quali esiste almeno un legame forma-lemma con una forma presente nel corpus di arrivo; in altre parole, non vengono trasferiti i lemmi *inutili*, nel senso che non risultano associati alle forme presenti nel corpus di arrivo.

È stato detto che gli elementi muti possono essere usati anche in interrogazione: a prima vista questa affermazione può lasciare perplessi dal momento che si tratta di oggetti i quali per definizione non *puntano* a niente. Il meccanismo di ricerca chiamato in causa in questo caso si basa sull'uso di opzioni presenti nell'ACCUMULATORE, che consentono di utilizzare una ricerca fatta a partire da una coppia forma-lemma, non limitandola alle occorrenze della forma lemmatizzate col lemma, ma estendendola a tutte le altre occorrenze della forma. Quindi, ad esempio, partendo dalla coppia (supposta muta) lemma **casa (s.f.)** + forma **case**, si possono localizzare tutte le occorrenze non lemmatizzate della forma **case** anche se questa forma non era stata mai lemmatizzata nel corpus corrente.

Una volta completata la lemmatizzazione del corpus tutti gli elementi muti non utilizzati risultano ancora *muti*, cioè non associati ad alcuna occorrenza; se si ritiene che la loro ulteriore presenza sia inutile o ambigua, è possibile eliminarli con apposite operazioni.

LEZIONE 25. COME TRASFERIRE IL LEMMARIO (ED EVENTUALMENTE L'IPERLEMMARIO) DI UN CORPUS IN UN ALTRO CORPUS

Questa Lezione richiede l'uso di due corpus; si useranno **Demo** come corpus dal quale estrarre il lemmario e **Prova** come corpus nel quale importarlo. Quando sarà richiesto di costruire **Prova**, vedere la Lezione 11 per la creazione del corpus e le Lezioni 13 e 14 per l'inserimento dei testi **es1**, **es2** e **es3**; non è necessario inserire anche i testi associati.

Per prima cosa aprire **Demo** e portarsi nell'ambiente **Gestione base dati** mediante [GEST. BASE DATI]; effettuare l'esportazione del lemmario tramite [LEMMARIO.. + ESPORTAZIONE LEMMARIO]; al termine dell'operazione un messaggio avverte che il lemmario esportato si trova nel file **c:\demo.gat\out\democopialemmi.mdb**.

Procedere ora alla costruzione del corpus Prova e all'inserimento al suo interno dei testi es1, es2 e es3 senza testi associati (non servono).

Una volta inseriti i 3 testi nel corpus, importarvi il lemmario estratto da Demo tramite [LEMMARIO.. + IMPORTAZIONE LEMMARIO + IPERLEMMI NON INCLUSI] (se si ha pratica di iperlemmi si può anche scegliere l'alternativa che include gli iperlemmi); specificare disco, cartella e nome del file **c:\demo.gat\out\democopialemmi.mdb**. Al termine si può vedere subito che i lemmi sono stati inseriti per il fatto che tutti i sottomenu del menu [LEMMARIO] ora sono attivati.

Passare alla finestra iniziale dell'ambiente **Ricerche** con [ALTRO AMBIENTE] e [RICERCHE]. Avviare la ricerca per lemmi con [RICERCHE.. + PER LEMMI]; impostare la ricerca di **a*** e procedere con [AVVIO RICERCA]. Si ottengono **32 lemmi muti**, ovvero caratterizzati dal fatto che hanno 0 occorrenze, come ovvio (e sono evidenziati in rosso; cliccando su **mostra forme** si visualizzeranno 55 coppie forma-lemma mute, caratterizzate dagli stessi elementi).

Il corpus **Demo** è stato così dotato di un lemmario virtuale, ovvero di un insieme di lemmi direttamente utilizzabili per lemmatizzarlo, in quanto inseriti automaticamente nei dizionari di macchina delle forme incluse nelle coppie *forma-lemma mute* importate. Tali elementi sono utilizzabili anche nell'ambiente **Ricerche**, purché si abbia l'accortezza di attivare il *checkbox non lemmatizzate* presente nell'ACCUMULATORE prima di passare alla visualizzazione dei contesti.

PARTE IV - ESTRAZIONE DI INFORMAZIONI LESSICOGRAFICHE

CAP. 14 - ESTRAZIONE DI INFORMAZIONI DA UN CORPUS GATTO

CORPUS E SOTTOCORPUS

GATTO consente di estrarre da un corpus testuale informazioni di varia natura. Quasi tutte le finestre di interfaccia predisposte per queste operazioni sono incluse nell'ambiente **Ricerche**, al quale si accede al pari degli altri dalla finestra di avvio di GATTO.

Prima di passare in rassegna le funzioni disponibili per l'interrogazione della base di dati è opportuno chiarire che GATTO opera su testi raggruppati in corpus ma questo non significa che un corpus vada considerato come un tutto indivisibile e che le ricerche vengano sempre e comunque estese alla sua totalità. Al contrario, è possibile definire in maniera dinamica dei sottocorpus sulla base dei dati bibliografici associati ai testi, consentendo peraltro ad uno stesso testo di appartenere contemporaneamente a più sottocorpus. Una volta definiti dei sottocorpus è possibile effettuare ricerche alternativamente in tutto il corpus, in un sottocorpus o in una qualunque combinazione dei sottocorpus definiti; alcune ricerche operano su testi singoli.

L'insieme di testi ai quali si applica di volta in volta una certa ricerca è indicato col termine *dominio di ricerca*.

Per completare il quadro delle informazioni relative a corpus e sottocorpus va ricordato che sia gli uni che gli altri possono essere costituiti da un minimo di 1 a un massimo di 16.000 testi.

PANORAMICA SULLE RICERCHE EFFETTUABILI

Il tipo di ricerca più immediato è quello che parte dalle forme presenti nel corpus e ne ricerca le occorrenze (*ricerca per forme*).

Le ricerche per forme possono chiamare in causa i lemmi utilizzati nelle operazioni di lemmatizzazione, consentendo di discriminare i contesti cercati sulla base dell'esistenza o meno di una lemmatizzazione e, in caso affermativo, sulla base del lemma usato. Si possono anche impostare le ricerche partendo direttamente dai lemmi associati alle forme (*ricerca per lemmi*): in questo modo possono essere localizzate contemporaneamente occorrenze caratterizzate non tanto dal fatto di afferire alle stesse forme quanto da quello di essere state lemmatizzate con gli stessi lemmi.

Le ricerche per lemmi possono essere effettuate anche a partire da specifiche proprietà di questi ultimi (*ricerca per categorie grammaticali, ricerca per disambiguatori*) che consentono di raggrupparli in base ai valori di tali proprietà.

Normalmente all'interno di GATTO i lemmi sono utilizzati come uno strumento per cercare delle forme, per cui i contesti ottenuti al termine della ricerca sono comunque ordinati per testo (ordine cronologico o basato su altri dati bibliografici) o per ordine alfabetico di forma, a prescindere dai lemmi associati (o meno) alle singole occorrenze localizzate. Esiste però un particolare tipo di ricerca per lemmi che consente di mostrare le occorrenze risultanti ordinandole in base ai lemmi con cui sono state lemmatizzate (*concordanza lemmatizzata*).

Anche gli iperlemmi possono costituire una base dalla quale avviare la ricerca, aumentando in maniera considerevole la possibilità di raggruppare forme od occorrenze sulla base dei loro significati (*ricerca per iperlemmi*).

Non sempre il risultato cercato include la esplicita visualizzazione di contesti: in certi casi è sufficiente localizzare le occorrenze nei testi ottenendone un indice, tenendo conto o meno delle lemmatizzazioni effettuate (generazione di *index locorum per forme o lemmatizzati*).

Ricerche di tipo diverso possono essere combinate tra loro, nel senso che possono esserlo i risultati ottenuti dalle singole ricerche, prima di passare alla visualizzazione dei relativi contesti. La chiave sulla quale si basa la combinazione di più ricerche è L'ACCUMULATORE, ovvero la finestra nella quale possono essere fatti confluire tutti gli elementi localizzati (forme, lemmi, coppie forma-lemma). Sono previsti due tipi di ricerca combinata, rispettivamente basati sul criterio di somma degli elementi localizzati nelle varie

ricerche (logica OR) e su quello di accettazione dei soli elementi comuni a più ricerche (logica AND). Un esempio di ricerche combinate è riportato nella Lezione 36.

Le ricerche descritte fin qui mirano a localizzare singole occorrenze di forme, lemmatizzate o meno. Si possono peraltro cercare cooccorrenze di più parole, collocate sequenzialmente nel testo o in prossimità tra loro, con l'eventuale vincolo di appartenere allo stesso periodo (*ricerca di cooccorrenze*). In questo caso la ricerca può coinvolgere anche i segni di interpunzione, come elementi da cercare direttamente o rispetto ai quali definire le posizioni delle parole cercate.

Questo primo pacchetto di ricerche ha come elemento comune l'obiettivo di localizzare comunque occorrenze di parole nei testi, eventualmente con l'aggiunta dei segni di interpunzione. Un secondo pacchetto mira invece ad ottenere informazioni di altra natura, ovvero formari, lemmari, lemmari esclusivi (cioè costituiti da lemmi che compaiono solo all'interno del *dominio di ricerca*) o incipitari.

Il contenuto dei testi non è estraibile integralmente dal corpus, ma dei singoli testi possono essere lette le parti afferenti a specifici riferimenti organici.

Per quanto riguarda le informazioni statistiche si possono conoscere le frequenze assolute o relative e le periodicità delle forme all'interno di un singolo dominio di ricerca o, in forma comparata, mettendo a confronto sottocorpus con altri sottocorpus o col corpus.

Infine si possono ottenere alcune uscite grafiche, ovvero curve forme-frequenze relative a singoli testi o a sottocorpus e istogrammi della distribuzione temporale delle occorrenze di forme o gruppi di forme.

FASI DI UNA RICERCA LESSICOGRAFICA

Normalmente una ricerca di contesti in GATTO si articola in quattro fasi successive:

1. impostazione e avvio della ricerca tramite digitazione degli elementi cercati, siano essi forme, lemmi o altro, con l'eventuale ricorso a caratteri jolly;
2. selezione, tra gli elementi così localizzati nel corpus, di quelli di reale interesse; questa fase permette di escludere il *rumore* prodotto dalla comparsa di elementi non previsti, magari presenti con frequenze elevate; gli elementi selezionati vengono estratti e raccolti in un contenitore detto *Accumulatore*, eventualmente in combinazione con altri ricavati da ricerche precedenti anche di tipo diverso;
3. selezione, tra gli elementi presenti nell'*Accumulatore*, di quelli dei quali si vogliono vedere i contesti e passaggio alla presentazione di questi ultimi;
4. visualizzazione, (eventuale) stampa su carta o trascrizione su file RTF (**R**ich **T**ext **F**ormat) del materiale lessicografico così ottenuto.

Non essendo interessati alle elaborazioni dei dati intermedi è sempre possibile passare direttamente dalla fase 1 alla 4, ottenendo direttamente i contesti degli elementi digitati.

OPZIONI DI RICERCA E CARATTERI JOLLY

Le ricerche lessicografiche previste all'interno di GATTO possono essere modulate sulla base di una serie di opzioni da utilizzarsi singolarmente o in combinazione per rendere più flessibili e mirate le interrogazioni. Queste opzioni differiscono tra loro in funzione dei tipi di ricerca e verranno descritte nel seguito. Tutte le opzioni sono configurate inizialmente sulla base di impostazioni predefinite che vengono ripresentate all'avvio di ogni nuova ricerca; molte di queste opzioni offrono la possibilità di definire differenti valori iniziali: vedere per questo l'Appendice B - *Impostazione delle opzioni generali*.

Le finestre di inserimento degli elementi da cercare (forme, lemmi, ecc.) contengono 10 riquadri a sfondo bianco, detti *selettori* (o *selettori semplici*), in ognuno dei quali è possibile digitare un elemento; la successiva ricerca riguarderà contemporaneamente i contenuti di tutti i selettori.

Le ricerche di occorrenze possono anche fare uso di caratteri jolly che sono sostanzialmente gli stessi per tutti i tipi di ricerche: per questo motivo tali caratteri verranno ora descritti sulla base di una serie di esempi, prima di passare alla considerazioni più dettagliate delle Lezioni successive.

Esempio 1: abate

Questo caso è il più semplice: non ci sono caratteri jolly e la risposta sarà, se esiste nei testi del corpus, l'unica forma **abate**.

Esempio 2: abat?

Il carattere *jolly* ? sta per un carattere qualunque, pertanto questa ricerca fornirà ad esempio le forme **abate** e **abati**.

Esempio 3: casa*

Il carattere *jolly* * sta per una stringa qualunque, anche vuota, pertanto questa ricerca fornirà tutte le forme che cominciano per **casa**, come **casale**, **casato**, **casacca**, ma anche **casa**.

Esempio 4: fo[n-s]t?

Questa ricerca troverà tutte le parole di 5 caratteri che cominciano per **fo**, hanno come terza lettera un carattere compreso tra **n** e **s** (inclusi), seguito da **t** e da un carattere qualunque. Perciò tra i risultati di questa ricerca vi potranno essere anche **fonte**, **forte**, **forti**, **foste**.

Questo esempio mostra come i caratteri *jolly* siano combinabili tra loro, per imporre più condizioni alle stesse forme.

Esempio 5: pa[rtc]to

Questa ricerca troverà tutte le parole di 5 caratteri che cominciano per **pa**, hanno come terza lettera **r**, **t** o **c** e terminano con **to**. Perciò si troveranno, se presenti nel corpus, **parto**, **patto** e **pacto**.

Esempio 6: re<gat,g?l,m*>are

Le parentesi angolari consentono di racchiudere al loro interno più stringhe alternative di lunghezza diversa, anche nulla, separate da virgole. A loro volta tali stringhe possono includere i caratteri *jolly* ? e *.

L'esempio in questione localizzerebbe, tra l'altro, **regatare**, **regalare**, **regolare**, **remare**, **remigare**, **remunerare**.

Esempio 7: me<n,,nom><dic*,s*,><a,e>re

In questo caso sono usate più parentesi angolari contemporaneamente; la prima e la seconda contengono anche stringhe nulle, indicate rispettivamente con due virgole consecutive e con una virgola non seguita da altri caratteri a fine parentesi. Applicata al corpus **Demo** questa ricerca fornirebbe: **medicare**, **medichinare**, **menare**, **menomare**, **mesere**, **messere**, **mestiere**. La ricerca mostrata riguarda tutte le parole che:

- devono cominciare con **me**,
- possono proseguire con **n** o **nom**,
- successivamente possono proseguire con una stringa qualunque che inizi per **dic** o **s**,
- successivamente devono contenere una **a** o una **e**,
- devono terminare con **re**.

Notare che la presenza della *stringa nulla* nella prima e nella seconda coppia di parentesi rende la presenza delle stringhe indicate al loro interno facoltativa (...*possono*... anziché ...*devono*...).

CONTENUTI DEGLI ALTRI CAPITOLI DI QUESTA PARTE DEL MANUALE

Capitolo 15

Ricerche per forme; opzioni nelle ricerche per forme; ricerche limitate al rimario; uso delle varianti; visualizzazione e stampa dei contesti ottenuti; visualizzazione parallela di note e traduzioni; ricerche a partire da brani di testo; ricerche per forme mai lemmatizzate.

Capitolo 16

Ricerche per lemmi; opzioni relative; ricerche per categorie grammaticali; ricerche per disambiguatori; produzione di concordanze lemmatizzate. Ricerche combinate.

Capitolo 17

Generazione di index locorum ordinati sulla base delle forme cercate o dei lemmi associati ad esse.

Capitolo 18

Ricerche per iperlemmi.

Capitolo 19

Ricerche di cooccorrenze.

Capitolo 20

Costruzione di sottocorpus; limitazione delle ricerche a sottocorpus; informazioni statistiche sui sottocorpus costruiti.

Capitolo 21

Analisi di frequenze; frequenze assolute; frequenze relative; periodicità.

Capitolo 22

Visualizzazione di parti di testo; generazione di formari, lemmari, lemmari esclusivi, incipitari di un singolo testo; formari, lemmari, lemmari esclusivi, incipitari di sottocorpus.

Capitolo 23

Visualizzazione dei dati bibliografici e statistici di un singolo testo; visualizzazione dei dati bibliografici complessivi dei testi appartenenti ad un corpus o a dei sottocorpus.

Il modo per conoscere le statistiche e le impostazioni iniziali assegnate ad un corpus è descritto nella Lezione 1.

NOTA SULL'UTILIZZO DEL CORPUS DEMO NEGLI ESEMPI DI RICERCHE

Le Lezioni incluse in questa parte del manuale fanno uso del corpus **Demo** così come è stato inizialmente inserito nel CD di GATTO o scaricato da WEB. Se di questo corpus è stato fatto uso seguendo gli esempi inseriti nella parte precedente, relativa alla lemmatizzazione, è consigliabile reinstallare il corpus in modo da ripristinare l'originale configurazione di lemmi e iperlemmi. Questo per mantenere la corrispondenza tra i risultati delle ricerche riportati nel seguito con quelli ottenibili effettuandole all'interno di **Demo**.

CAP. 15 - RICERCHE PER FORME

Questo capitolo fornirà una carrellata sui vari aspetti di una ricerca per forme effettuabile tramite GATTO e sulle opzioni disponibili nelle diverse fasi della ricerca stessa ai fini di modificarne lo svolgimento; le Lezioni incluse nel capitolo mostreranno come utilizzare concretamente gli strumenti descritti. Molte delle funzioni e opzioni presentate si applicano in realtà non solo alle ricerche per forme ma anche ad altri tipi di ricerche.

Il tipo di ricerca più immediato è quello che parte dalle forme e ne ricerca le occorrenze all'interno del corpus (*ricerca per forme*). GATTO prevede un gruppo di 10 riquadri, detti *selettori* o anche *selettori semplici*, in ognuno dei quali è possibile digitare una forma da cercare o, facendo uso di caratteri jolly, un insieme di forme caratterizzate da ridotta varianza grafica esprimibile con tali caratteri. La ricerca verrà effettuata contemporaneamente su tutte le forme digitate.

Le interrogazioni per forme offrono in partenza una prima possibilità di scelta tra ricerca di tutte le forme presenti nel corpus, soddisfacenti ovviamente i requisiti richiesti, e limitazione dell'indagine alle forme mai lemmatizzate. Questa seconda possibilità ha lo scopo di fornire più rapidamente un quadro sullo stato di lemmatizzazione del corpus, consentendo di localizzare rapidamente le forme non ancora lemmatizzate; si tratta di una funzione particolarmente utile se si considera che un corpus GATTO è dinamico, ovvero che l'inserimento di nuovi testi potrebbe ad esempio trasformare un corpus completamente lemmatizzato in uno con un certo numero di forme nuove non associate ad alcun lemma e quindi non localizzabili in una ricerca per lemmi. Evidentemente qualora non si sia interessati alla lemmatizzazione questa modalità di ricerca può essere ignorata.

Un'altra opzione riguarda l'ambito entro il quale cercare le occorrenze: mentre il default è la ricerca all'interno di tutto il corpo di un testo, in alternativa si può limitarla al rimaio, ovvero alle occorrenze caratterizzate dal trovarsi a fine verso oppure a quelle marcate come *rima al mezzo* o alla combinazione delle due categorie.

Normalmente in risposta ad un'interrogazione per forme viene mostrato l'elenco di quelle che soddisfano i requisiti richiesti, precisando per ciascuna il numero complessivo di occorrenze presenti nel corpus. È possibile affinare la richiesta attivando l'opzione **mostra lemmi**, la quale specifica che per ogni forma localizzata le occorrenze mostrate vanno ripartite in più gruppi, distinguendo tra occorrenze non lemmatizzate e occorrenze lemmatizzate con lemmi diversi, indicando per ogni gruppo il lemma associato (o l'assenza di lemma) e il numero di occorrenze nel dominio di ricerca; in questo modo la ricerca potrà proseguire coinvolgendo tutti i gruppi di occorrenze o al contrario solo quelli desiderati. A questa opzione si può fare ricorso anche in un secondo tempo, dopo aver preso visione delle forme localizzate senza distinzioni basate sulla lemmatizzazione.

Se nel corpus è stato precedentemente importato il lemmario di un altro corpus, l'operazione ha probabilmente introdotto negli archivi del corpus ricevente delle coppie forma-lemma che in precedenza non vi comparivano; data una di queste coppie, nel corpus non esistono occorrenze della forma che vi compare, lemmatizzate col lemma incluso in quella coppia. Per questo motivo tali coppie sono dette *mute*. Se nel corpus sono presenti coppie forma-lemma mute, diviene visibile l'opzione **lemmi muti** che, se attivata, estende la ricerca anche a questi elementi; tale opzione, in quanto riferita alla ricerca di coppie forma-lemma e non a quella di semplici forme (non esistono *forme mute*), si abilita automaticamente solo quando è attiva l'opzione **mostra lemmi**. Una ricerca basata sulle coppie forma-lemma mute ha senso se, nell'ACCUMULATORE, si fa poi uso delle opzioni **non lemmatizzate** e/o **escl. lemmi selez.**

Le opzioni descritte finora si applicano contemporaneamente a tutte le forme cercate. Esiste poi una serie di opzioni che possono essere applicate o meno ai singoli selettori, permettendo di diversificare le impostazioni in funzione delle forme cercate. Tali opzioni verranno ora discusse brevemente.

OPZIONI PER SINGOLO SELETTORE

Ricerca espansa

Parole differenti solo per gli attributi (accenti, dieresi, cediglie) vengono normalmente trattate come distinte: è possibile quindi limitare la ricerca delle parole sulla base di particolari attributi (es. cercare solo **che** o solo **ché** o solo **chè**); attivando la *ricerca espansa* si può invece precisare che si vogliono trovare tutte le parole costituite dalla stessa sequenza di lettere, a prescindere dagli attributi (in tal caso la ricerca di **ché** o di **che** fornirebbe in risposta **che**, **ché**, **chè**, ecc.).

Varianti

Si può costituire una tabella associativa, unica per tutti i corpus, nella quale a forme indicate dall'utente come *forme base* è possibile associare liberamente altre forme, dette *forme varianti*. Una volta preparata questa tabella, peraltro modificabile in ogni momento, è possibile estendere automaticamente la ricerca di una forma coincidente con una forma base inclusa nella tabella a tutte le forme varianti ad essa associate, così come rimane possibile limitare la ricerca alla sola forma base.

Iniziale raddoppiata

Se la lettera iniziale di una forma è una consonante, si può estendere automaticamente la ricerca alla parola uguale a quella impostata ma caratterizzata dal raddoppiamento dell'iniziale (ad esempio, la ricerca di **che** può essere così estesa a **cche** senza inserire quest'ultima in un altro selettore come parola da cercare a se stante).

Escludi elemento

Questa opzione definisce il selettore corrispondente come *negativo*, ovvero le forme selezionate in base a quel selettore verranno escluse dalla ricerca.

Il meccanismo di esclusione opera così: dapprima vengono localizzate tutte le forme derivanti dagli altri selettori, poi da queste vengono tolte quelle che soddisfino le condizioni di un qualunque selettore negativo presente.

L'ordine di successione dei selettori (*positivi* e *negativi*) non ha rilievo.

Ovviamente nella finestra di ricerca deve essere sempre presente almeno un selettore *positivo*.

Opzione '<=>'

Normalmente le forme cercate sono esattamente quelle impostate nei selettori; cliccando il simbolo di freccia in basso posto a sinistra del selettore appare un *menu a tendina* nel quale, oltre al simbolo = di

default, si possono cliccare i simboli < e >. Attivando il primo dei due la ricerca localizzerà tutte le forme alfabeticamente precedenti quella inserita nel selettore; attivando il secondo si otterranno tutte le forme alfabeticamente successive. L'uso di questi simboli è incompatibile con la presenza di caratteri jolly nel selettore.

SELETTORI MULTIPLI

Può capitare di voler imporre più condizioni alla stessa forma cercata. Si supponga ad esempio di cercare tutte le forme che comincino per **arma** e finiscano per **are**.

Una prima soluzione potrebbe essere quella di digitare in un selettore la stringa **arma*are**, per indicare la ricerca di forme caratterizzate dalle parti iniziale e finale volute; il risultato sarebbe simile a quello desiderato ma si perderebbe la forma **armare**, che si vorrebbe invece ritrovare, se presente nel dominio di ricerca, in quanto soddisfacente entrambi i requisiti richiesti (il problema è causato dalla presenza della doppia **a** nella stringa di ricerca permetterebbe di trovare, al massimo, la forma **armaare** se esistesse).

Una seconda soluzione potrebbe allora consistere nel collocare in due selettori diversi le due stringhe **arm*** e ***are**; questa volta il problema sarebbe opposto, nel senso che verrebbero localizzate tutte le forme soddisfacenti almeno uno dei due requisiti imposti e tra queste solo un sottoinsieme, la cui individuazione sarebbe rimessa all'utente, li soddisferebbe entrambi.

GATTO offre uno strumento aggiuntivo per risolvere questo specifico problema: si tratta del ricorso al meccanismo dei *selettori multipli*, ovvero ad una tecnica che consiste nel digitare le diverse condizioni imposte in selettori distinti, come è stato fatto nel caso precedente, precisando però che si tratta di condizioni da applicare contemporaneamente alle stesse forme, che quindi andranno presentate solo se li soddisfano entrambi. In questo modo nell'esempio precedente verrebbero localizzate tutte le forme che cominciano per **arma** e contemporaneamente finiscono per **are**, compresa **armare**.

RICERCA PER BRANI

Il ricorso ai selettori non è l'unico modo per impostare una ricerca per forme. Esiste un secondo meccanismo, indicato col termine *ricerca per brani*, che consente l'introduzione di testo, trasferito in un'apposita finestra di GATTO a partire da un documento accassibile dal computer tramite meccanismi di *copia e incolla* o per digitazione diretta. Avviando la ricerca il brano verrà analizzato e verranno estratte tutte le forme che vi compaiono, scartando numeri o altre stringhe non riconosciute come forme valide; delle forme così individuate potranno essere cercate le occorrenze presenti nel dominio di ricerca.

Questo tipo di ricerca mostra anche, nella risposta, le forme presenti nel brano di testo ma non nel corpus. Queste forme vengono evidenziate in rosso; ad esse non sono associate occorrenze né lemmi, come ovvio, e sono riportate solo in quanto può essere di interesse sapere quali, tra le forme citate nel testo di partenza, non risultano presenti nel corpus. Queste forme, anche se selezionate, non vengono copiate nell'ACCUMULATORE, trattandosi di un'operazione inutile.

RICERCA DI FORME MAI LEMMATIZZATE

Questo tipo di ricerca è regolato dalle stesse opzioni descritte finora ma, come accennato in precedenza, limita le forme trovate a quelle che risultano non essere state mai lemmatizzate all'interno del dominio di ricerca; in altri termini questa dicitura si riferisce alle forme delle quali non esistono occorrenze lemmatizzate all'interno del dominio di ricerca, a prescindere dalla situazione esistente al suo esterno; in particolare questo significa che le forme in oggetto possono anche disporre di un dizionario di macchina non nullo, dal momento che questo si applica all'intero corpus, nel caso in cui le associazioni forma-lemma che lo hanno popolato siano state originate in testi del corpus esterni al dominio di ricerca considerato al momento.

RELAZIONE TRA RICERCA PER FORME E USO DEI LEMMI NELLA RICERCA

Questo capitolo tratta di *ricerche per forme*, trattandole come cosa diversa e distinta dalle *ricerche per lemmi*. Ciò nonostante come si sarà notato lo stato di lemmatizzazione delle occorrenze cercate entra in gioco frequentemente anche in questo tipo di ricerca. Per capire chiaramente le interrelazioni tra questi elementi è opportuno porre attenzione al fatto che la ricerca per forme è uno strumento che opera ricerche di occorrenze

a partire dalle forme indicate, salvo definire in un secondo tempo e in misura più restrittiva le occorrenze effettivamente volute, magari in base al loro stato di lemmatizzazione. La *ricerca per lemmi* parte invece dalla digitazione di questi ultimi come elementi di avvio della ricerca, la quale può puntare quindi ad occorrenze di forme diverse, collegate solo dalla comune associazione ai lemmi indicati; tuttavia specifiche opzioni disponibili consentono di riportare o estendere l'obiettivo finale alla ricerca delle sole occorrenze non lemmatizzate e/o lemmatizzate con lemmi diversi da quelli indicati in partenza.

Con ciò i due tipi di ricerche (per forme e per lemmi) non sono affatto interscambiabili ma si integrano nel consentire di ottenere una ampia gamma di risultati diversi.

PASSAGGIO DIRETTO AI CONTESTI

A prescindere dalle modalità della ricerca impostata, spesso quest'ultima concerne forme specifiche delle quali non interessano varianti, lemmatizzazioni o altro: ad esempio potrebbe essere il caso della ricerca di tutte le occorrenze della forma **placido**, caratterizzata da univoca e ben definita accezione. In questo caso il passaggio attraverso le varie fasi della ricerca sarebbe un'inutile perdita di tempo, facilmente evitabile ricorrendo alla *ricerca con passaggio diretto ai contesti*.

VISUALIZZAZIONE CONTEMPORANEA DI NOTE E TRADUZIONI

Quando le occorrenze trovate sono incluse in parti di testo cui è associata una nota è sempre possibile vedere quest'ultima in associazione col contesto presentato. Allo stesso modo è possibile vedere insieme ai contesti anche quelli corrispondenti inseriti in un testo associato contenente una traduzione o una diversa edizione: in questo caso i brani corrispondenti possono essere allargati o ristretti con effetti di zoom parallelo.

DATI ASSOCIATI AI CONTESTI ESTRATTI

Ai contesti estratti dal dominio di ricerca a seguito di un'interrogazione sono associate informazioni aggiuntive quali dati bibliografici e statistici dei testi dai quali sono stati prelevati, oltre a lemmi e iperlemmi utilizzati nel caso di occorrenze lemmatizzate e alla possibilità di prendere visione dell'intera parte di testo afferente allo stesso riferimento organico.

LEZIONE 26. COME EFFETTUARE RICERCHE PER FORME

Una serie di esempi mostrerà come effettuare concretamente le ricerche e come utilizzare le opzioni disponibili. Le finestre coinvolte contengono talvolta menu o altri oggetti grafici caratterizzati da diciture e azioni che differiscono solo per il fatto di essere applicate ad ambiti diversi (come ad esempio il menu [SEL/DESEL.] presente sia nella finestra FORME LOCALIZZATE che nell'ACCUMULATORE): il funzionamento di questi elementi verrà esemplificato in un ambito solo, intendendo che il meccanismo di funzionamento descritto valga anche negli altri casi.

Esempio 1: ricerca di fratello

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE].

Tramite [CORPUS] andare a selezionare il corpus **Demo**, aprirlo con [APRI], quindi tornare alla finestra MODALITÀ RICERCA con [CHIUDI FINESTRA].

Attivare le ricerche per forme tramite [RICERCHE.. + PER FORME] e digitare **fratello** nel primo selettore. Per avviare la ricerca usare [AVVIO RICERCA] o digitare <Invio>. La forma **fratello** risulta presente in **Demo** con 5 occorrenze. Selezionarla cliccando in un punto qualunque della riga di **fratello**: nella colonna **sel.** apparirà una **x** a indicare che la riga della griglia è stata selezionata. Copiare nell'ACCUMULATORE la riga selezionata tramite [COPIA IN ACC.] o digitando <Invio>. La nuova finestra è simile alla precedente ma mentre quella conteneva solo i risultati dell'ultima ricerca, la finestra attuale, ovvero l'ACCUMULATORE, contiene i risultati di tutte le ricerche effettuate finora da quando si è entrati nell'ambiente **Ricerche**: naturalmente essendo l'attuale ricerca anche l'unica effettuata i due contenuti in questo caso coincidono. Selezionare nuovamente la riga di **fratello** cliccando nella riga corrispondente e passare alla visualizzazione dei contesti con [MOSTRA CONTESTI] o digitando <Invio>: si arriva così ai 5 contesti presenti nel corpus.

Tornare all'impostazione di una nuova ricerca con [VAI A.. + IMPOSTAZIONE RICERCA]. Si può notare che il contenuto della finestra di partenza non è stato cancellato.

Esempio 2: ricerca di **casa** con la variante **chasa**, singolare e plurale

Vuotare il selettore contenente la forma **fratello** con [CANCELLA TUTTI I DATI].

Digitare in un selettore qualunque (non occorre utilizzarli nell'ordine) la stringa **c<h>as<a,e>** con la quale si cercano le occorrenze di **casa** e della sua variante **chasa**, sia al singolare che al plurale (infatti la prima coppia di parentesi angolari indica che tra **c** e **as** deve essere presente una **h** oppure nessun carattere, mentre la seconda richiede che la forma termini per **a** oppure per **e**). Attivare l'opzione **mostra lemmi** cliccando sul *check box* relativo per distinguere le occorrenze trovate in base alla loro lemmatizzazione e avviare la ricerca con [AVVIO RICERCA]. Questa volta la finestra COPPIE FORMA-LEMMA LOCALIZZATE contiene 7 record, relativi ad altrettante forme distinte anche in funzione della lemmatizzazione. Ogni riga della griglia è relativa alle occorrenze di una certa forma lemmatizzata con un certo lemma o non lemmatizzata; nel caso in questione GATTO ha individuato nel corpus 77 occorrenze non lemmatizzate di **casa**, 6 occorrenze di **casa** lemmatizzate con **casa (s.f.)**, 4 di **case** di cui due non lemmatizzate e due lemmatizzate ancora con **casa (s.f.)**, 5 di **chasa**, di cui 2 lemmatizzate e 3 no e 1 di **chase**, lemmatizzata; il numero complessivo di occorrenze presenti si ottiene tramite [N. ELEMENTI]: sono 93. Volendo vedere le occorrenze raggruppate per forma a prescindere dalla lemmatizzazione cliccare sul bottone (**Solo forme**): le righe si riducono a 4, tante quante sono le forme coinvolte, mentre le occorrenze sono sommate per ogni forma. Ripristinare la visualizzazione a 7 record con (**Mostra lemmi**).

Come indica il riquadro a sfondo celeste a destra i 7 record sono ordinati **per sole forme** (ovvero senza tener conto dell'ordinamento alfabetico dei diversi lemmi associati alla stessa forma). Volendo proseguire la ricerca trascurando per il momento le 77 occorrenze non lemmatizzate di **casa** selezionare tutti i 7 record con [SEL/DESEL.. + SELEZIONA TUTTO] poi, tenendo premuto <Ctrl>, cliccare sulla prima riga per deselezionare **casa** non lemmatizzata; i numeri degli elementi presenti e selezionati si possono conoscere con [N. ELEMENTI] il quale indica l'esistenza di 6 elementi selezionati su 7 presenti; riutilizzando [N. ELEMENTI] il riquadro a sfondo giallo si richiude. Per procedere con la ricerca degli elementi indicati nelle 6 righe selezionate cliccare su [COPIA IN ACC.]. L'ACCUMULATORE contiene ora 7 righe, relative all'esito di questa ricerca combinato con quello della precedente (**fratello**). Selezionare tutte le righe con [SEL/DESEL.. + SELEZIONA TUTTO] poi, tenendo premuto <Ctrl>, cliccare sulla riga di **fratello** per deselezionarlo. Procedere con [MOSTRA CONTESTI]: si ottengono così i 16 contesti cercati, ordinati cronologicamente per testo e, a parità di testo, alfabeticamente per forma (le occorrenze della stessa forma compaiono nell'ordine di apparizione nel testo).

Tornare all'impostazione di una nuova ricerca con [VAI A.. + IMPOSTAZIONE RICERCA].

I risultati di una ricerca, ovvero i contesti ottenuti, rimangono disponibili finché non viene effettuata una nuova ricerca: per vederlo, da RICERCA PER FORME tornare a RISULTATI RICERCA mediante [VAI A.. + CONTESTI ESTRATTI]: i 16 contesti precedentemente ottenuti sono ancora memorizzati e utilizzabili per successive stampe o altro.

Esempio 3: ricerca delle forme riconducibili alla *famiglia lessicale adulterare*

Vuotare la finestra RICERCA PER FORME con [CANCELLA TUTTI I DATI] e disattivare il *check box mostra lemmi*.

Digitare in un selettore la stringa seguente: <,a><b,d,v><o,u><l,r><t,d><e,o,u>r*

Procedere con [AVVIO RICERCA]: nel caso del corpus **Demo** si trovano solo le 2 forme **adulteru** e **vulturnu**, mentre in un corpus di dimensioni maggiori una ricerca di questo tipo consentirebbe di reperire con una sola operazione un elevato numero di forme diverse riferite alla famiglia lessicale indicata. L'esempio riportato ha lo scopo di mostrare come, utilizzando opportunamente certe funzionalità, sia possibile impostare ricerche molto ampie: quella presentata cerca, contemporaneamente, 192 (2 x 3 x 2 x 2 x 2 x 4) gruppi di forme, ognuno definito in base ai caratteri iniziali (il selettore termina con *).

Esempio 4: ricerca di forme inizianti per **f** e terminanti per **are**

Vuotare la finestra RICERCA PER FORME con [CANCELLA TUTTI I DATI] e disattivare eventualmente il *check box mostra lemmi*.

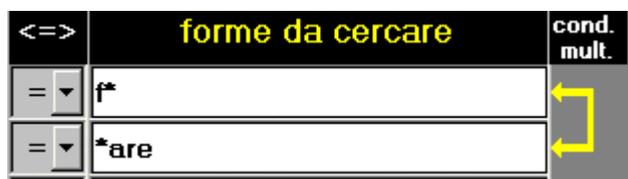
Digitare **f*** in un selettore e ***are** nel successivo, per trovare le forme che soddisfino entrambe le condizioni, e procedere con [AVVIO RICERCA].

Nella finestra FORME LOCALIZZATE il menu [N. ELEMENTI] indica che sono state localizzate 1229 forme diverse e il riquadro di destra indica che questi elementi sono **non ordinati**. La causa del mancato ordinamento risiede nel fatto che gli elementi localizzati superano il limite oltre il quale l'operazione non è più eseguita automaticamente; questo comportamento è mirato ad ottimizzare i tempi di risposta evitando di effettuare operazioni non espressamente volute e che potrebbero richiedere un certo tempo. Il limite di default di GATTO è 300 elementi ma può essere modificato agendo sulle *opzioni generali* del programma (vedere Appendice B - *Impostazione delle opzioni generali*). Volendo ordinare gli elementi trovati in ordine crescente di forme cliccare su (A-Z); volendo invece ordinarli ad esempio in ordine decrescente di occorrenze cliccare sulla freccetta a destra della scritta **forme** nella coppia di riquadri **ordina per e forme**, cliccare su **occorrenze** e quindi su (Z-A): si trova che la forma a più alta frequenza tra quelle localizzate è **fo** con 322 occorrenze, seguita da **fu** con 239.

Come è evidente è stato commesso un errore di impostazione nella ricerca poiché sia **fo** che **fu** iniziano per **f** ma non terminano per **are** e quindi non soddisfano entrambe le condizioni, come era invece la richiesta di partenza; se si ordinano le forme localizzate in ordine alfabetico decrescente utilizzando il riquadro di ordinamento si noterà che tra le forme ottenute compaiono **votare**, **volgarizzare** e via dicendo, che terminano per **are** ma non cominciano per **f**. La spiegazione risiede nel fatto che sono state effettuate in pratica due ricerche separate anche se contemporanee, quella delle *forme che iniziano per f* e quella delle *forme che finiscono per are*, come si può vedere tornando alla finestra RICERCA PER FORME con [VAI A.. + IMPOSTAZIONE RICERCA].

Volendo vedere ancora le forme ottenute tornare alla finestra FORME LOCALIZZATE con [VAI A.. + ELEMENTI LOCALIZZATI]. Riportarsi nella finestra RICERCA PER FORME con [VAI A.. + IMPOSTAZIONE RICERCA]. Notare che se in partenza sono stati usati due selettori diversi dai primi, tornando a RICERCA PER FORME le stringhe digitate sono state compattate verso l'alto, trasferendole nei primi due selettori.

Per correggere l'errore e imporre che entrambe le condizioni si applichino alle stesse forme cliccare nella zona grigia della colonna **cond. mult.** a destra, in corrispondenza dell'ultimo selettore usato, ovvero quello contenente la stringa ***are**. In quel riquadro e in quello sovrastante comparirà una coppia di frecce gialle collegate, per indicare che ora i due selettori sono connessi e formano un unico *selettore multiplo*, il quale applica tutte le condizioni così collegate alle forme trovate.



I *selettori multipli* possono essere formati da due o più selettori semplici: cliccando nella parte grigia di un selettore alternativamente lo si connette o lo si sconnette con quello immediatamente precedente. Pertanto cliccando in corrispondenza del selettore più in alto non si ottiene alcun risultato.

Se a questo punto si riparte con [AVVIO RICERCA] il risultato sarà completamente diverso: 13 forme tutte inizianti con **f** e terminanti con **are**. Peraltro in questo caso specifico non ci sarebbe stata alcuna controindicazione ad effettuare la ricerca inserendo semplicemente in un unico selettore **f*are**: si sarebbe ottenuto lo stesso risultato, come è facile verificare.

Esempio 5: ricerca di forme inizianti per **fa** e terminanti per **are**

A titolo di esercizio, collegandosi all'ultima affermazione fatta a proposito dell'esempio precedente, provare ad effettuare la ricerca delle forme inizianti per **fa** e terminanti per **are** nei due modi:

1. utilizzando un selettore semplice contenente **fa*are**,
2. utilizzando un selettore multiplo formato dai selettori semplici **fa*** e ***are**.

Si trova che nel primo caso si ottengono 5 forme, nel secondo caso 6: la differenza è la forma **fare** non localizzata nel primo caso, coerentemente con quanto detto in precedenza a proposito dell'utilità dei selettori multipli.

Esempio 6: ricerca di **re**, **regine** e **principi**

Oltre a quelli visti esistono altri tasti utilizzabili durante le ricerche: impostare nel primo selettore di RICERCA PER FORME **re**, portarsi sul secondo selettore con <↓>, digitare **regin?**, passare al terzo digitandovi **princip?** e avviare la ricerca con <Invio>.

Il risultato è 1 occorrenza di **principi**, 235 di **re** e 3 di **regina**. Per proseguire la ricerca le tre righe andrebbero selezionate e trasferite nell'ACCUMULATORE con [COPIA IN ACC.]. In quella griglia si sono però ormai assommati diversi risultati precedenti e potrebbe essere opportuno vuotarla preliminarmente. Per vedere la situazione portarsi nell'ACCUMULATORE con [VAI A.. + ACCUMULATORE]; poi tornare indietro con [VAI A.. + ELEMENTI LOCALIZZATI]. Vuotare l'ACCUMULATORE con [SVUOTA ACC.] (la disabilitazione di questo menu è una conferma che ora l'ACCUMULATORE è vuoto), poi selezionare le tre righe. Questa operazione può essere effettuata in 3 modi diversi, sia pur con identico risultato (vedere anche *Comandi di uso generale* nel capitolo 29):

1. usare il menu [SEL/DESEL.. + SELEZIONA TUTTO] (usando [SEL/DESEL.. + DESELEZIONA TUTTO] si ottiene l'effetto opposto, ovvero quello di rimuovere tutte le selezioni).
2. in alternativa, cliccare nella prima riga (quella contenente **principi**), quindi tenendo premuto <Ctrl> effettuare la stessa operazione in ciascuna delle due righe successive;
3. in alternativa attivare il *check box* **blocco selezionati** in basso a destra e quindi cliccare in ciascuna delle tre righe, senza necessità del tasto <Ctrl>;

A questo punto volendo proseguire la ricerca trasferire il contenuto nella griglia nell'ACCUMULATORE, risSelected i 3 record in uno dei modi descritti e passare alla visualizzazione dei record con [MOSTRA CONTESTI].

Esempio 7: ricerca di **lo** e controllo del numero di contesti da visualizzare

Con le procedure già viste inserire **lo** in un selettore di RICERCA PER FORME. Con le tecniche ormai note avviare la ricerca, trasferire la forma **lo** nell'ACCUMULATORE e poi passare alla visualizzazione dei contesti.

Una finestra centrale di dimensioni ridotte (CONTESTI LOCALIZZATI) aggiorna il numero dei contesti via via trovati, fermandosi a 300: questo è infatti il limite in corrispondenza del quale GATTO interrompe la ricerca e chiede come procedere; il valore 300 è preimpostato ma può essere modificato, come verrà visto tra breve. Va precisato che CONTESTI LOCALIZZATI appare sempre, non solo quando si effettua il passaggio diretto ai contesti: negli esempi precedenti il numero di contesti localizzati era inferiore e la finestra era quindi comparsa per un tempo brevissimo.

Tornando alla ricerca in corso, le scelte prospettate dal programma sono 3, indicate dai bottoni (**Continua**), (**Mostra**) e (**Stop**). Verranno ora illustrati gli effetti delle tre possibili continuazioni.

1. (**Continua**). Cliccando su (**Continua**) GATTO prosegue cercando le occorrenze ulteriori e arrestandosi dopo altre 300; arrivati a 600 disabilitare il *check box* **ripresenta questa richiesta** e cliccare di nuovo su (**Continua**): questa volta GATTO non si ferma a 900 ma procede fino al termine dei contesti esistenti. Per vederlo, nella finestra RISULTATI RICERCA visualizzarne il numero tramite [CONTESTI.. + NUM. CONT. TROVATI, SELEZ.]: sono 1074.
2. (**Stop**). Per provare questa seconda strada, dalla finestra RISULTATI RICERCA tornare preliminarmente all'ACCUMULATORE mediante [VAI A.. + ACCUMULATORE], quindi con la riga di **lo** selezionata ripetere [MOSTRA CONTESTI]. Cliccare su (**Stop**). Il risultato è che vengono mostrati i primi 300 contesti localizzati e la ricerca ha termine. Se (**Stop**) fosse stato scelto in seconda battuta, dopo 600 contesti, sarebbero stati trovati e mostrati 600 contesti e così via (per conoscere il numero dei contesti presenti nella finestra RISULTATI RICERCA usare come prima [CONTESTI.. + NUM. CONT. TROVATI, SELEZ.]. In sostanza (**Stop**) termina la ricerca e mostra i contesti localizzati fino a quel momento. Volendoli vedere tutti la ricerca va ripetuta, per lo meno a partire dall'ACCUMULATORE.
3. (**Mostra**). Come sopra, riportarsi sull'ACCUMULATORE e riavviare la localizzazione dei contesti. Arrivati a 300 cliccare su (**Mostra**). Vengono visualizzati i primi 300 contesti, esattamente come si otterrebbe con (**Stop**). In questo caso però la finestra RISULTATI RICERCA dispone di un menu non presente nei due casi precedenti, ovvero [VAI A.. + ULTERIORI CONTESTI], che consente volendolo di riprendere la ricerca dei contesti a partire dal punto nel quale era stata interrotta, ovvero dal contesto 301. Arrivati a 600, al solito, viene riproposta la domanda su come procedere.

Come accennato sopra il limite di 300 è preimpostato in GATTO ma è modificabile nell'ACCUMULATORE, tramite il menu [DEF. CONTESTI.. + NUMERO]: i valori accettati sono compresi tra 1 e 32.000. Il nuovo valore impostato rimane in vigore per tutta la sessione o fino ad ulteriore modifica. Volendo invece modificare il default di GATTO per mantenere il nuovo dato anche nelle sessioni successive vedere l'Appendice B - *Impostazione delle opzioni generali*.

Esempio 8: ricerca di **lo** con passaggio diretto ai contesti

Ripetere adesso la ricerca di **lo** illustrata nell'esempio precedente ma questa volta, tenuto anche conto che la ricerca coinvolge un'unica forma e non vi sono ulteriori selezioni da operare, saltare le finestre FORME LOCALIZZATE e ACCUMULATORE, passando direttamente ai contesti mediante [CONTESTI] (in luogo di [AVVIO RICERCA]). Con questa tecnica è sempre possibile scavalcare i passaggi intermedi della ricerca, passando direttamente dall'impostazione ai risultati. In questo caso si nota che CONTESTI LOCALIZZATI offre due sole scelte anziché tre, escludendo il bottone (**Mostra**): il comportamento è quello atteso e rappresenta un piccolo prezzo da pagare a fronte della maggior snellezza dell'operazione.

Per uscire dall'ambiente **Ricerche**, trovandosi in una delle finestre descritte utilizzare [ESCI.. + ALTRO AMBIENTE] oppure [ESCI] a seconda dei casi.

Esempio 9: scrittura su file o su carta dei contenuti delle griglie

I contenuti delle griglie presenti nelle finestre FORME LOCALIZZATE e ACCUMULATORE possono essere scritti su file o stampati. Per esplorare questa funzionalità impostare la ricerca di **caccia***, attivare il *check box mostra lemmi* e avviare la ricerca con [AVVIO RICERCA]. Per stampare su carta il contenuto della griglia usare [STAMPA.. + TUTTE LE RIGHE]: nella finestra IMPOSTAZIONI DI STAMPA cliccare sul *check box stampante* e scegliere (**OK**).

La stampa può essere limitata alle righe desiderate: a titolo di prova selezionare con le tecniche descritte negli esempi precedenti una o più coppie forma-lemma presenti nella griglia e provare a ripetere le stampe precedenti, avendo cura di selezionare il menu [STAMPA.. + RIGHE SELEZIONATE] in luogo di [STAMPA.. + TUTTE LE RIGHE].

In modo del tutto analogo si possono riportare i contenuti delle griglie su file. Provare ad usare [STAMPA.. + TUTTE LE RIGHE] selezionando poi il *check box file*: appare sulla destra un riquadro ulteriore il quale precisa che il file che verrà usato è di tipo **rtf**; inoltre viene fornita la possibilità di scegliere fra 3 strade:

1. creare un file nuovo,
2. inserire i dati inclusi nella griglia in un file esistente, accodandoli al suo contenuto attuale,
3. inserire i dati inclusi nella griglia in un file esistente, sostituendoli al suo contenuto attuale.

Mantenere l'opzione **file nuovo** e scegliere (**OK**): indicare un disco e una cartella a piacere e dare al file il nome **prova1**, poi eseguire l'operazione con (**Salva**).

Per vedere ed utilizzare il risultato, aprire Word o un altro word processor oppure semplicemente Notepad o altro editore, poi cercare nella cartella scelta il file **prova1.txt** (l'estensione viene aggiunta automaticamente). Il contenuto del file è illustrato qui di seguito:

corpus DEMO Coppie forma-lemma localizzate <data>

forma	lemma	cat. gr.	disambiguatore	occ.	commento
caccia				4	
caccia	caccia	s.f.		2	
cacciare	cacciare	v.		1	
cacciaro	cacciare	v.		1	
cacciato				1	
cacciato	cacciare	v.		1	
cacciatoricacciatore		s.m.		1	

Le prime righe riportano il nome del corpus usato, la natura della griglia della quale si è stampato il contenuto e la data.

Facendo uso di Word è possibile convertire questo testo in una tabella vera e propria. Per fare questo, da Word selezionare nel file tutta la parte compresa tra la parola **forma** e la fine del file; cliccare su **TABELLA + CONVERTI IL TESTO IN TABELLA...**, verificare che sia attivo l'*option box separa il testo a Tabulazioni* e poi cliccare su (**OK**). Il contenuto del file assumerà l'aspetto seguente (ulteriormente aggiustabile con i comandi di Word):

corpus DEMO Coppie forma-lemma localizzate <data>

forma	lemma	cat. gr.	disambiguatore	occ.	commento
caccia				4	

caccia	caccia	s.f.		2	
cacciare	cacciare	v.		1	
cacciato	cacciato	v.		1	
cacciato				1	
cacciato	cacciare	v.		1	
cacciatori	cacciatore	s.m.		1	

Con la stessa tecnica si possono riportare su file i soli elementi selezionati, tramite [STAMPA.. + RIGHE SELEZIONATE] e poi procedendo come prima.

LEZIONE 27. ALTRE OPZIONI DISPONIBILI. RIMARIO. ORDINAMENTI DEI RISULTATI

Nel corso della Lezione precedente sono stati passati in rassegna diversi casi di ricerche per forme, descrivendo il funzionamento delle opzioni via via incontrate; queste, a parte il meccanismo dei *selettori multipli*, si applicavano contemporaneamente a tutti gli elementi cercati. In questa Lezione verranno illustrate altre opzioni, impostabili in maniera diversa per ognuna delle forme cercate in modo da fornire un ulteriore grado di libertà all'utente. Alcune di queste opzioni, come **Ricerca espansa**, **Varianti**, **Iniziale raddoppiata** e **Escludi elemento** sono riportate in RICERCA PER FORME con diciture abbreviate che includono una lettera maiuscola, non sempre coincidente con la lettera iniziale (ad esempio *vAr*); il significato delle lettere maiuscole è che le corrispondenti opzioni possono essere alternativamente attivate o disattivate, con riferimento al solo selettore nel quale si trova il cursore, pigiando il tasto <Ctrl> insieme al tasto della lettera. Ad esempio, <Ctrl> + <r> attiva o disattiva la **Ricerca espansa** (naturalmente lo stesso risultato si può ottenere cliccando direttamente sul *check box* coinvolto).

Verrà poi visto come limitare i contesti da estrarre a quelli inclusi nel rimario di un testo.

Infine verranno presentati in maniera più completa i meccanismi di ordinamento delle righe, ovvero delle coppie forma-lemma, presenti nell'ACCUMULATORE e nelle altre finestre intermedie delle ricerche.

Opzione '<=>' (minore - uguale - maggiore)

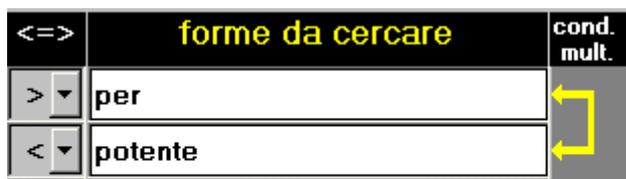
Per cominciare, portarsi in RICERCA PER FORME e digitare in un selettore **caccia**. Disattivare eventualmente il *check box mostra lemmi*.

A sinistra della parola digitata, ovvero **caccia**, c'è una colonna che riporta in testata il disegno <=>, indicante la possibilità di scegliere tra i simboli che indicano *minore*, *uguale*, *maggiore*. Per default la scelta è =, come si può vedere nella casella a sfondo grigio a sinistra di **caccia**; il significato di questa impostazione è che risulta impostata la ricerca di *forme uguali a caccia*. Cliccando sulla freccetta diretta in basso e posta immediatamente a sinistra di **caccia** si apre un *box a tendina* contenente le tre scelte possibili, ovvero =, < o >. Cliccare su <; in questo modo l'impostazione è stata convertita nella ricerca di *tutte le forme alfabeticamente precedenti (minori di) caccia*. Cliccando su [AVVIO RICERCA] troveremo questa volta 2358 forme, a partire da ' fino a **cacceranno** (per verificarlo, ordinare i risultati per forme alfabeticamente crescenti).

In modo analogo funziona l'opzione >.

Come detto prima, questa impostazione è specifica del selettore cui è applicata, e non ha nulla a che vedere con gli altri. A titolo di esempio tornare a RICERCA PER FORME, cancellarne il contenuto e impostare in un selettore **aba** con l'opzione < e in un altro **caccia** con l'opzione di default =. Si otterranno 8 elementi, ovvero le 7 forme alfabeticamente precedenti **aba**, più la forma **caccia**.

Come ultimo esempio di applicazione di questa opzione tornare a RICERCA PER FORME, cancellarne il contenuto e impostare in un selettore **per** con l'opzione > e nel successivo **potente** con l'opzione <, poi connettere i due selettori in un *selettore multiplo* cliccando sul riquadro grigio a destra di **potente**, ottenendo:



Avviare la ricerca: si otterranno le 690 forme incluse nel corpus e alfabeticamente comprese tra **pera** e **potendu**, ovvero quelle comprese tra **per** e **potente** con esclusione degli estremi.

Opzione 'Ricerca espansa'

Impostare la ricerca della forma **che** e procedere con [AVVIO RICERCA]: come prevedibile si otterrà come risposta la forma **che**, con 1935 occorrenze. Tornare a RICERCA PER FORME e ripetere la stessa ricerca attivando però l'opzione **Ricerca espansa** nella riga corrispondente al selettore nel quale è digitata la forma **che** e ripetere la ricerca. Questa volta si otterranno 4 forme, la prima delle quali è **che**, mentre le altre hanno in comune con essa le lettere componenti, ma differiscono per altri elementi.

La **Ricerca espansa** consente di ricercare tutte le forme che sono costituite dalla stessa sequenza di lettere di quella impostata, anche se ne differiscono per altri elementi come accenti, cediglie, pseudocaratteri.

Lo stesso risultato si sarebbe ottenuto avviando la ricerca da **ché**, **chè** o **che**.

Opzione 'Iniziale raddoppiata'

Se la forma impostata comincia con una consonante, l'attivazione di questa opzione ricerca automaticamente anche la variante costituita dalla stessa forma con la lettera iniziale raddoppiata. Ad esempio, impostare la ricerca di **casa** attivando questa opzione: si otterranno in risposta sia **casa** (83 occorrenze) che **ccasa** (2 occorrenze).

Opzione 'Escludi elemento'

È possibile escludere dai risultati alcune forme non desiderate utilizzando questa opzione. Tecnicamente lo stesso risultato si può ottenere in altri modi, ad esempio non trasferendo le forme indesiderate dalle FORME LOCALIZZATE all'ACCUMULATORE; tuttavia nel caso in cui la ricerca sia destinata a produrre un elevato numero di forme diverse può essere più semplice escludere a priori quelle non volute e la cui presenza è nota.

Si supponga a titolo di esempio di cercare le forme attinenti a **castello** e a questo fine si parta con la ricerca di **cast***: si otterranno 10 forme, solo le prime 6 delle quali fanno in qualche modo riferimento a **castello** a differenza delle altre, che sono peraltro caratterizzate dal cominciare per **casti** o **casto**. Uno dei modi per escludere i risultati non voluti è quello di reimpostare la ricerca nel modo seguente:

- in un selettore digitare, come prima, **cast***
- in un altro selettore digitare **casti*** e attivare l'opzione **Escludi elemento**
- in un terzo selettore digitare **casto*** e attivare l'opzione **Escludi elemento**

Ripetere la ricerca: questa volta si otterranno solo le 6 forme volute.

L'opzione **Escludi elemento** funziona così: dapprima vengono localizzate nel *dominio di ricerca* tutte le forme che soddisfano i requisiti fissati dai selettori *positivi* (cioè dai selettori normalmente impostati), poi da queste vengono rimosse tutte quelle che risultino eventualmente soddisfacenti uno o più selettori *negativi* (cioè aventi l'opzione **Escludi elemento** attiva); quindi, ad esempio, un selettore negativo che faccia riferimento a forme non incluse tra quelle cercate semplicemente non ha effetto: ad esempio la ricerca di **cas?**, in presenza di un selettore contenente **torre** con opzione **Escludi elemento** attiva, sarebbe del tutto equivalente alla ricerca di **cas?** in assenza di questo secondo selettore.

RICERCA LIMITATA AL RIMARIO

Normalmente le occorrenze localizzate nel corso di una ricerca possono trovarsi ovunque all'interno dei testi.

Nel caso di opere in versi esiste la possibilità di limitare le occorrenze estratte a quelle incluse nel rimario, collocate cioè a fine verso o caratterizzate dal codice di *rima al mezzo*.

Ad esempio, si imposti normalmente la ricerca di **vivi** e **stingua** e si arrivi ai contesti: si troveranno 8 occorrenze di **vivi** e 2 di **stingua**. Si torni ora a RICERCA PER FORME e, lasciando impostati i selettori con **vivi** e **stingua**, si attivi [RIMARIO.. + FINE VERSO]; innanzi tutto si noterà che il menu [RIMARIO..] è ora in maiuscoletto, per ricordare che l'opzione di rimario è attiva; procedere con [AVVIO RICERCA]: i numeri di occorrenze (2 e 8) vengono confermati e così pure quando si arriva all'ACCUMULATORE con [COPIA IN ACC.]; quando però si passa alla visualizzazione dei contesti ne viene mostrato solo uno relativo a **stingua**, l'unico riguardante un'occorrenza collocata a fine verso.

Si provi a ripetere la stessa ricerca tornando a RICERCA PER FORME e attivando invece [RIMARIO.. + RIMA AL MEZZO]: questa volta si otterrà una sola occorrenza di **vivi**, codificata come *rima al mezzo*.

Ripetere infine la stessa ricerca con [RIMARIO.. + F VERSO + RIMA ½]: come era facile aspettarsi il risultato consiste nella combinazione delle due occorrenze localizzate nel corso delle ricerche precedenti.

Da quanto visto dovrebbe risultare chiaro che la ricerca limitata al rimario opera in partenza come una ricerca estesa a tutto il testo e per questo motivo i numeri di occorrenze dichiarati sono gli stessi: è solo nella fase finale che le occorrenze non soddisfacenti i vincoli di appartenenza al rimario vengono scartate, potendosi perciò produrre come esito limite anche la completa assenza di contesti effettivamente localizzati.

ORDINAMENTO DI FORME, LEMMI E COPPIE FORMA-LEMMA PRESENTI NELLE FINESTRE *FORME LOCALIZZATE, LEMMI LOCALIZZATI, COPPIE FORMA-LEMMA LOCALIZZATE E ACCUMULATORE*

Sono stati già visti alcuni casi di ordinamento dei risultati ottenuti in queste finestre. Qui verrà fornita una panoramica completa degli ordinamenti possibili, spiegando il significato esatto di ognuno di essi.

Innanzitutto come conseguenza diretta di una ricerca gli elementi ottenuti vengono ordinati o meno a seconda del loro numero: in particolare vi è un numero limite di elementi trovati al di sotto del quale l'ordinamento ha luogo automaticamente mentre, se il loro numero è superiore, l'ordinamento è rimesso all'utente per evitare l'esecuzione non voluta di operazioni che richiedano un certo tempo. Questo numero limite (in partenza 300) può essere modificato dall'utente (vedere l'Appendice B - *Impostazione delle opzioni generali*).

In un secondo tempo l'ordinamento dei risultati può essere modificato dall'utente sulla base dei tipi di dati inclusi nella griglia, in ordine crescente o decrescente, utilizzando il *box a tendina* e i bottoni (A-Z) e (Z-A) posti sulla destra.

I tipi di ordinamento possibili sono i seguenti:

- **non ordinati:** gli elementi sono inseriti in maniera casuale;
- **ordinati per sole forme:** ordinamento per forme; le eventuali coppie costituite dalle stesse forme e da lemmi diversi si susseguono a caso;
- **ordinati per forme:** ordinamento per forme; se sono presenti anche i lemmi a parità di forma l'ordinamento è fatto per lemma, a parità anche di questo per categoria grammaticale, ad ulteriore parità per disambiguatore;
- **ordinati per forme inverse:** ordinamento per forme; se sono presenti anche i lemmi a parità di forma l'ordinamento è fatto per lemma, a parità anche di questo per categoria grammaticale, ad ulteriore parità per disambiguatore; l'ordinamento alfabetico di forme e lemmi non è effettuato sulla base delle loro lettere iniziali ma a partire da quelle finali, ovvero dalle terminazioni;
- **ordinati per soli lemmi:** ordinamento per lemmi; le eventuali coppie costituite dagli stessi lemmi e da forme diverse si susseguono a caso;
- **ordinati per lemmi:** ordinamento per lemmi; a parità di lemma l'ordinamento è fatto per categoria grammaticale, a parità anche di questa per disambiguatore, ad ulteriore parità per forme (se presenti);
- **ordinati per lemmi inversi:** ordinamento per lemmi; a parità di lemma l'ordinamento è fatto per categoria grammaticale, a parità anche di questa per disambiguatore, ad ulteriore parità per forme (se presenti); l'ordinamento alfabetico di forme e lemmi non è effettuato sulla base delle loro lettere iniziali ma a partire da quelle finali, ovvero dalle terminazioni;
- **ordinati per categorie grammaticali:** ordinamento per categorie grammaticali; a parità di categoria grammaticale l'ordinamento è fatto per lemma, a parità anche di questo per disambiguatore, ad ulteriore parità per forme (se presenti);
- **ordinati per disambiguatori:** ordinamento per disambiguatori; a parità di disambiguatore l'ordinamento è fatto per lemma, a parità anche di questo per categoria grammaticale, ad ulteriore parità per forme (se presenti);
- **ordinati per occorrenze:** ordinamento per numero di occorrenze; a parità di occorrenze l'ordinamento è fatto per forme, se presenti, altrimenti per lemmi.

LEZIONE 28. USO DELLE VARIANTI. SALVATAGGIO E RECUPERO DELL'ACCUMULATORE

Questa Lezione esemplificherà l'utilizzo di due funzioni accessorie relative alle operazioni di ricerca. La prima delle due, ovvero il ricorso alle varianti, è specifico delle ricerche per forme mentre la possibilità di

salvare su file il contenuto dell'ACCUMULATORE e recuperarlo nel corso di una sessione successiva, applicandolo allo stesso corpus od anche ad un corpus diverso, è possibile per tutti i tipi di ricerca.

USO DELLE VARIANTI

Si consideri la ricerca seguente, per forme: <c,cc,ch>as<a,e>. Se effettuata sul corpus **Demo** fornirà la seguente lista di forme: **casa, case, ccasa, chasa, chase**.

In questo caso con una ricerca non particolarmente complessa si è cercato di recuperare le diverse varianti di **casa**. Se si tratta di una ricerca destinata ad essere ripetuta frequentemente, magari in domini di ricerca diversi e in corpus diversi e immaginando che la varianza da controllare sia più ampia, l'operazione potrebbe essere pesante, soprattutto per il rischio di dimenticarsi talvolta qualcuna delle varianti. Il problema potrebbe essere semplificato in presenza di una tabella associativa del tipo:

forma base	forma variante
casa	ccasa
case	ccase
casa	chasa
case	chase

che in qualche modo avvertisse il programma di estendere una ricerca formalmente effettuata sulla *forma base* **casa** anche alle *forme varianti* riportate.

Come verrà visto, il meccanismo delle *varianti* inserito in GATTO offre esattamente questa possibilità.

Portarsi in RICERCA PER FORME. Cliccare su [VARIANTI]. Appare una finestra di dimensioni ridotte occupata per la maggior parte da una griglia probabilmente inizialmente vuota, destinata a contenere coppie forma base-variante. I menu nella parte alta consentono di effettuare le 3 operazioni fondamentali, ovvero inserimento, modifica e rimozione di una coppia, oltre a permettere l'uscita dalla finestra GESTIONE DATA BASE VARIANTI. Per inserirvi la tabella riportata sopra, attivare [AGGIUNGI] poi digitare nel riquadro **f. base** in basso la forma **casa** e nel sottostante riquadro **variante** la forma **ccasa**. Completare l'operazione con **(Registra)** (o annullarla con **(Annulla)**). La coppia di forme associate risulterà inserita nella tabella delle varianti. Ripetere l'operazione per le altre 3 coppie di forme riportate in tabella. Chiudere la finestra con [ESCI].

A questo punto impostare normalmente la ricerca di **cas?** e procedere con [AVVIO RICERCA]: vengono localizzate le 5 forme **casa, case, casi, casj** e **casu**. Ovviamente non vi è traccia di **ccasa, chasa** e relativi plurali.

Ripetere ora la stessa ricerca, attivando però l'opzione **vAr** in corrispondenza del selettore usato. Questa volta vengono localizzate anche **ccasa, chasa** e **chase** in quanto la tabella associativa ha provveduto ad estendere la ricerca di **casa** e **case** alle forme varianti inserite.

Tornare a RICERCA PER FORME, aprire la finestra GESTIONE DATA BASE VARIANTI con [VARIANTI], cliccare sulla riga contenente **case** e **ccase** e cliccare su [CANCELLA]: la coppia di forme è rimossa dalla tabella; ripetere l'operazione (si può anche usare il tasto <Canc>) con la coppia **case** e **chase**. Ora selezionare cliccandovi sopra la coppia **casa** e **ccasa** e usare [MODIFICA]; portarsi nel riquadro in basso **variante** e correggere **ccasa** in **ccas?**, poi completare l'operazione con **(Registra)**. Ripetere la sequenza con la coppia **casa** e **chasa**, trasformandola nella coppia **casa** e **chas?**. Chiudere la finestra.

A questo punto ripetere la solita ricerca, con l'opzione **varianti** attiva: anche in questo caso le forme varianti sono tutte localizzate. La possibilità di uso dei caratteri jolly nelle forme base e nelle forme varianti incrementa notevolmente la potenza della tabella stessa.

Per concludere, tornare nella finestra GESTIONE DATA BASE VARIANTI e rimuovere le due righe presenti: si trattava di un'esercitazione e va tenuto presente che la tabella non è specifica del corpus in uso, in questo caso **Demo**, ma dell'installazione di GATTO, per cui può essere utilizzata in tutte le ricerche successive.

Poiché la tabella associativa gestita con questo meccanismo è unica per tutti i corpus disponibili, a prescindere dalle loro impostazioni, al fine di evitare conflitti nel campo **variante** non sono ammesse le cifre arabe, a prescindere dai parametri stabiliti per il corpus correntemente in uso.

SALVATAGGIO E RECUPERO DELL'ACCUMULATORE

Come detto in precedenza l'ACCUMULATORE può contenere i risultati di più ricerche precedenti. L'analisi dei contesti relativi agli elementi presenti nell'ACCUMULATORE, specie se il dominio di ricerca è ampio e i contesti stessi sono numerosi, può richiedere molto tempo e suggerire di interrompere ad un certo punto il lavoro per riprenderlo in una sessione successiva. Questo implicherebbe la ripetizione delle ricerche eseguite fino a quel momento per ricostruire il contenuto dell'ACCUMULATORE.

Per aggirare questo problema esiste la possibilità di salvare il contenuto dell'ACCUMULATORE, tutto o in parte, in un file per richiamarlo nelle sessioni successive. Tra l'altro, l'ACCUMULATORE così recuperato può essere applicato anche ad un corpus diverso da quello di partenza.

Per sperimentare questo meccanismo, portarsi in RICERCHE PER FORME, digitare **cari***, attivare il *check box* **mostra lemmi**, avviare la ricerca e copiare le 14 coppie forma-lemma ottenute nell'ACCUMULATORE.

Cliccare su [FILE.. + SALVA + INTERO ACCUMULATORE]: si può scegliere se creare un nuovo file o utilizzare un file già esistente. Mantenere l'opzione **file nuovo** e scegliere (**OK**): indicare un disco e una cartella a piacere e dare al file il nome **prova2**, poi eseguire l'operazione con (**Salva**). Viene creato il file **prova2.acc** (l'estensione **acc** viene aggiunta automaticamente).

Tramite [ESCI] seguito da [FINE] uscire da GATTO.

Riavviare GATTO, andare nell'ambiente **Ricerche**, aprire il corpus **Demo**, impostare la ricerca per forme di **pensav*** con **mostra lemmi** attivo, avviare la ricerca e trasferire le 4 coppie forma-lemma così ottenute nell'ACCUMULATORE. In questo modo è stata chiusa la precedente sessione di lavoro e ne è stata aperta un'altra; il fatto che si stia utilizzando lo stesso corpus non è particolarmente significativo.

A questo punto, per recuperare il contenuto del file **prova2.acc** e riportarlo nell'ACCUMULATORE cliccare su [FILE.. + CARICA], localizzare **prova2.acc** e aprirlo con (**Apri**). Il contenuto del file viene recuperato e reinserito nell'ACCUMULATORE aggiungendolo al contenuto preesistente; qualora si fosse voluto sostituire il contenuto del file a quello attuale dell'ACCUMULATORE bastava far precedere l'operazione dalla cancellazione di quest'ultimo tramite [SVUOTA ACC.]; volendolo fare ora, svuotare l'ACCUMULATORE e ricaricare il contenuto di **prova2.acc**.

Come si può notare le coppie forma-lemma recuperate dal file presentano una caratteristica nuova, ovvero in luogo del numero di occorrenze riportano la dicitura **indef.** (per *indefinito*); questo rispecchia il fatto che il programma non *sa* se l'attuale dominio di ricerca è lo stesso dal quale tali coppie forma-lemma e relativi numeri di contesti sono stati ricavati, quindi questi ultimi potrebbero essere non più esatti; anzi, alle coppie forma-lemma riportate potrebbe non corrispondere in questo caso nessuna occorrenza.

LEZIONE 29. VISUALIZZAZIONE DEI CONTESTI OTTENUTI

Il risultato di una ricerca consiste in uno o più contesti estratti dal dominio di ricerca. Questi contesti possono essere mostrati in formati diversi, a seconda che si desideri privilegiarne il numero o le dimensioni. Associate ad ogni occorrenza sono inoltre disponibili informazioni direttamente presentate a video o ottenibili cliccando su specifiche zone del contesto.

Questa Lezione si propone, operando sui risultati di alcune ricerche, di mostrare le modalità di spostamento tra i contesti estratti, le diverse rappresentazioni dei contesti, i comandi per cambiarne le dimensioni e per accedere alle informazioni che vi sono associate.

Sempre operando sul corpus **Demo**, si effettui la ricerca per forme di **perd***, senza attivare il *check box* **mostra lemmi** né altre opzioni. Si ottengono 41 forme. Selezionarle e copiarle tutte nell'ACCUMULATORE svuotato dell'eventuale contenuto precedente; selezionarle ancora tutte e mostrarne i contesti. Nella finestra RISULTATI RICERCA si ottengono 80 contesti, come si può verificare tramite [CONTESTI.. + NUM. CONT. TROVATI, SELEZ.].

SPOSTAMENTI TRA I CONTESTI

In partenza vengono visualizzati i primi contesti localizzati: il numero di quelli visibili dipende ovviamente dalle dimensioni della finestra. Per passare agli altri contesti usare:

- i tasti <↓> e <↑> per spostarsi in avanti o indietro di un contesto;
- i tasti <Pag↓> e <Pag↑> per spostarsi in avanti o indietro di una schermata;
- il tasto <↖> per portarsi al contesto n. 1;

- il menu [VAI A.. + CONTESTO N.] per portarsi ad un determinato contesto, definito dal suo numero d'ordine;
- la barra di scorrimento laterale per spostarsi in avanti o indietro in maniera più generale.

Un altro caso di passaggio ad altri contesti, già descritto in precedenza, si ha quando nella finestra CONTESTI LOCALIZZATI si fa uso del bottone (**Mostra**), chiedendo di visualizzare i contesti localizzati fino a quel momento ma riservandosi la possibilità di procedere successivamente alla ricerca e visualizzazione dei successivi: in questo caso nella finestra RISULTATI RICERCA è presente il menu [VAI A.. + ULTERIORI CONTESTI] che consente di riprendere la ricerca e localizzare altri contesti.

Superamento del numero di 32.000 contesti ottenuti

La barra di scorrimento laterale, così come gli altri comandi, consente di muoversi tra tutti i contesti ottenuti con la ricerca purché il loro numero non superi 32.000.

Qualora questo limite venga superato (GATTO è in grado di generare fino a un massimo di circa 20 milioni di contesti, a prescindere da considerazioni di tempo necessario e di spazio sufficiente su disco), viene automaticamente generata una seconda barra di scorrimento, a destra della prima e più corta. In questo caso le due barre funzionano come segue:

- quella standard (più a sinistra e più lunga) consente di muoversi entro un intervallo di 32.000 contesti: tutti i meccanismi di spostamento descritti in precedenza rimangono validi e si intendono riferiti all'intervallo di 32.000 contesti controllato da questa prima barra di scorrimento;
- l'altra barra consente di passare agli altri gruppi di 32.000 contesti: i comandi di navigazione precedenti consentono allora di muoversi entro l'intervallo corrente di 32.000 contesti di volta in volta selezionato tramite la barra di destra.

ALTRE MODALITÀ DI VISUALIZZAZIONE DEI CONTESTI

Ogni contesto viene presentato su 3 righe, la prima di testata e le altre due contenenti il contesto vero e proprio, con l'occorrenza localizzata evidenziata in rosso. Se si contano le parole costituenti il primo contesto mostrato, da **tarda la speranza...** a **...più li agrata**, si nota che le parole sono in tutto 31, ovvero l'occorrenza localizzata, preceduta e seguita da 15 parole; normalmente gli stessi numeri definiscono anche le lunghezze degli altri contesti mostrati; se si modificano le dimensioni della finestra cambia il numero dei contesti mostrati ma non la loro struttura e lunghezza a meno che, restringendo la finestra, la sua larghezza non divenga insufficiente per mostrare i contesti completi utilizzando due righe: in questi casi ne viene tagliata la parte finale.

Alcuni contesti possono essere più corti di 31 parole. È il caso ad esempio del n. 3, nel quale le parole successive a quella cercata sono 10: questo è dovuto al fatto che i contesti sono comunque limitati dall'eventuale raggiungimento dei limiti della zona di testo associata ad un riferimento organico; in altre parole uno stesso contesto non può essere a cavallo di un riferimento organico.

La lunghezza preimpostata di 31 parole non è fissa; per modificarla si può utilizzare il menu [CONTESTI.. + LUNGHEZZA] scegliendo un qualunque valore compreso tra 1 e 81. Se il valore scelto è pari viene trasformato nel numero dispari immediatamente superiore, per garantire la simmetria del contesto rispetto alla parola centrale. Lunghezze maggiori sarebbero inutili perché essendo comunque riservate due righe per ogni contesto l'eccessivo allungamento della sua prima metà finirebbe per *espellere* dallo schermo proprio la parola centrale, togliendo significato alla ricerca. Il valore preimpostato può essere modificato anche nell'ACCUMULATORE, quindi prima di visualizzare i contesti, col menu [DEF. CONTESTI.. + LUNGHEZZA] (volendo invece modificare questo valore in maniera stabile, senza doverlo riassegnare ogni volta, vedere l'Appendice B - *Impostazione delle opzioni generali*).

Contesti kwic

La tipologia di visualizzazione dei contesti sulla base di 3 righe di schermo è indicata nella terminologia di GATTO come *standard*. Esistono altre modalità di visualizzazione. Per vederle, nella finestra RISULTATI RICERCA cliccare su [CONTESTI.. + TIPO + KWIC (RIF. TOP.)]: i contesti vengono ora mostrati in modalità kwic (**KeyWord In Context**), ovvero riservando a ciascuno una riga ed allineando al centro le parole localizzate; nella parte sinistra è riportato il riferimento topografico. Utilizzando ora [CONTESTI.. + TIPO + KWIC (RIF. ORG.)] si otterrà una visualizzazione molto simile alla precedente, nella quale però la parte sinistra riporta i caratteri iniziali del riferimento organico, in luogo di quello topografico. Il tipo di rappresentazione dei

contesti (standard o kwic) può essere impostato anche prima di visualizzarli, trovandosi nell'ACCUMULATORE e facendo uso in quella finestra del menu [DEF. CONTESTI.. + TIPO] e relativi sottomenu.

Contesti singoli

Se, trovandosi in una qualunque delle 3 modalità di visualizzazione descritte (standard, kwic con riferimento topografico o kwic con riferimento organico) si clicca in un punto qualunque del testo, escluse cioè la testata in modalità standard e la parte iniziale nei contesti kwic, il contesto verrà mostrato da solo. La sua geometria cambia e in particolare la lunghezza non è calcolata in base al numero di parole ma è estesa al periodo comprendente la parola centrale. In questa finestra le dimensioni del contesto possono essere modificate utilizzando i bottoni (**Allarga**) e (**Restringi**): il primo aggiunge ogni volta un periodo all'inizio e uno in coda, mentre il secondo ha l'effetto opposto. Il valore massimo è 15 periodi, come si può vedere portando in questa finestra ad esempio il contesto n. 1. Anche in questo caso un eccessivo allargamento del contesto può portare alla perdita della visualizzazione della parola centrale. Per tornare alla finestra a contesti multipli usare (**Esci**). Il limite non valicabile di appartenenza ad un unico riferimento organico vale anche in questo caso: per vederlo portare a contesto singolo il contesto n. 8: una volta allargato a 3 periodi le ulteriori espansioni non toccano l'inizio del contesto (**La namoranza - disiosa...**) che si trova evidentemente subito dopo il riferimento organico.

Se, sempre restando nella finestra a contesto singolo, si vuole dimensionare il contesto stesso in base al numero di parole, è sufficiente cliccare in basso a sinistra sull'*option box a parole*; se si vuole cambiare la lunghezza in parole preimpostata, digitarla nel riquadro **lung. contesto**: e poi attivarla con (**Applica**).

Per tornare alla visualizzazione per periodi utilizzare l'*option box a periodi*.

È anche possibile definire con esattezza il contesto voluto scegliendone la parola iniziale e finale a piacere, fermo restando che la parola centrale deve continuare ad essere compresa al suo interno. Ad esempio, si porti a contesto singolo il contesto n. 25 (**strolagi providero che,...**) e lo si espanda a 3 periodi. Si supponga di voler ritagliare il contesto trasformandolo in: **s'elli non stesse anni dieci che non vedesse il sole, che perderebbe il vedere. Allora il re il fece guardare in tenebrose spelonche il tempo detto**. Per ottenere questo risultato basta cliccare col mouse sulla prima parola voluta (**s'**) e sull'ultima (**detto**); l'operazione è ripetibile volendo ridurlo ulteriormente. In questo caso si attiva automaticamente il terzo *option box* in basso (**specifico**), non selezionabile direttamente. Per tornare alle altre visualizzazioni cliccare sugli *option box* corrispondenti. Notare che se si volesse ridurre in questo modo il contesto all'unica parola **perderebbe**, questa andrebbe cliccata una volta col tasto sinistro del mouse per eliminare la parte di contesto precedente e una volta col tasto destro per eliminare quella successiva.

L'utilità di queste diverse possibilità di definizione della geometria del contesto sarà vista nella prossima Lezione, parlando di stampa e trascrizione su file dei contesti; per ora è importante aver chiaro *come* effettuare queste operazioni.

Geometria assegnata al contesto entrando nella finestra a contesto singolo

Trasferendo un contesto dalla finestra a contesti multipli a quella a contesto singolo non sempre questo presenta la stessa geometria; è bene quindi aver chiaro il meccanismo che la determina:

- se il contesto non è selezionato per l'esportazione viene inizialmente visualizzato nella finestra a contesto singolo con lunghezza pari a un periodo;
- se il contesto risulta già selezionato per l'esportazione ma non in modalità kwic viene visualizzato nell'esatta geometria con cui verrà esportato; anche modifiche alla sua lunghezza successivamente effettuate in questa finestra non avranno effetto sull'aspetto del contesto al momento della stampa o scrittura su file (a meno che esso non venga deselezionato e rifelezionato con una diversa struttura); si provi ad esempio a cambiare nella finestra a contesto singolo la geometria di un contesto (già selezionato in precedenza) cliccando sugli *option box* in basso a sinistra, passando subito dopo alla finestra a contesti multipli e infine ritornando al contesto singolo: il contesto ripresenterà la stessa geometria mostrata nel precedente passaggio alla finestra a contesto singolo, ignorando le modifiche successive;
- se il contesto risulta già selezionato per l'esportazione in una delle due modalità kwic, non ottenibili nella finestra a contesto singolo, ai fini dell'esportazione verrà mantenuta la geometria fissata, mentre il contesto verrà visualizzato con lunghezza pari a un periodo.

Simboli usati per indicare la geometria assegnata a un contesto per l'esportazione.

I seguenti simboli possono comparire alla sinistra del numero d'ordine di un contesto per indicare la geometria con la quale verrà esportato:

simbolo	significato
N	a parole
P	a periodi
B	a brani ⁶
K	kwic, con riferimento topografico
W	kwic, con riferimento organico

STRUTTURE DEI CONTESTI E VISUALIZZAZIONE DELLE INFORMAZIONI ASSOCIATE

Finestra a contesti multipli, visualizzazione standard

Come detto in precedenza questa visualizzazione mostra ogni contesto su 3 righe, la prima delle quali è indicata come *riga di testata* e contiene una serie di informazioni, in parte visibili direttamente e in parte ottenibili cliccando su parti specifiche di tale riga.

Gli elementi che possono comparire nella struttura della riga di testata verranno ora passati in rassegna procedendo da sinistra verso destra.

1. **marcatore di contesto selezionato**, in blu scuro. Se presente, indica che il contesto è stato selezionato per l'esportazione e la lettera di cui è costituito indica la geometria fissata per questa operazione, in base allo schema riportato nella tabella precedente.
2. **numero di contesto**, in verde (in rosso se il contesto è stato selezionato per l'esportazione). Attribuisce al contesto un numero identificativo nell'insieme dei contesti localizzati. La numerazione serve, ad esempio, per identificare i contesti che si vogliono stampare o copiare su file;
3. **simbolo A (presenza di brano associato)**, in marrone. Cliccandolo si apre una finestra contenente una nota che riguarda tra le altre la parola centrale del contesto, oppure un brano tratto dal testo associato contenente una traduzione o una diversa edizione; in questo caso il brano mostrato è quello che corrisponde, nel testo base, al brano comprendente la parola centrale del contesto. La finestra può anche contenere sia una nota che un brano associato. Per distinguerli tenere presente che il brano associato è sempre preceduto da una riga contenente la sequenza "====". Se il brano associato consiste in una nota ma, per qualche motivo (non escluso un errore di preparazione dei testi) la nota stessa risultasse vuota, cioè priva di un testo, al suo posto apparirà la dicitura "NOTA VUOTA"; se ad un testo è associata una traduzione o una diversa edizione il simbolo **A** viene automaticamente inserito in tutti i suoi contesti, inclusi quelli per i quali non sia effettivamente previsto un brano corrispondente nel testo associato: in questo caso cliccando su **A** si otterrebbe in risposta la dicitura "BRANO PRIVO DI TRADUZIONE". Chiudere la finestra con <Esc> o cliccando sull'icona di chiusura **X** in alto a destra;
4. **simbolo # (indicatore di presenza di nota editoriale)**, in blu scuro. Avverte che tra i dati bibliografici del testo dal quale il contesto è stato estratto è presente una nota editoriale; cliccandolo si apre una finestra contenente i dati bibliografici e statistici dell'opera;
5. **titolo abbreviato**, in azzurro. Contiene l'inizio del campo **abbrev. titolo** incluso tra i dati bibliografici associati al testo. Cliccandolo anche in questo caso si apre la finestra contenente i dati bibliografici e statistici dell'opera dalla quale il contesto è stato estratto; ad esempio, sempre con riferimento agli 80 contesti derivanti dalla ricerca di **perd***, portarsi sul contesto n. 64; a sinistra del titolo abbreviato **Contr. Croce e Verg.** compare il simbolo **#**; cliccando su una di queste due zone si apre una finestra contenente i dati del testo **Contrasto fra la Croce e la Vergine**, di autore anonimo, a partire dalla sigla **d3**; l'ottavo campo contiene effettivamente una nota editoriale (**La stessa lauda...**) come indicato dalla presenza del simbolo **#**. Gli ultimi campi della finestra contengono informazioni statistiche, ultima delle quali è l'indicazione che nel testo sono presenti 16 occorrenze lemmatizzate. Chiudere la finestra con <Esc> o cliccando sull'icona di chiusura **X** in alto a destra;
6. **parte iniziale del riferimento organico**, in viola. È il riferimento organico corrente, al quale è associata la parte di testo da cui il contesto è stato estratto;

⁶ Vedere la Lezione 33, relativa alla visualizzazione di note o traduzioni.

7. **numero di verso**, in verde. Questa dicitura compare ovviamente solo per i brani in versi. Ad esempio, nel contesto 58 dopo il riferimento organico **28** si trova l'indicazione **v. 13**, indicante appunto che la parola localizzata si trova nel **verso 13** del riferimento organico **28**. In questo caso il numero di verso è stato preceduto dalla lettera **v**, ad indicare che il conteggio è stato fatto per versi singoli. Il quadro completo delle abbreviazioni che possono comparire per indicare il conteggio del verso è riportato nella tabella che segue:

dicitura	significato
v.	verso
d.	distico
t.	terzina
q.	quartina
o.	ottava
s.	stanza

8. **riferimento organico completo**. È una finestra che permette di leggere, sotto il titolo *riferimento organico completo*, tutto il riferimento organico nei casi in cui lo spazio nella riga di testata non sia sufficiente. Per aprirla cliccare in un punto qualunque della parte di riferimento organico riportata (in viola) o del numero di verso se presente. Ad esempio, portarsi sul contesto n. 71 e cliccare su questa zona: il riferimento organico è **L. 4, cap. 61**: in questo caso il dato riportato coincide con quello visualizzato ma, se quest'ultimo fosse stato troncato per eccesso di lunghezza, nella finestra RIFERIMENTO ORGANICO COMPLETO sarebbe stato possibile leggerlo integralmente.. Per chiudere questa finestra usare **<Esc>** o cliccare sull'icona di chiusura **x** in alto a destra;
9. **riferimento topografico**, in blu (numero di volume in celeste). È costituito dal numero di pagina seguito dal numero di riga nella pagina. Ad esempio, l'occorrenza cui corrisponde il contesto n. 15 si trova nella riga 17 della pagina 141. Dopo il numero di pagina può comparire un ulteriore dato numerico, indicante il numero di colonna per un documento tipograficamente ripartito appunto in colonne. Una scrittura come **310.2.16** indicherebbe la riga 16 della seconda colonna di pagina 310. Il dato può essere preceduto dal numero di volume, riportato in celeste: ad esempio **3.34.12** starebbe ad indicare la riga 12 della pagina 34 del terzo volume. Se il numero di pagina è preceduto dalla lettera **r** significa che le pagine sono indicate in numeri romani. Se il numero di pagina è seguito da una **r** o da una **v**, queste stanno rispettivamente ad indicare **recto** e **verso**; ad esempio **30r . 1** indicherebbe pagina 30 recto, riga 1 e **42v . 3** indicherebbe pagina 42 verso, riga 3. Nei casi in cui il riferimento topografico non viene visualizzato (funzione impostata come disabilitata o non obbligatoria all'atto della creazione del corpus) al suo posto compare la dicitura **TXT**;
10. **testo associato al riferimento organico**. Cliccando sul riferimento topografico si apre una finestra che permette di prendere visione dell'intera parte di testo associata al riferimento organico dalla quale il contesto è stato estratto. La finestra include una barra di scorrimento per muoversi nel testo; nei casi in cui la parte di testo mostrata sia particolarmente estesa vengono automaticamente abilitati i bottoni (**Avanti**) e (**Indietro**) per scorrere il testo a blocchi, muovendosi poi con la barra di scorrimento all'interno del blocco corrente. Nella stessa finestra è presente il bottone (**Dati bib.**) che permette di aprire anche da qui la finestra contenente i dati bibliografici e statistici del testo. Per chiudere la finestra usare **<Esc>** o l'icona di chiusura **x** in alto a destra;
11. **indice di qualità**, in marrone. Si tratta di un indicatore della qualità del testo, nel senso di attendibilità e pertinenza linguistica, congruità cronologica, ecc. Indica in sostanza un testo significativo per la documentazione della specifica varietà linguistica. Per i testi OVI valgono le seguenti codifiche: TS = testi significativi per la documentazione della specifica varietà linguistica; TPC = testi da citare solo in casi particolari per ragioni di cronologia; TPL = testi da citare solo in casi particolari per ragioni linguistiche;
12. **simboli @ o # (indicatori di lemmatizzazione)**, in verde. Dove presenti, stanno ad indicare che l'occorrenza in questione è stata lemmatizzata; in particolare, **#** indica che all'occorrenza è stato associato non solo un lemma, come nel caso di presenza di **@**, ma anche un iperlemma. Cliccandoli si apre una finestra contenente i dati del lemma e dell'eventuale iperlemma associati all'occorrenza. Vedere ad esempio il contesto n. 18, dove l'occorrenza di **perdeo** risulta lemmatizzata con **perdere** (**v.**). Chiudere la finestra con **<Esc>** o con l'icona di chiusura **x** in alto a destra.

Finestra a contesti multipli, visualizzazione kwic con riferimento topografico

In questo caso le informazioni collaterali, ovviamente in numero minore per problemi di spazio, si trovano nella stessa riga del contesto, alla sua sinistra.

1. **marcatore di contesto selezionato**, in blu scuro. Se presente, indica che il contesto è stato selezionato per l'esportazione e la lettera di cui è costituito indica la geometria fissata per questa operazione, in base allo schema riportato nella tabella precedente.
2. **numero di contesto**, in verde (in rosso se il contesto è stato selezionato per l'esportazione). Attribuisce al contesto un numero identificativo nell'insieme dei contesti localizzati;
3. **simbolo *** (**presenza di brano associato**), in nero. Cliccandolo si apre una finestra contenente una nota che riguarda tra le altre la parola centrale del contesto, oppure un brano tratto dal testo associato contenente una traduzione o una diversa edizione; in questo caso il brano mostrato è quello che corrisponde, nel testo base, al brano comprendente la parola centrale del contesto. La finestra può anche contenere sia una nota che un brano associato. Per distinguerli, tenere presente che il brano associato è sempre preceduto da una riga contenente la sequenza "====". Se il brano associato consiste in una nota ma, per qualche motivo (non escluso un errore di preparazione dei testi) la nota stessa risultasse vuota, cioè priva di un testo, al suo posto apparirà la dicitura "NOTA VUOTA"; se ad un testo è associata una traduzione o una diversa edizione il simbolo **A** viene automaticamente inserito in tutti i suoi contesti, inclusi quelli per i quali non sia effettivamente previsto un brano corrispondente nel testo associato: in questo caso cliccando su **A** si otterrebbe in risposta la dicitura "BRANO PRIVO DI TRADUZIONE". Chiudere la finestra con <Esc> o cliccando sull'icona di chiusura **X** in alto a destra;
4. **sigla del testo**, in celeste; cliccandola si apre la finestra contenente i dati bibliografici e statistici del testo da cui è tratto il contesto. Chiudere la finestra con <Esc> o cliccando sull'icona di chiusura **X** in alto a destra;
5. **riferimento topografico**, in blu (numero di volume in celeste). È costituito dal numero di pagina seguito dal numero di riga nella pagina. Ad esempio, l'occorrenza cui corrisponde il contesto n. 15 si trova nella riga 17 della pagina 141. Dopo il numero di pagina può comparire un ulteriore dato numerico, indicante il numero di colonna per un documento tipograficamente ripartito appunto in colonne. Una scrittura come **310.2.16** indicherebbe la riga 16 della seconda colonna di pagina 310. Il dato può essere preceduto dal numero di volume, riportato in celeste: ad esempio **3.34.12** starebbe ad indicare la riga 12 della pagina 34 del terzo volume. Se il numero di pagina è preceduto dalla lettera **r** significa che la numerazione delle pagine utilizza i numeri romani. Se il numero di pagina è seguito da una **r** o da una **v**, queste stanno rispettivamente ad indicare **recto** e **verso**; ad esempio **30r.1** indicherebbe pagina 30 recto, riga 1 e **42v.3** indicherebbe pagina 42 verso, riga 3;
6. **lemma associato**. Cliccando sul riferimento topografico, se l'occorrenza in questione è stata lemmatizzata si apre una finestra contenente i dati del lemma e dell'eventuale iperlemma utilizzati.

Finestra a contesti multipli, visualizzazione kwic con riferimento organico

Le informazioni riportate hanno una struttura simile a quella del caso precedente.

1. **marcatore di contesto selezionato**, in blu scuro. Se presente, indica che il contesto è stato selezionato per l'esportazione e la lettera di cui è costituito indica la geometria fissata per questa operazione, in base allo schema riportato nella tabella precedente.
2. **numero di contesto**, in verde (in rosso se il contesto è stato selezionato). Attribuisce al contesto un numero per identificarlo nell'insieme dei contesti localizzati;
3. **simbolo *** (**presenza di brano associato**), in nero. Cliccandolo si apre una finestra contenente una nota che riguarda tra le altre la parola centrale del contesto, oppure un brano tratto dal testo associato contenente una traduzione o una diversa edizione; in questo caso il brano mostrato è quello che corrisponde nel testo base, al brano comprendente la parola centrale del contesto. La finestra può anche contenere sia una nota che un brano associato. Per distinguerli, tenere presente che il brano associato è sempre preceduto da una riga contenente la sequenza "====". Se il brano associato consiste in una nota ma, per qualche motivo (non escluso un errore di preparazione dei testi) la nota stessa risultasse vuota, cioè priva di un testo, al suo posto apparirà la dicitura "NOTA VUOTA"; se ad un testo è associata una traduzione o una diversa edizione il simbolo **A** viene automaticamente inserito in tutti i suoi contesti, inclusi quelli per i quali non sia effettivamente previsto un brano corrispondente nel testo associato: in questo caso cliccando su **A** si otterrebbe in risposta la dicitura

"BRANO PRIVO DI TRADUZIONE". Chiudere la finestra con <Esc> o cliccando sull'icona di chiusura **X** in alto a destra;

4. **sigla del testo**, in celeste; cliccandola si apre la finestra contenente i dati bibliografici e statistici del testo da cui è tratto il contesto. Chiudere la finestra con <Esc> o cliccando sull'icona di chiusura **X** in alto a destra;
5. **riferimento organico**, in viola (occorrenza non lemmatizzata) o ciclamino (occorrenza lemmatizzata). È il riferimento organico corrente, al quale è associata la parte di testo da cui il contesto è stato estratto;
6. **numero di verso**, in verde. Questa dicitura compare ovviamente solo per le opere in versi. Ad esempio, nel contesto 58 dopo il riferimento organico **28** si trova l'indicazione **v13**, indicante appunto che la parola localizzata si trova nel **verso 13** del riferimento organico **28**. In questo caso il numero di verso è stato preceduto dalla lettera v, ad indicare che si il conteggio è stato fatto per versi singoli. Il quadro completo delle abbreviazioni che possono comparire per indicare il conteggio del verso è riportato nella tabella che segue:

dicitura	significato
v	verso
d	distico
t	terzina
q	quartina
o	ottava
s	stanza

7. **lemma associato**. Se l'occorrenza è stata lemmatizzata, e la cosa è segnalata dal colore più chiaro (ciclamino) del riferimento organico, cliccare sul riferimento organico stesso o se presente sul numero di verso per aprire la finestra contenente i dati del lemma e dell'eventuale iperlemma utilizzati. Chiudere la finestra con <Esc> o cliccando sull'icona di chiusura **X** in alto a destra.

Finestra a contesto singolo

In questo caso la rappresentazione a video riproduce esattamente l'aspetto tipografico del testo. La testata è strutturata su due righe.

Prima riga:

1. **simbolo # (indicatore di presenza di nota editoriale)**, in blu scuro. Avverte che tra i dati bibliografici del testo dal quale il contesto è stato estratto è presente una nota editoriale; cliccandolo si apre la finestra contenente i dati bibliografici e statistici dell'opera;
2. **titolo abbreviato**, in azzurro. Contiene l'inizio del campo **abbrev. titolo** incluso tra i dati bibliografici associati al testo; cliccandolo si apre la finestra contenente i dati bibliografici e statistici dell'opera;
3. **indice di qualità**, in marrone. Si tratta di un indicatore della qualità del testo, nel senso di attendibilità e pertinenza linguistica, congruità cronologica, ecc. Indica in sostanza un testo significativo per la documentazione della specifica varietà linguistica. Per i testi OVI valgono le seguenti codifiche: TS = testi significativi per la documentazione della specifica varietà linguistica; TPC = testi da citare solo in casi particolari per ragioni di cronologia; TPL = testi da citare solo in casi particolari per ragioni linguistiche;

Seconda riga:

4. **marcatore di contesto selezionato**, in blu scuro. Se presente, indica che il contesto è stato selezionato per l'esportazione e la lettera di cui è costituito indica la geometria fissata per questa operazione, in base allo schema riportato nella tabella precedente.
5. **numero di contesto**, in verde (in rosso se il contesto è stato selezionato per l'esportazione). Attribuisce al contesto un numero identificativo nell'insieme dei contesti localizzati; la numerazione serve, ad esempio, per identificare i contesti che si vogliono stampare o copiare su file;
6. **parte iniziale del riferimento organico**, in viola. È il riferimento organico corrente, al quale è associata la parte di testo da cui il contesto è stato estratto;

7. **numero di verso**, in verde. Questa dicitura compare ovviamente solo per le opere in versi. Il quadro completo delle abbreviazioni che possono comparire per indicare il conteggio del verso è riportato nella tabella che segue:

dicitura	significato
v.	verso
d.	distico
t.	terzina
q.	quartina
o.	ottava
s.	stanza

8. **riferimento organico completo**. È una finestra che permette di leggere, sotto il titolo *riferimento organico completo*, tutto il riferimento organico nei casi in cui lo spazio nella riga di testata non sia sufficiente. Per aprirla cliccare in un punto qualunque della parte di riferimento organico riportata (in viola) o del numero di verso se presente; per chiuderla usare <Esc> o l'icona di chiusura **X** in alto a destra;
9. **riferimento topografico**, in blu (numero di volume in celeste). È costituito dal numero di pagina seguito dal numero di riga nella pagina. Ad esempio, l'occorrenza cui corrisponde il contesto n. 15 si trova nella riga 17 della pagina 141. Dopo il numero di pagina può comparire un ulteriore dato numerico, indicante il numero di colonna per un documento tipograficamente ripartito appunto in colonne. Una scrittura come **310.2.16** indicherebbe la riga 16 della seconda colonna di pagina 310. Il dato può essere preceduto dal numero di volume, riportato in celeste: ad esempio **3.34.12** starebbe ad indicare la riga 12 della pagina 34 del terzo volume. Se il numero di pagina è preceduto dalla lettera **r** significa che la numerazione delle pagine utilizza i numeri romani. Se il numero di pagina è seguito da una **r** o da una **v**, queste stanno rispettivamente ad indicare **recto** e **verso**; ad esempio **30r.1** indicherebbe pagina 30 recto, riga 1 e **42v.3** indicherebbe pagina 42 verso, riga 3. Nei casi in cui il riferimento topografico non viene visualizzato (funzione impostata come disabilitata o non obbligatoria all'atto della creazione del corpus) al suo posto compare la dicitura **TXT**;
10. **testo associato al riferimento organico**. Cliccando sul riferimento topografico si apre una finestra che permette di prendere visione dell'intera parte di testo associata al riferimento organico dalla quale il contesto è stato estratto. La finestra include una barra di scorrimento per muoversi nel testo; nei casi in cui la parte di testo mostrata è particolarmente estesa vengono automaticamente abilitati i bottoni (**Avanti**) e (**Indietro**) per scorrere il testo a blocchi, muovendosi poi con la barra di scorrimento all'interno del blocco corrente. Nella stessa finestra è presente il bottone (**Dati bib.**) che permette di aprire anche da qui la finestra contenente i dati bibliografici e statistici del testo. Per chiudere la finestra usare <Esc> o l'icona di chiusura **X** in alto a destra;
11. **lemma associato** (se l'occorrenza è stata lemmatizzata). È rappresentato dall'**entrata lessicale** in verde, dalla **categoria grammaticale** in kaki e dal simbolo **I** in nero se è stato usato anche un iperlemma; i dati completi di lemma e iperlemma si possono ottenere cliccando in un punto qualunque dei campi precedenti per aprire la finestra che li visualizza. Per chiuderla usare <Esc> o l'icona di chiusura **X** in alto a destra.

VISUALIZZAZIONE DEI BRANI ASSOCIATI NELLA FINESTRA A CONTESTO SINGOLO

Se si passa alla finestra a contesto singolo per visualizzare un contesto al quale è associata una nota (simbolo **A** nella testata del contesto, nella finestra a contesti multipli) quest'ultima viene automaticamente presentata in un riquadro sottostante a quello principale; la visualizzazione della nota non dipende in alcun modo dalla geometria adottata per il contesto, in quanto la nota si intende sempre e comunque riferita alla parola centrale, sempre presente nel contesto considerato. Se per qualche motivo (magari per un errore nella preparazione dei testi) la nota risulta vuota, cioè priva di testo, al suo posto compare la dicitura "NOTA VUOTA".

Nel caso in cui il testo dal quale il contesto è tratto includa nel corpus una traduzione o una diversa edizione tutti i contesti sono caratterizzati dalla presenza del simbolo **A**. In tal caso, passando alla finestra a

contesto singolo, la traduzione viene automaticamente mostrata nel riquadro che si apre al di sotto di quello principale, esattamente come nel caso di una nota. Se sono contemporaneamente presenti sia una nota che una traduzione, vengono visualizzate entrambe (prima la nota, poi il separatore "===", poi la traduzione). Può capitare che il testo tradotto non riporti una traduzione per lo specifico contesto considerato nella finestra: in questo caso nel riquadro riservato alla traduzione compare la dicitura "BRANO PRIVO DI TRADUZIONE".

Analogamente alla nota anche la traduzione non dipende in genere dalla geometria scelta per il contesto principale; questo è vero finché si utilizzano gli *option box* **a parole, a periodi** o **specifico**, posti nel riquadro **tipo di contesto** in basso a sinistra. Se il brano prevede una traduzione viene però abilitato automaticamente anche un quarto *option box*, **a brani assoc.**, il quale consente di visualizzare il contesto base *per brani*, ovvero secondo una visualizzazione molto simile a quella *a periodi*, salvo il fatto che il testo mostrato non sarà costituito da 1, 3, 5,.. periodi, ma da altrettanti *brani*, ovvero parti in cui il testo è stato segmentato e alle quali corrispondono brani paralleli nel testo associato (in pratica, ad un *brano* presente nel testo base corrisponde nel testo associato un *brano* che ne costituisce la traduzione); la ripartizione in brani può anche essere fatta in modo da coincidere con quella in periodi, ma si tratta di una scelta rimessa al preparatore dei testi.

Se si attiva la visualizzazione a **brani assoc.** gli effetti di allargamento o restringimento del contesto base operati tramite i pulsanti (**Allarga**) e (**Restringi**) si applicheranno in parallelo anche al testo associato, le cui dimensioni varieranno di conseguenza, così da assicurare in questo caso corrispondenza tra il contesto riportato nel riquadro superiore e la relativa traduzione contenuta in quello inferiore.

ORDINAMENTO DEI CONTESTI

Nella finestra a contesti multipli i contesti sono mostrati ordinati **per testi** (con *ordinamento normale*). Vengono cioè mostrate prima tutte le forme presenti nel primo testo, poi quelle incluse nel secondo e così via; all'interno di un testo le diverse forme sono ordinate alfabeticamente e di ciascuna vengono presentati i contesti in ordine di apparizione.

Tramite [CONTESTI.. + ORDINATI PER FORME + ORD. NORMALE] oppure [CONTESTI.. + ORDINATI PER FORME + ORD. INVERSO] è possibile modificare l'ordine di presentazione dei contesti: questi vengono allora ordinati prima alfabeticamente per forme, poi per testo e, all'interno dello stesso testo, in ordine di apparizione. Il secondo menu effettua l'ordinamento *inverso*, ordinando alfabeticamente le forme non sulla base delle loro lettere iniziali ma a partire da quelle finali, utilizzando cioè le terminazioni (per cui, ad esempio, **lodato** verrà dopo scritta poiché la terminazione **o** segue la **a**).

Con [CONTESTI.. + ORDINATI PER TESTI + ORD. NORMALE] oppure [CONTESTI.. + ORDINATI PER TESTI + ORD. INVERSO] che a quel punto sostituiscono i precedenti è possibile tornare all'ordinamento di partenza. Valgono le considerazioni già fatte circa l'ordinamento normale e quello inverso.

Con [CONTESTI.. + ORDINATI PER SUCCESSIONE] si ottiene un terzo tipo di ordinamento, da usare essenzialmente in combinazione con l'opzione Rimario in IMPOSTAZIONE RICERCA. I contesti verranno ordinati per testo e, a parità di testo, per successione nel testo, a prescindere dalla forma.

Infine con [CONTESTI.. + ORDINATI PER SINTAGMI] si ottiene un ordinamento volto ad evidenziare, raggruppandone le occorrenze, eventuali sintagmi cui appartenga la parola localizzata nel corso della ricerca (questa modalità di ordinamento non è presente nelle cooccorrenze). I contesti vengono presentati in ordine alfabetico di forma; a parità di forma vale l'ordinamento alfabetico delle parole circostanti; in caso di ulteriore parità l'ordine è per testo e infine per successione al suo interno. Questo tipo di ordinamento presenta una certa complessità di funzionamento: per questo motivo verrà trattato a parte nel prossimo paragrafo.

Nel caso delle concordanze lemmatizzate gli ordinamenti sono differenti: quello di partenza e al quale si può in seguito tornare mediante [CONTESTI.. + ORDINATI PER LEMMI + ORD. NORMALE] oppure [CONTESTI.. + ORDINATI PER LEMMI + ORD. INVERSO] prevedono un ordinamento per lemmi, a parità di lemmi per forme, a parità di forme per testi; quelli alternativi, ai quali si accede tramite [CONTESTI.. + ORDINATI PER TESTI + ORD. NORMALE] oppure [CONTESTI.. + ORDINATI PER TESTI + ORD. INVERSO], prevedono un ordinamento per testi, a parità di testi per lemmi e a parità di lemmi per forme. Valgono le solite considerazioni in merito agli ordinamenti normale e inverso.

Resta da spiegare come siano ordinati fra loro i testi.

L'ordinamento preimpostato dei testi è quello **cronologico**, basato sul confronto tra le *date codificate* associate. Questo criterio può essere modificato. A questo scopo partendo dalla finestra a contesti multipli

tornare alla finestra MODALITÀ RICERCA mediante [VAI A.. + MODALITÀ RICERCA] e cliccare su [ORDINAMENTO]; tramite il *box a tendina* selezionare per esempio il campo **Autore** e definire l'ordinamento crescente cliccando su (**A - Z**) (la direzione attuale di ordinamento, crescente o decrescente, è evidenziata dal fatto che le lettere nel *bottone* corrispondente sono maiuscole). Chiudere il riquadro con (**Chiudi**). Adesso andare a *ricerca per forme* e ripetere la ricerca di **perd***, proseguendo fino ad ottenere i contesti. Questa volta, ad esempio, il contesto n. 1 riporta come autore "=", i successivi sono di autore **Anonimo**, il 53° è il primo di **Giovanni Campulu** e così via, in ordine alfabetico crescente di autore (questi dati sono rilevabili dalle schede dei dati bibliografici dei contesti, visibili cliccando sulle testate dei contesti stessi nel campo *Titolo abbreviato* in celeste).

In ogni momento si può sapere qual'è l'ordinamento attuale cercando, tra i sottomenu di [CONTESTI..], quello contrassegnato dal segno ">" ed eventualmente, al livello più interno, quello marcato col segno di spunta.

ORDINAMENTO PER SINTAGMI

Scegliendo questo tipo di ordinamento mediante [CONTESTI.. + ORDINATI PER SINTAGMI] i contesti sono disposti non solo secondo l'ordine alfabetico della forma trovata, ma anche secondo l'ordine alfabetico di un certo numero di forme adiacenti.

L'utente deve definire un numero di forme (da 2 a 10, inclusa quella trovata) e la posizione che la forma trovata deve occupare dentro la stringa così definita. Il sistema ordina i contesti secondo la prima forma della stringa, in subordine secondo la successiva e così fino al termine della stringa definita.

Per esempio, se si definisce una stringa di **3** parole, con la forma trovata in **prima posizione**, la stringa sarà costituita da 3 parole:

1. *forma trovata*,
2. *forma successiva 1*,
3. *forma successiva 2*.

L'ordinamento verrà fatto secondo *forma trovata*, in subordine secondo *forma successiva 1*, in subordine secondo *forma successiva 2*.

Definendo invece una stringa di **3** parole, con la forma trovata in posizione **2**, la stringa sarà:

1. *forma precedente*,
2. *forma trovata*,
3. *forma successiva*.

L'ordinamento verrà fatto secondo *forma trovata*, in subordine secondo *forma precedente*, in subordine secondo *forma successiva*.

Vi sono due casi nei quali l'ordinamento può risultare alterato rispetto a quello sin qui descritto:

1. presenza di punteggiatura forte all'interno della stringa,
2. inserimento in campo formula di una o più parole le cui posizioni le portano ad essere incluse nella stringa.

Se una parola che per posizione nella frase rientrerebbe nel sintagma non è stata indicizzata e non è quindi cercabile, non viene considerata appartenente al sintagma stesso e quindi non viene evidenziata.

Presenza di punteggiatura forte all'interno della stringa.

Le parole incluse per posizione all'interno della stringa definita dall'utente, ma separate dalla parola trovata da un segno di punteggiatura forte, non vengono normalmente prese in considerazione ai fini dell'ordinamento: tali parole *occupano* le rispettive posizioni, ma non sono prese in considerazione e non appaiono mutuamente ordinate: peraltro in quella posizione precedono alfabeticamente le parole che invece appartengono regolarmente al testo indicizzato.

Ad esempio, supponiamo che la stringa sia stata definita di lunghezza pari a 4 parole, con la stessa forma trovata in posizione 2; supponiamo anche che la ricerca abbia localizzato 30 occorrenze di tale forma e che in 10 dei 30 contesti la forma trovata sia preceduta immediatamente da una punteggiatura forte. L'ordinamento, come detto, parte dalla prima parola della stringa ovvero, in questo caso, dalla parola (posiz. 1) precedente a quella trovata (posiz. 2). Per primi verranno presentati i 10 contesti nei quali la parola in posizione 1 non viene presa in considerazione perché seguita da una interpunzione forte, poi gli altri 20. Questi ultimi verranno ordinati alfabeticamente secondo la parola in posizione 1, in subordine secondo la parola in posizione 3 (quella successiva a quella centrale), in subordine secondo quella in posizione 4. Invece

nei 10 contesti iniziali viene trascurata la posizione 1 e quindi essi saranno ordinati solo secondo le parole in posizione 3 e 4.

Quanto detto vale anche se la punteggiatura forte si trova in posizione successiva alla forma trovata, sempre all'interno della stringa.

Volendo ignorare questo vincolo ed estendere comunque l'ordinamento a tutta la stringa, a prescindere dalla presenza di punteggiatura forte, basta disattivare il *check box punteggiatura forte interrompe* presente nella finestra ottenuta tramite [CONTESTI.. + ORDINATI PER SINTAGMI].

Appartenenza ad un campo formula di una o più parole interne alla stringa

Il trattamento di queste parole è analogo a quello utilizzato in presenza di punteggiatura forte.

Le parole che si trovino all'interno della zona considerata utile per il sintagma ma che siano inserite in un campo formula *occupano* le rispettive posizioni, ma non sono prese in considerazione e non appaiono mutuamente ordinate: peraltro in quella posizione precedono alfabeticamente le corrispondenti parole presenti negli altri contesti, le quali appartengono invece regolarmente al testo indicizzato.

Fare attenzione alle particolarità descritte nei precedenti sottoparagrafi, le quali possono ingenerare la falsa impressione che l'ordinamento per sintagmi non stia funzionando correttamente, mentre in realtà il programma si comporta come se le parole poste prima di una punteggiatura forte che preceda la parola centrale (e quelle successive ad una che la segua), così come quelle incluse in un campo formula, non esistano nel testo, pur occupando le relative posizioni.

È infine abbastanza evidente che il modo migliore per utilizzare questo tipo di ordinamento si ottiene visualizzando i contesti in modalità **kwic** anziché standard.

Per agevolare il riconoscimento dei sintagmi localizzati dal programma, le parole che li costituiscono (esclusa la forma trovata che è formattata in rosso) sono formattate in blu.

INDICATORI A BARRETTA ORIZZONTALE

Mantenendo l'ultimo ordinamento impostato, ovvero quello per autore, si può vedere che nei contesti n. 1, 2, 4, 47, 53 ecc. subito sopra il titolo abbreviato compare una **barretta nera orizzontale**: la sua presenza sta ad indicare che il testo è diverso da quello relativo al contesto immediatamente precedente. Si tratta di un ausilio visivo inserito per agevolare l'individuazione del punto di stacco fra gruppi di contesti relativi a testi diversi; infatti poiché del titolo abbreviato viene mostrata solo la parte iniziale si potrebbe essere tratti in inganno qualora si susseguissero contesti estratti da testi differenti, caratterizzati però da titoli abbreviati uguali nella parte iniziale, come può capitare ad esempio nel caso di volumi diversi di una stessa opera. Questo indicatore non compare nelle concordanze lemmatizzate ordinate per lemmi, nell'ordinamento per sintagmi e nel raffinamento.

Con criterio e finalità analoghe una **identica barretta nera** compare sopra il riferimento topografico ad indicare che la forma localizzata è diversa da quella riportata nel contesto precedente. Nel caso di concordanze lemmatizzate la stessa barretta indica invece il primo contesto lemmatizzato con un altro lemma. Questo indicatore non compare nelle occorrenze ordinate per successione o per sintagmi, nelle cooccorrenze e nel raffinamento.

Il primo contesto mostrato riporta sempre entrambe le barrette, quando previste.

LEZIONE 30. RAFFINAMENTO

La possibilità di intervenire a più riprese sui risultati che si ottengono nel corso di una ricerca, facendo uso delle griglie presenti nelle finestre FORME LOCALIZZATE e ACCUMULATORE, permette di limitare il numero dei contesti ottenuti sulla base di criteri piuttosto selettivi, tutti comunque riferiti alla parola *centrale* di ogni contesto.

Un'impostazione del tutto diversa è invece quella che di cercare *cooccorrenze* di due o più parole poste in prossimità tra loro. Questo tipo di ricerca, prevista in GATTO e trattata nel capitolo 19, consente di modulare la richiesta utilizzando una serie di opzioni e di codifiche che la rendono piuttosto potente e versatile ma un po' meno immediata nell'uso rispetto alla ricerca per forme.

Può capitare che si vogliano cercare cooccorrenze di parole specifiche, senza però particolari esigenze circa il loro ordinamento, la posizione rispetto alla punteggiatura e via dicendo. Può anche succedere che una

ricerca venga avviata come una normale ricerca per forme e che solo successivamente, osservando i contesti ottenuti, si abbia interesse a ridurre questi ultimi, specie se presenti in numero elevato, sulla base di criteri di prossimità ad altre parole.

GATTO offre una funzione, il *raffinamento*, che agevola il lavoro qualora ci si trovi in una di queste ultime situazioni; il raffinamento può essere ripetuto più volte, per limitare i contesti sulla base di più parole vicine tra loro, così come permette di indicare contemporaneamente più forme alternative da ricercare in prossimità di quella centrale, definendo anche una distanza limite variabile. Se il risultato non soddisfa o si desidera ripetere il raffinamento sulla base di altre parole, si può tornare ai contesti individuati sulla base della sola parola centrale e da lì ripartire.

Le parole incluse nei contesti sono visualizzate in nero, con eccezione di quella centrale rossa: le parole localizzate in prossimità di quella centrale come risultato di un raffinamento sono evidenziate in nocciola.

L'ordinamento dei contesti presentati a seguito dell'uso di [RAFFINA] è quello che risulta essere stato impostato per ultimo prima di avviare il raffinamento; per modificarlo è necessario tornare ai contesti di partenza mediante [CONTESTI BASE].

Esempio 1: ricerca di cooccorrenze di **terra** con **mare**

Effettuando una normale ricerca per forme localizzare i contesti della forma **terra**: con [CONTESTI.. + NUM. CONT. TROVATI, SELEZ.] si vede che ne risultano **182**. A questo punto avviare il raffinamento con [RAFFINA], digitare **mare** nel riquadro bianco e lasciare la distanza massima impostata a **10**; procedere con (OK). Si ottiene l'unico contesto che soddisfa la condizione cercata.

Esempio 2: ricerca di cooccorrenze di **testamento** o **tesstamento** con **Baldovino**

Effettuare una ricerca per forme fino a localizzare le **33** occorrenze complessive di **testamento** e **tesstamento**, quindi con [RAFFINA] impostare **aldovino** seguito da (OK): si ottengono **25** contesti. Come si può notare, osservando ad esempio i contesti **21** e **22**, tutte le occorrenze di **aldovino** incluse tra le 10 parole precedenti e le 10 parole successive a **testamento** vengono evidenziate.

Esempio 3: ricerca di cooccorrenze di **sancta, sancte, sancti**, ecc. con **Maria** e con **Magdalena**

Effettuare la ricerca per forme di **sanct?**, individuando così le **1946** occorrenze presenti in **Demo**. Con [RAFFINA] cercare le cooccorrenze con **maria**: i contesti si riducono a **20**; utilizzando nuovamente [RAFFINA] e digitando **magdalena** si estraggono gli unici due contesti che, a una distanza massima di 10 parole da **sancta** (o **sancti**,...), includono anche **maria** e **magdalena**.

Esempio 4: ricerca di cooccorrenze di **sancta, sancte, sancti**, ecc. con **patre, patri**,... e con **monaco, monaci**,..., con riduzione della distanza massima ammessa

Effettuare la stessa ricerca per forme dell'esempio precedente (si ottengono **1946** occorrenze), poi effettuare il raffinamento con **patr?**: si ottengono **380** contesti; effettuare un ulteriore raffinamento con **monac?**: i contesti, centrati su **sancta** (e simili) e includenti sia **patre** (e simili) che **monaco** (e simili) scendono a **13**.

A questo punto si immagini di voler ripetere la stessa analisi, riducendo però a **6** parole la distanza massima ammessa.

Prima di tutto, tornare alle 1946 occorrenze iniziali con [CONTESTI BASE]. A questo punto ripetere i passaggi precedenti, avendo cura di digitare nel riquadro **distanza massima** il valore **6**: si ottengono, in successione, **349** e **6** contesti.

Esempio 5: ricerca di cooccorrenze di **sancta, sancte, sancti**, ecc. con **patre, patri**,... oppure con **monaco, monaci**,...

Gli esempi precedenti hanno localizzato i contesti che prevedevano la presenza contemporanea di **sanct?**, **patr?** e **monac?** (? sta per una lettera qualunque).

In questo esempio verranno invece cercate le cooccorrenze di **sanct?** con **patr?** e di **sanct?** con **monac?**, ovvero tutti i contesti centrati su **sanct?** nei quali compaiano anche **patr?** oppure **monac?**.

Riportarsi sulle solite 1946 occorrenze che si ottengono tramite la ricerca per forme di **sanct?**.

A questo punto attivare [RAFFINA] e digitare, nel riquadro bianco la stringa **<patr?,monac?>** (oppure **<patr,monac?>**); controllare che la distanza massima sia riportata a **10** e dare (OK). I contesti ottenuti questa volta sono **478**, ovvero molti di più dei 13 del caso precedente, che richiedevano la presenza simultanea di parole tratte da tutti i tre pacchetti (**sanct?**, **patr?**, **monac?**).

Questo esempio utilizza i caratteri jolly descritti negli esempi 6 e 7 del capitolo 14. Le parentesi angolari erano lì descritte come mezzo per inserire parti variabili all'interno delle forme cercate, come può essere il caso della stringa **pas<col,s>are**, che cerca entrambe le forme **pascolare** e **passare**: niente però impedisce di inserire tra parentesi angolari parole intere in luogo di parti di parole. Nell'esempio ora citato si sarebbe ottenuto lo stesso risultato digitando **<pascolare,passare>**.

Nel caso della finestra di avvio delle ricerche la presenza di 10 selettori disponibili per digitare forme diverse rende antieconomico inserire stringhe di questo tipo in un unico selettore, ma nel caso del raffinamento (un solo selettore presente) questo è l'unico modo per specificare più parole contemporaneamente.

Esempio 6: ricerca di cooccorrenze di **tanto** con tutte le occorrenze di **co**, a prescindere dai diacritici con cui questa seconda forma grafica compare.

Eseguire la ricerca per forme di **tanto**, fino ad ottenere i **132** contesti che la contengono.

Cliccare su [RAFFINA] e inserire, nel riquadro bianco la forma **co**; attivare il flag **attiva ricerca espansa** e procedere con (OK). Si ottengono **5** contesti, **1** con la forma **co**, **3** con **co** e **1** con **ço**. Notare che lo stesso risultato sarebbe stato ottenuto inserendo nel riquadro **co** o **ço** in luogo di **co**.

Per verifica, tornare ai 132 contesti di partenza con [CONTESTI BASE] e ripetere lo stesso raffinamento senza selezionare il flag **attiva ricerca espansa**: si otterrà come risultato il solo primo contesto dei 5 precedenti.

Esempio 7: ricerca delle occorrenze di **messere** con contemporanea presenza di **cavallo**, esclusi però i contesti che includono **guisa**.

Eseguire la ricerca per forme di **messere**, fino ad ottenere i **102** contesti che la contengono.

Cliccare su [RAFFINA], inserire nel riquadro bianco la forma **cavallo** e procedere con (OK). Si ottengono **3** contesti. Cliccare su [RAFFINA]. A questo punto attivare l'opzione **solo i contesti che non contengono**: (con ciò disattivando automaticamente l'opzione standard **solo i contesti che includono anche**:) e inserire nel riquadro bianco la forma **guisa**. Prima di procedere, notare che nel primo dei 3 contesti, alcune parole dopo **Messere**, compare la forma **guisa** non presente negli altri 2 contesti. Cliccare su (OK). Sopravvivono dopo la selezione solo il secondo e terzo contesto, unici a soddisfare contemporaneamente le 3 condizioni imposte (presenza di **messere**, presenza di **cavallo**, assenza di **guisa**).

Da quest'ultimo esempio si vede che l'opzione di *raffinamento* può essere rieseguita più volte, definendo ogni volta la *direzione* del raffinamento (ricerca oppure esclusione di cooccorrenze) indipendentemente dalla direzione seguita in precedenza.

LEZIONE 31. VISUALIZZAZIONE DEI CONTESTI A PARTIRE DALLA LISTA DEI TESTI CHE LI CONTENGONO

I contesti che si ottengono come risultato di una ricerca vengono presentati in ordine dei testi di appartenenza. Questo fatto fornisce immediatamente un modo per raggruppare i risultati sulla base della loro fonte.

Se la provenienza dei contesti deve delimitare strettamente i risultati che si cercano il modo migliore di procedere è quello di limitare la ricerca a un sottocorpus (vedere a questo riguardo il capitolo 20).

Esiste una terza soluzione, intermedia tra le precedenti, che permette di separare più agevolmente i contesti sulla base dei testi di provenienza, senza peraltro precisare a priori questi ultimi. Questa soluzione risulta poi utile nei casi in cui non si vogliono vedere in dettaglio i singoli contesti ma semplicemente conoscere la lista dei testi che li contengono.

Si supponga di voler trovare la lista dei testi che includono occorrenze della forma **casa**.

Impostare nel solito modo la ricerca per forme di **casa**. Nella finestra RICERCA PER FORME per avviare la ricerca anziché utilizzare i menu [AVVIO RICERCA] o [CONTESTI] fare uso di [LISTA TESTI]; il risultato immediato consiste in una griglia che contiene una lista di 4 testi con l'informazione relativa al numero di contesti della forma cercata presenti in ognuno di essi (in questo caso, nell'ordine, 1, 16, 1, 65). Il contenuto delle righe correntemente selezionate nella griglia può essere stampato con [STAMPA SEL.]. Per vedere i contesti di uno o più di tali testi, selezionare nel solito modo, ovvero cliccando sulle righe corrispondenti, i testi interessati. Ad esempio, volendo vedere i due contesti inclusi nel primo e nel terzo testo, selezionare le corrispondenti righe della griglia (ricordarsi di tenere premuto il tasto <Ctrl> per la selezione multipla) e

procedere con [MOSTRA CONTESTI]. Per vedere ora i 65 contesti tratti dal quarto testo tornare alla lista dei testi con [LISTA TESTI], selezionare l'ultima riga e ripetere la procedura precedente.

Per selezionare o deselegionare simultaneamente tutti i testi della lista usare [SEL/DESEL.. + SELEZIONA TUTTO] o [SEL/DESEL.. + DESELEZIONA TUTTO].

Il contenuto della griglia può essere stampato o scritto su file mediante [STAMPA SEL.]. Nel secondo caso la colonna 'sigla' viene stampata con un formato speciale e in testo nascosto (formato speciale per usi redazionali del TLIO).

La lista dei testi è attivabile non solo dalla finestra RICERCA PER FORME ma anche dall'ACCUMULATORE. Per verificarlo, tornare a RICERCA PER FORME mediante [VAI A.. + IMPOSTAZIONE RICERCA] e ripetere la ricerca della forma **casa**, questa volta usando [AVVIO RICERCA]; selezionare la riga di **casa** e continuare con [COPIA IN ACC.]; selezionare di nuovo la riga di **casa** e procedere ulteriormente con [LISTA TESTI].

L'ordinamento dei contesti presentati a seguito dell'uso di [LISTA TESTI] è quello che risulta essere stato impostato per ultimo prima di passare a visualizzare i contesti; per modificarlo è necessario tornare alla lista dei testi [LISTA TESTI].

Si può tornare alla finestra contenente la lista dei testi, a partire dalle altre finestre di ricerca, tramite [VAI A.. + LISTA TESTI]; a questo riguardo va precisato che i menu [VAI A.. + CONTESTI ESTRATTI] e [VAI A.. + LISTA TESTI] presenti nelle varie finestre sono abilitati alternativamente, in funzione dell'ultima visualizzazione effettuata.

LEZIONE 32. COME ESPORTARE I CONTESTI TROVATI

I contesti ottenuti come risultati delle ricerche possono essere esportati al fine di utilizzarli più agevolmente. Sono previste 4 modalità di esportazione dei risultati:

1. stampa su carta,
2. scrittura su file RTF,
3. scrittura su file RTF nel formato usato per la redazione del TLIO,
4. estrazione di un grafico di distribuzione temporale dei contesti.

Prima di passare alla illustrazione di come procedere è opportuno notare alcuni punti.

Mentre la quarta modalità fornisce un'informazione correlata globalmente a tutti i contesti ottenuti, le prime 3 si applicano ai soli contesti *selezionati*; in altre parole tra l'ottenimento dei contesti e la loro esportazione è previsto un passaggio intermedio nel quale l'utente *seleziona* i contesti per i quali si deve procedere all'esportazione, indipendentemente dalle modalità di quest'ultima.

Un secondo punto da tenere presente è che l'utente è libero di definire non solo *quali* contesti esportare ma anche *quale geometria assegnare* loro, singolarmente o per gruppi, in funzione dell'esportazione.

L'ultimo concetto associato a queste operazioni è quello di *congelamento* del contesto: nel momento che un contesto viene selezionato per l'esportazione ne viene congelata la struttura; successivamente è possibile visualizzarlo con diversa forma ed estensione ma, a meno di non procedere a una sua deselegione e rilesione, il contesto verrà esportato con la struttura che aveva al momento della selezione.

Per seguire gli esempi che seguono, impostare ed eseguire la ricerca per forme di **perd***: si otterranno 80 contesti.

Esempio 1: selezione e stampa di singoli contesti dalla finestra RISULTATI RICERCA (a contesti multipli)

Trovandosi nella finestra RISULTATI RICERCA (a contesti multipli) per selezionare un contesto per l'esportazione è sufficiente cliccare in corrispondenza (o a sinistra) del suo numero d'ordine: quest'ultimo diverrà di colore rosso e alla sua sinistra apparirà la lettera **N**, ad indicare che il contesto è stato selezionato per l'esportazione in modalità *a parole*. Selezionare in questo modo il contesto n. 1. Cliccando di nuovo il contesto verrà alternativamente deselegionato e rilesionato.

Tramite [CONTESTI.. + TIPO + KWIC (RIF. TOP.)] portarsi nella modalità di visualizzazione kwic con riferimento topografico e selezionare alla stessa maniera il contesto n. 9; mentre il n. 1 rimane caratterizzato dalla lettera **N**, poiché la sua configurazione per l'esportazione è stata congelata a quella formata da 31 parole, il contesto n. 9 viene selezionato nella modalità *kwic più riferimento topografico*, indicata dalla lettera **K**. Portarsi ora alla modalità kwic + riferimento organico con [CONTESTI.. + TIPO + KWIC (RIF. ORG.)] e selezionare il contesto n. 7; questa volta il simbolo è la lettera **W**. Tornare alla modalità standard con [CONTESTI.. + TIPO + STANDARD] e selezionare ancora i contesti n. 3 e 4.

Per sapere quanti sono i contesti complessivamente selezionati al momento cliccare su [CONTESTI.. + NUM. CONT. TROVATI, SELEZ.]: come prevedibile la risposta è che esistono **80** contesti localizzati, dei quali **5** selezionati per l'esportazione.

Per procedere alla stampa dei contesti così selezionati, dalla finestra RISULTATI RICERCA cliccare su [STAMPA]; selezionare **stampante** come **Tipo di uscita** e cliccare su **(OK)**.

Analizzando la stampa ottenuta si vede che la prima riga riporta sigla e titolo abbreviato del testo; questo tipo di dato non compare più nel seguito in quanto tutti i contesti stampati appartengono allo stesso testo. I 5 contesti sono stampati nell'ordine con cui sono stati visualizzati e rispettando per ciascuno la geometria che aveva al momento della selezione. La nuova numerazione dei contesti assegna semplicemente a ciascuno un numero progressivo di esportazione, non correlato col numero d'ordine che il contesto aveva in precedenza. Il penultimo contesto è in modalità kwic con riferimento organico e infatti è caratterizzato dall'appartenere al **verso 48 del riferimento organico 4**, mentre l'ultimo (kwic con riferimento topografico) appartiene alla **riga 6 di pagina 78**.

Con [AZZERA SEL.] rimuovere tutte le 5 selezioni di contesti per l'esportazione.

Esempio 2: selezione, stampa e scrittura su file di singoli contesti dalla finestra RISULTATI RICERCA (a contesto singolo)

Cliccare in un punto qualunque del testo incluso nel contesto n. 5, così da portarlo nella finestra a contesto singolo. Tramite l'*option box a parole* impostare il tipo di contesto a parole, poi ridefinirne la lunghezza a **17** parole modificando il contenuto del riquadro **lung. contesto** e applicare la modifica con **(Applica)**; il contesto risultante sarà **ebbe tanta aversitate: chi fu soffrent' e no perdeo memoria per grave pene c' a lui fosser**. A questo punto selezionarlo cliccando a sinistra del numero d'ordine 5, che diventerà di colore rosso, mentre alla sua sinistra comparirà il simbolo **N** (contesto strutturato *a parole*). Tornare alla finestra a contesti multipli con **(Esci)**. Riaprire la finestra a contesto singolo sul solito contesto n. 5: la sua configurazione rimane quella congelata con la selezione. Ora attivare la modalità **a periodi** con l'apposito *option box* ed espandere il contesto fino a 5 periodi con **(Allarga)**; la sua dimensione cambia ma il simbolo a sinistra del numero d'ordine rimane **N**; non deselegionare il contesto e tornare invece alla finestra a contesti multipli con **(Esci)**.

Ora cliccare sul testo del contesto n. 6 per portare questo nella finestra a contesto singolo, espanderne le dimensioni a **3** periodi e selezionarlo per l'esportazione: a sinistra del numero d'ordine compare il simbolo **P** (contesto *a periodi*); il contesto così selezionato è **Vorria c'...perderia...fora pietosa**. Tornare alla finestra a contesti multipli.

Cliccare sul testo del n. 8 e, nella finestra a contesto singolo, espanderlo a **3** periodi poi cliccare sul primo dei due "**per**" (una riga sopra a **perdo**) e su "**data**" (3 righe sotto), così da definire il contesto *specifico*: **[...] per paura...perdo...m' à data..**. Selezionarlo e tornare alla finestra a contesti multipli.

Attivare la stampa con l'omonimo menu e attivare come **Tipo di uscita** sia **stampante** che **file**; lasciare impostato **file nuovo** e dare **(OK)**; indicare come nome di file **prova3** (solo il nome, poiché l'estensione **RTF** viene assegnata automaticamente) e avviare con **(Salva)**. Verificare che il foglio stampato contenga effettivamente i 3 contesti con le geometrie volute, poi fare lo stesso controllo sul nuovo file **prova3.rtf**, utilizzando un programma in grado di leggere file RTF (**Rich Text File**), come ad esempio Word.

Con [AZZERA SEL.] rimuovere tutte le selezioni di contesti per l'esportazione.

Esempio 3: selezione, stampa e scrittura su file in formato di redazione del TLIO di contesti dalla finestra SELEZIONE/DESELEZIONE RISULTATI

Gli esempi precedenti hanno esemplificato la selezione di singoli contesti. È possibile usare uno strumento più efficiente dovendo selezionare un numero elevato di contesti, ovvero la finestra SELEZIONE/DESELEZIONE RISULTATI.

Partendo dalla finestra RISULTATI RICERCA (a contesti multipli) cliccare su **[Selezione]**. Nella nuova finestra sono presenti 4 riquadri:

- quello in alto a sinistra definisce l'operazione da eseguire,
- quello in basso a sinistra riporta il numero di contesti al momento selezionati per l'esportazione,
- quello in alto a destra permette di scegliere i contesti da selezionare o deselegionare,
- quello in basso a destra permette di definire la geometria da attribuire ai contesti da esportare.

Nel riquadro in alto a destra cliccare su **Tutti**: come risultato tutti gli 80 contesti vengono selezionati, con dimensione 31 parole, come indicato nel riquadro in basso a destra; per rendere operativa la selezione impostata cliccare su **(OK e chiudi)**.

Annullare la selezione con [**Azzera sel.**], poi riaprire la finestra SELEZIONE/DESELEZIONE RISULTATI. Questa volta cliccare su **contesti n.** e digitare nel riquadro a destra la stringa **-2,4,6,8-10** poi cliccare su **a periodi** e nel corrispondente riquadro a destra sostituire **1** con **3**; cliccare su **(OK)**: in questo modo è stato indicato di selezionare i primi 2 contesti, il quarto, il sesto e quelli compresi tra l'ottavo e il decimo e di eseguire la selezione senza però chiudere la finestra; tutti i contesti sono stati selezionati con dimensione pari a 3 periodi. Il riquadro in basso a sinistra riporta il numero di contesti selezionati al momento, ovvero **7**. Ora cliccare di nuovo su **contesti n.** e questa volta digitare **3,12**. Nel riquadro in basso a destra cliccare su **a parole** e cambiare **31** in **11**; infine cliccare su **(OK e chiudi)**: i contesti selezionati sono ora passati a **9** come si può vedere con [**CONTESTI.. + NUM. CONT. TROVATI, SELEZ.**], di cui **7** configurati a 3 periodi e **2** a 11 parole. Cliccare su [**STAMPA**], attivare **stampante** e **file redaz. TLIO**, lasciare **file nuovo** e cliccare su **(OK)**. Assegnare come nome di file **prova4** e avviare con **(Salva)**. La stampante scrive una pagina contenente i contesti indicati, con le geometrie assegnate.

Aprire il file **prova4.rtf**: i contesti hanno la geometria voluta ma il formato di stampa è notevolmente diverso da quello riportato nel file **prova3.rtf**. Si tratta di un formato speciale particolarmente adatto a trasferire i contesti così ottenuti in voci di vocabolario. Questo formato non è disponibile per i contesti in formato kwic.

Esempio 4: uscita grafica

Tornare alla finestra RISULTATI RICERCA (a contesti multipli) e cliccare su [**GRAFICA**]. Verrà visualizzato un istogramma con la distribuzione temporale degli 80 contesti localizzati: ad ogni contesto è associato l'anno (o l'anno finale) del dato cronologico associato al testo da cui il contesto è tratto; l'intervallo temporale complessivo entro cui ricadono i testi chiamati in causa viene ripartito in periodi di uguale estensione e per ognuno di essi viene riportato il numero di contesti che vi risultano inclusi.

La finestra USCITE GRAFICHE permette alcune manipolazioni:

- cliccare su [**SCRITTE.. + DICITURA ASSE Y**], digitare **Occorrenze trovate** e **<Invio>**: la dicitura verticale viene sostituita da quella digitata. Lo stesso può essere fatto col titolo e con la dicitura orizzontale;
- l'istogramma è ripartito inizialmente in 10 intervalli temporali. Cliccare su [**NUMERO PUNTI**], digitare **5** e **<Invio>**. L'istogramma viene ridisegnato di conseguenza;
- per stampare il grafico utilizzare [**COPIA.. + SU STAMPANTE**];
- per trasferirlo in un documento di Word, un immagine di Paint o altro, usare [**COPIA.. + NEGLI APPUNTI**] e poi recuperarlo con gli specifici comandi di inserimento immagini (ad esempio, in Word usare [**MODIFICA.. + INCOLLA SPECIALE..**] seguito da **Bitmap indipendente dalla periferica** e **OK**; in Paint usare [**MODIFICA.. + INCOLLA**], ecc.).

ULTERIORI ELEMENTI PRESENTI NELLA FINESTRA SELEZIONE/DESELEZIONE RISULTATI.

deseleziona

Utilizzare questo *option box* per *deselezionare* anziché selezionare i contesti indicati nella finestra.

aggiorna già selezionati

Supponiamo che prima di accedere a questa finestra il solo contesto n. 3 fosse stato selezionato, con lunghezza 3 periodi: se a questo punto si selezionano **tutti** i contesti con lunghezza **31 parole** si otterrà il nuovo selezionamento per tutti gli altri contesti, mentre il numero 3 rimarrà selezionato a 3 periodi. Esiste infatti una protezione per cui la finestra SELEZIONE/DESELEZIONE RISULTATI per default non seleziona di nuovo con diversa geometria i contesti che risultano già selezionati. L'attivazione del *checkbox* **aggiorna già selezionati** scavalca questa protezione e forza l'assegnazione della nuova estensione a tutti i contesti specificati.

solo già selezionati

Usata in combinazione con l'opzione precedente permette di rifelezionare con diversa geometria tutti e soli i contesti che risultano già selezionati.

a brani, kwic (rif. top.), kwic (rif. org.)

Sono evidentemente altre geometrie attribuibili ai contesti. In particolare per quanto concerne **a brani** vedere la Lezione 34.

Per uscire dalla finestra senza operare selezioni usare (**Esci**).

OPZIONI INCLUSE NELLA FINESTRA DI STAMPA

La finestra include alcuni elementi opzionali, dipendenti peraltro dal tipo di esportazione scelto, che verranno ora descritti (vedere anche la Lezione 34 per ulteriori elementi sull'esportazione dei brani associati).

Opzione 'stampa note/trad'

Se attivata, include nella stampa anche eventuali note e traduzioni associate al contesto stampato. Questa opzione non è utilizzabile per la scrittura su file in formato di redazione del TLIO. Anche nel caso di contesti selezionati in formato kwic la stampa di note e traduzioni non ha luogo, a prescindere dall'attivazione di questa opzione.

Per default la lunghezza del testo associato contenente una traduzione è di 1 brano; questo valore può essere modificato in fase di selezione del contesto per la stampa.

Opzione 'stampa lemmi/iperlemmi'

Se attivata, inserisce la stampa del lemma utilizzato e dei suoi attributi nel caso di occorrenze lemmatizzate; se è presente un iperlemma direttamente associato all'occorrenza stampa anche questo. Nel caso di contesti selezionati in formato kwic la stampa dei lemmi non ha luogo, a prescindere dall'attivazione di questa opzione.

LEZIONE 33. RICERCA DELLE FORME INCLUSE IN UN BRANO DI TESTO. RICERCA DELLE FORME MAI LEMMATIZZATE

RICERCA DI FORME A PARTIRE DA UN BRANO DI TESTO

Le ricerche mostrate finora sono sempre partite dalla digitazione di una o più forme all'interno di appositi *selettori*.

Esiste una modalità alternativa per specificare le forme da cercare, la quale parte dall'inserimento di un brano di testo in un'apposita finestra di GATTO tramite tecniche di *copia e incolla*. A quel punto diviene possibile localizzare all'interno del corpus le eventuali occorrenze delle parole incluse nel brano.

Si avvii GATTO, si entri nell'ambiente **Ricerche**, si apra il solito corpus **Demo** e si avvii la finestra RICERCA PER FORME. Ora, anziché digitare delle forme all'interno dei selettori, si clicchi su **[INPUT BRANO]**. A questo punto si prelevi un brano testuale da un documento qualunque, copiandolo e incollandolo nel grande riquadro **forme da cercare**.

Ad esempio, per prelevare un brano da un documento Word, selezionare al suo interno il testo da prelevare e cliccare su **[MODIFICA + COPIA]** di Word, poi trasferirsi su GATTO, cliccare col tasto destro in un punto del riquadro **forme da cercare** e scegliere **INCOLLA**.

Come prova si può utilizzare questo stesso manuale: si prelevi la parte iniziale della precedente Lezione 32, da **I contesti ottenuti...** a **...notare alcuni punti**. Questa parte include anche numeri e elementi di formattazione. Ora si avvii la ricerca con **[AVVIO RICERCA]**: si otterranno 20 forme presenti nel brano scelto le quali compaiono anche all'interno del corpus **Demo**. La ricerca può proseguire fino a trovarne i contesti con le normali tecniche viste in precedenza. Come si vede GATTO scarta gli elementi che non sono cercabili, come i numeri, senza peraltro segnalarne la presenza come *errata*.

Se la ricerca parte da un brano copiato da un documento, è molto probabile che al suo interno si trovi un certo numero di parole grammaticali, come articoli e preposizioni, la cui elevata frequenza finirebbe per rendere difficile isolare le relativamente limitate occorrenze delle forme di interesse: per questo motivo si sconsiglia l'uso dei menu **[CONTESTI]** e **[LISTA TESTI]**, sia pur presenti, in favore di **[AVVIO RICERCA]** che consente di sfrondare la ricerca dagli elementi non desiderati.

La *ricerca a partire da un brano di testo* dispone di un numero limitato di opzioni, solo in parte coincidenti con quelle incluse nella normale ricerca basata sui selettori.

Opzione 'mostra lemmi'

Questa opzione ha lo stesso significato che nella ricerca a partire dai selettori. Tornare alla finestra RICERCA PER FORME e ripetere l'ultima ricerca attivando il *checkbox mostra lemmi*: questa volta si otterranno 62 elementi, ovvero le stesse 20 forme di prima, ma con i contesti ripartiti a seconda della lemmatizzazione (ad esempio le 127 occorrenze della forma **come** risultano ora suddivise in 111 occorrenze non lemmatizzate, 9 lemmatizzate con **come (cong.)**, 5 con **come (avv.)**, 1 con **sì come (cong.)** e 1 con **come (s.m.)**). Si vede che i lemmi associati a **come** si susseguono in ordine casuale, non alfabetico: questo è consistente con l'indicazione **ordinati per sole forme** riportata a destra. Per ottenere un ordinamento completo dei risultati ovvero per forme, a parità di forme per lemmi, a parità di lemmi per categorie grammaticali e infine per disambiguatori, utilizzare il bottone **(A-Z)** in combinazione con l'indicazione **ordina per forme**: fatto questo le occorrenze non lemmatizzate di **come** prederanno quelle lemmatizzate; tra queste verranno prima quelle lemmatizzate col lemma **come** in ordine alfabetico di categoria grammaticale e poi quelle lemmatizzate con **sì come**.

Opzione 'muti'

Se nel corpus è stato precedentemente importato il lemmario di un altro corpus, l'operazione ha probabilmente introdotto negli archivi del corpus ricevente delle coppie forma-lemma che in precedenza non vi comparivano; data una di queste coppie, nel corpus non esistono occorrenze della forma che vi compare, lemmatizzate col lemma incluso in quella coppia. Per questo motivo tali coppie sono dette *mute*. Un lemma collegato alle forme solo tramite *coppie mute* è chiamato a sua volta *lemma muto*. Se nel corpus sono presenti coppie forma-lemma mute, diviene visibile l'opzione **muti** che, se attivata, estende la ricerca anche a coppie e lemmi muti. Una ricerca basata sugli elementi *muti* ha senso se, nell'ACCUMULATORE, si fa poi uso delle opzioni **non lemmatizzate e/o escl. lemmi selez.**

Opzione 'ricerca espansa'

Tornare alla finestra RICERCA. PER FORME, disattivare **mostra lemmi** e attivare **ricerca espansa**.

Questa opzione ha lo stesso significato che aveva nella ricerca per selettori, con la differenza che si applica a tutte le forme contemporaneamente.

Avviare la ricerca: si otterranno 36 elementi tra i quali sono chiaramente visibili forme distinte solo per la presenza di accenti o pseudocaratteri, come apici e punti in alto. Cliccando su **mostra lemmi** si combinano le due opzioni; il risultato consta di 85 elementi.

Opzione 'spezza dopo diacritico'

È stato illustrato nella parte relativa alle codifiche dei testi che le parole che includono un apice o un punto in alto possono essere trattate diversamente: **ch'elli** può essere memorizzata come la singola forma **ch'elli** o come la coppia di forme **ch'** e **elli**; lo stesso vale ad esempio per **co·llui**, memorizzabile come **co·llui** o come **co·** e **llui**.

Corrispondentemente, la ricerca di forme a partire da un brano di testo richiede di precisare come vadano interpretate le parole presenti al suo interno. Per la precisione:

- quando l'opzione **spezza dopo diacritico** non è attiva, parole includenti al loro interno i caratteri ' (apice, ANSI 39), · (punto in alto, ANSI 183) e ' (apice retroverso, ANSI 146) vengono mantenute unite. Quindi, ad esempio, la presenza nel testo di **co·llui** genererà proprio la ricerca della forma **co·llui**;
- quando l'opzione **spezza dopo diacritico** è attiva, le suddette parole vengono spezzate e ad esempio la presenza di **co·llui** genererà la ricerca di **co·** e di **llui**;
- in entrambi i casi l'apice retroverso viene convertito in apice normale.

Un altro effetto dell'opzione **spezza dopo diacritico** riguarda il trattamento del punto: il punto immediatamente successivo a dei caratteri costituenti una parola è considerato punto di abbreviazione se l'opzione **spezza dopo diacritico** è attiva, è considerato invece interpunzione se l'opzione è disattivata.

L'opzione **spezza dopo diacritico** presenta l'intrinseca limitazione di applicare lo stesso trattamento a tutte le parole presenti in un brano. Per risolvere i problemi che ne potrebbero derivare è possibile inserire nel testo, anche ripetutamente, il carattere @: questo carattere inverte temporaneamente, ovvero solo per la

parola che lo segue, lo stato dell'opzione **spezza dopo diacritico**; al termine della parola successiva al carattere @ lo stato iniziale dell'opzione **spezza dopo diacritico** viene automaticamente ripristinato.

Per chiarire con un esempio il comportamento di questa opzione, tornare alla finestra di impostazione di ricerca a brano e rimuoverne il contenuto con [CANCELLA TUTTI I DATI]; disabilitare le opzioni eventualmente attive.

Il contenuto di questa finestra può anche essere digitato direttamente: in questo caso le parole vanno separate da spazi o segni di punteggiatura. Digitare la stringa (notare il punto di abbreviazione dopo **fi** e di interpunzione dopo **ca**):

andà'lo casa. dara'li,dara'line;falla-i fio. po po·lo

Eeguire una prima ricerca con l'opzione **spezza dopo diacritico** disattivata: si troveranno nel corpus **Demo** occorrenze delle forme **andà'lo, casa, dara'li, dara'line, falla-i, po, po·lo**, ovvero di tutte le 8 forme digitate, considerate intere, con esclusione di **fi**.: quest'ultima è in realtà presente all'interno di **Demo** ma poiché il punto è stato assunto essere un segno di interpunzione la parola effettivamente cercata è stata **fi**, che non vi compare.

Ripetere ora la stessa ricerca attivando l'opzione **spezza dopo diacritico**. Il risultato è ora molto diverso e vengono trovate occorrenze delle forme **dara', fio., i, li, line, lo, po**. In particolare, questa volta vengono localizzate come forme distinte le parti che precedono e seguono i diacritici apice e punto in alto. **fi** è stata trovata, mentre si è persa **ca** in quanto la forma cercata è stata **ca**, includente un punto di abbreviazione.

Come ultima prova includere un carattere @ prima di **ca** e ripetere l'ultima ricerca, ovvero mantenendo attiva l'opzione **spezza dopo diacritico**: la stringa da cercare diviene allora (@ può essere a contatto o meno con la parola seguente):

andà'lo @ casa. dara'li,dara'line;falla-i fio. po po·lo

Il risultato questa volta è pressoché identico al precedente, salvo che viene localizzata anche la forma **ca** poiché è stato detto al programma di applicare, solo per essa, le convenzioni valide quando lo stato di **spezza dopo diacritico** è invertito rispetto all'impostazione del *checkbox*: ciò ha significato in questo caso cercare **ca** considerando l'opzione disattivata e quindi assumendo che il punto fosse un segno di interpunzione non appartenente alla forma.

Per tornare alla finestra RICERCA PER FORME tramite selettori cliccare su [INPUT STRINGHE] (notare che i contenuti delle due finestre di inserimento dati non vengono cancellati passando dall'una all'altra).

Attenzione: eventuali segni diacritici isolati o sequenze di soli segni diacritici, come potrebbe essere · (punto in alto più apice), vengono rimossi e quindi non considerati ai fini della ricerca.

RICERCA DELLE FORME MAI LEMMATIZZATE

Questo tipo di ricerca è in tutto e per tutto analoga a quello standard; è anche combinabile con le diverse alternative disponibili, compresa la ricerca a partire da un brano di testo appena discussa in questa Lezione.

Come esempio portarsi alla finestra MODALITÀ RICERCA e selezionare [RICERCHE.. + PER FORME], quindi digitare **tar*** e cliccare su [AVVIO RICERCA]: si otterranno 13 forme diverse presenti in **Demo**.

Ora tornare alla finestra MODALITÀ RICERCA con [VAI A.. + MODALITÀ RICERCA] e selezionare [RICERCHE.. + PER FORME MAI LEMMATIZZATE]; ripetendo la stessa ricerca solo 3 delle forme precedenti vengono localizzate e ciò significa che di queste 3 forme non esistono occorrenze lemmatizzate nel corpus, a differenza delle altre 10. Per verificarlo, tornare a MODALITÀ RICERCA e impostare nuovamente la ricerca di **tar*** con [RICERCHE.. + PER FORME], attivare **mostra lemmi** e usare [AVVIO RICERCA]. L'analisi delle 16 coppie forma-lemma ottenute conferma che solo **tardanza, tardi** e **tarsilla** sono prive di occorrenze lemmatizzate.

L'unica differenza tra questo tipo di ricerca e quello standard è l'assenza dei pulsanti **mostra lemmi**, chiaramente inutili volendo localizzare le forme caratterizzate proprio dal fatto di non essere associate a lemmi.

Questa modalità di ricerca è utile nella fase di lemmatizzazione di un corpus, in quanto consente di localizzare rapidamente le forme al momento completamente escluse dal processo di lemmatizzazione.

LEZIONE 34. COME VISUALIZZARE ED ESPORTARE LE NOTE, UNA DIVERSA EDIZIONE O UNA TRADUZIONE

In precedenza (capitolo 4 e Lezione 14) era stato mostrato come associare ai testi di partenza note e testi paralleli contenenti una diversa edizione o una traduzione.

Il contenuto di note e traduzioni non può costituire di per sé oggetto di una ricerca, però è sempre possibile prenderne visione quando tali elementi riguardano le occorrenze localizzate in una ricerca. Per vedere come fare, impostare ed effettuare nel corpus **Demo** la ricerca delle forme **filio** e **Christo**, senza attivare **mostra lemmi** o altre opzioni: si ottengono 4 contesti per ognuna delle due forme. Cinque di questi contesti mostrano un simbolo "A" di colore marrone nella riga di testata, dopo il numero d'ordine del contesto, ad indicare che ad essi sono associate note e/o traduzioni.

Visualizzazione di note e traduzioni nella finestra RISULTATI RICERCA (a contesti multipli)

Cliccare sul simbolo "A" in corrispondenza al contesto n. 7: apparirà la nota **Christo (u.f.) | Xpo (ed.)** associata a quell'occorrenza di **Christo**. Usare <Esc> o l'icona di chiusura **X** in alto a destra per chiudere la finestra.

Ora ripetere l'operazione in corrispondenza del contesto n. 1, anch'esso interessato da un brano associato. Nel riquadro che si apre, prima di **Eo tempore..**, è presente una riga contenente solo la sequenza "====": ciò sta ad indicare che il brano riportato subito dopo non è una nota, come nel caso del precedente contesto n.7, ma piuttosto un brano tratto da un testo contenente una traduzione (o una diversa edizione); nel caso specifico si tratta di una traduzione in latino.

Ripetere una terza volta l'operazione col contesto n. 4. Questa volta la finestra che si apre contiene due brani (**dixe...** e **Pelias timens...**). Tra i due vi è una riga contenente "====", la quale separa il primo brano che è una nota dal secondo che è una traduzione. La barra di scorrimento verticale permette di prendere visione dell'intero secondo brano, troppo esteso per le dimensioni della finestra.

Visualizzazione di note e traduzioni nella finestra RISULTATI RICERCA (a contesto singolo)

Aprire la finestra a contesto singolo cliccando su un punto del testo all'interno del contesto n. 7, al quale è associata una nota; come si vede, in questo caso la nota è mostrata automaticamente in un riquadro sottostante quello del contesto. Tutte le funzionalità di questa finestra rimangono per il resto invariate. Se ad esempio si allarga il contesto la nota rimane comunque invariata, in quanto legata esclusivamente al fatto che ad essa è associata la parola centrale del contesto, in questo caso **Christo**.

Tornare alla finestra a contesti multipli e riaprire quella a contesto singolo sul contesto n.1. Questa volta il riquadro inferiore contiene il brano di traduzione, al solito riconoscibile come tale dalla presenza della riga iniziale formata da "====".

Se si espande il contesto, ad esempio per periodi, la parte tradotta rimane invariata. La parte inferiore può benissimo non corrispondere come estensione alla traduzione di quella superiore in quanto quest'ultima è dimensionata a 1, 3, 5, .. *periodi*, mentre l'altra è costituita dal *brano*, incluso nel testo associato, che corrisponde solo al *brano* (del testo base) nel quale è inserita la parola centrale (in questo caso **filio**).

Nel caso di presenza di una traduzione associata la finestra a contesto singolo presenta tuttavia una peculiarità assente negli altri casi: è abilitato l'*option box a brani assoc.* in basso a sinistra.

Selezionandolo hanno luogo alcune modifiche:

- i due riquadri in cui la finestra è divisa, destinati al testo base e al brano associato, vengono ad avere le stesse dimensioni;
- la dimensione del testo base non è più in *periodi*, ma in *brani*. Va ricordato che qui il termine *brano* sta ad indicare ognuna delle suddivisioni arbitrarie del testo base operate dall'utente, alle quali corrispondono altrettanti brani nel testo associato. In questo modo al testo nella parte superiore, lungo un brano, corrisponde esattamente la traduzione riportata nella parte inferiore;
- se a questo punto si variano le dimensioni in brani del testo base operando con i bottoni (**Allarga**) e (**Restringi**), anche il testo associato varierà corrispondentemente, fornendo in ogni istante l'esatta traduzione (nel senso dell'estensione del brano di testo) del brano superiore.

Tornando alla finestra a contesti multipli e ripetendo le operazioni descritte applicandole al contesto n. 4, si ritroveranno gli stessi comportamenti: l'unica differenza è che in questo caso oltre alla traduzione è presente anche una nota (**dixe..**), che viene sempre riportata, distinta dalla traduzione mediante la riga contenente "====".

Esportazione di contesti cui siano associate note e traduzioni partendo dalla finestra a contesti multipli

Trovandosi nella finestra a contesti multipli cliccare su [SELEZIONE], attivare **contesti n.**, digitare **1,7** e cliccare su **(OK)**, lasciando i contesti dimensionati a 31 parole. Poi attivare di nuovo **contesti n.**, digitare **4**, selezionare il formato **a brani**, correggerne la lunghezza da 1 a **3**, infine uscire con **(OK e chiudi)**.

Avviare la stampa con [STAMPA], selezionare come tipo di uscita sia stampante che file, questa volta attivare la sottostante opzione **stampa note/trad.**, lasciare impostato **file nuovo** e procedere con **(OK)**. Assegnare come nome di file **prova5** e avviare con **(Salva)**. I 3 contesti vengono stampati insieme ai rispettivi brani associati (solo la traduzione per il primo, nota e traduzione per il secondo, solo una nota per il terzo). Aprendo il file **prova5.rtf** si troveranno gli stessi dati, solo un po' differenti per quanto riguarda la formattazione.

Notare che per permettere di distinguere il testo base dalle note e dalle traduzioni, le note sono precedute da una riga contenente "---" e le traduzioni da una contenente "====".

Esportazione di contesti cui siano associate note e traduzioni partendo dalla finestra a contesto singolo

Con [AZZERA SEL.] rimuovere le selezioni precedenti, poi cliccare in un punto del testo del contesto n. 4 per portarlo nella finestra a contesto singolo. Attivare l'*option box a brani assoc.*, allargare fino a **5** brani, selezionare il contesto cliccando alla sinistra del numero d'ordine (la lettera **B** evidenzia la selezione a brani) e uscire con **(Esci)**. Stampare su carta il contesto con [STAMPA], attivando l'uscita **stampante** e l'opzione **stampa note/trad.** e avviando la stampa con **(OK)**. Il lungo contesto (da **In quello tempo..** a **..tutto loro tenimento.**) e la relativa traduzione (da **Eo tempore fuerunt..** a **..turpiter expulit.**) sono lunghi 5 brani ciascuno.

CAP. 16 - RICERCHE BASATE SUI LEMMI

In alternativa alle ricerche effettuate direttamente per forme GATTO consente di localizzare le occorrenze all'interno del corpus partendo da altri elementi, direttamente o indirettamente collegati ad esse, ovvero lemmi, categorie grammaticali, disambiguatori e iperlemmi.

Si possono inoltre effettuare ricerche di cooccorrenze di due o più parole individuate sulla base della loro grafia o dei lemmi associati, correlandole eventualmente con la punteggiatura circostante.

Queste modalità di ricerca si basano per lo più su interfacce e opzioni uguali o simili a quelle già viste in precedenza: per una descrizione degli elementi illustrati nelle *ricerche per forme* si rimanda quindi al capitolo precedente, mentre in questo verranno descritte solo le opzioni non incontrate in precedenza.

RICERCA PER LEMMI

L'impostazione è molto simile a quella delle *ricerca per forme*, salvo che in questo caso gli elementi digitati nei selettori e cercati nel corpus sono lemmi anziché forme. Più precisamente come primo obiettivo la *ricerca per lemmi* mira a localizzare tutti i lemmi presenti nel lemmario e le cui entrate lessicali siano in accordo con le stringhe digitate nei selettori e successivamente tutte le occorrenze lemmatizzate con tali lemmi.

Opzione 'mostra forme'

In maniera del tutto speculare all'opzione **mostra lemmi** presente nella *ricerca per forme*, qui l'opzione **mostra forme** separa le occorrenze lemmatizzate con uno stesso lemma in funzione delle forme di appartenenza. Ad esempio, la ricerca del lemma **casa** nel corpus **Demo** fornisce come risposta la presenza di 15 occorrenze; se si attiva l'opzione **mostra forme** si scopre che tali occorrenze sono relative a 8 forme diverse.

Opzione 'lemmi muti'

Se nel corpus è stato precedentemente importato il lemmario di un altro corpus, l'operazione ha probabilmente introdotto negli archivi del corpus ricevente delle coppie forma-lemma che in precedenza non vi comparivano; data una di queste coppie, nel corpus non esistono occorrenze della forma che vi compare, lemmatizzate col lemma incluso in quella coppia. Per questo motivo tali coppie sono dette *mute*. Un lemma collegato alle forme solo tramite *coppie mute* è chiamato a sua volta *lemma muto*. Se nel corpus sono presenti coppie forma-lemma mute, diviene visibile l'opzione **lemmi muti** che, se attivata, estende la ricerca

anche a coppie e lemmi muti. Una ricerca basata sugli elementi *muti* ha senso se, nell'ACCUMULATORE, si fa poi uso delle opzioni **non lemmatizzate** e/o **escl. lemmi selez.**

Opzione 'forma pronominale'

È un'opzione non presente nelle *ricerche per forme*, in quanto specifica di quelle *per lemmi*. Estende automaticamente la ricerca di un verbo alla forma riflessiva; ad esempio la ricerca di **lavare**, attivando questa opzione, localizza anche i lemmi **lavare(-si)** e **lavare(-se)** se presenti nel corpus.

Opzione 'relazione con iperlemmi'

I lemmi localizzati nel corso di una ricerca possono essere associati o meno a degli iperlemmi. Questo dato è ininfluenza ai fini della ricerca. L'opzione '*relazione con iperlemmi*' consente invece di restringere la ricerca dei lemmi (e dei relativi contesti) solo a quelli che sono privi di connessioni con iperlemmi o, al contrario, solo a quelli che ne presentano. Si tratta di un'opzione utile principalmente in fase di introduzione degli iperlemmi.

RICERCA PER CATEGORIE GRAMMATICALI

Anche in questo caso la ricerca effettuata prende le mosse dai lemmi, i quali vengono però cercati non sulla base dell'entrata lessicale ma della categoria grammaticale di appartenenza; non si tratta quindi della ricerca di lemmi specifici ma piuttosto di gruppi di lemmi. L'indicazione della categoria grammaticale può essere fatta per digitazione diretta o per selezione dalla lista delle categorie grammaticali presenti nel corpus, generata automaticamente.

RICERCA PER DISAMBIGUATORI

Come spiegato in precedenza il disambiguatore è una stringa che può essere associata ad un lemma per distinguerlo da altri caratterizzati da uguale entrata lessicale e uguale categoria grammaticale. Questa modalità di ricerca permette di risalire ai lemmi a partire dai disambiguatori.

CONCORDANZE LEMMATIZZATE

L'impostazione è la stessa della *ricerca per lemmi*. La differenza riguarda i contesti ottenuti, i quali saranno presentati ordinati alfabeticamente per lemma; per ogni lemma in ordine alfabeticamente di forma; per ogni forma in ordine di testo e di posizione all'interno del testo. Questo ordinamento differisce da quello standard di GATTO, basato primariamente sulle forme oppure sui testi.

LEZIONE 35. COME EFFETTUARE RICERCHE PER LEMMI

Le interrogazioni per lemmi seguono uno schema formalmente molto simile a quelle per forme, mentre cambia la logica della ricerca. La ricerca *per forme* localizza tutte le occorrenze delle forme specificate, mentre quella *per lemmi* trova tutte le occorrenze lemmatizzate con i lemmi indicati, a prescindere dalle forme di cui le occorrenze sono espressione. La ricerca per lemmi permette anche di risalire, più in generale, a tutte le occorrenze delle forme che siano state lemmatizzate almeno una volta (cioè che abbiano nel corpus almeno un'occorrenza lemmatizzata) col lemma indicato.

Esempio 1: ricerca delle occorrenze lemmatizzate con **andare**.

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE], quindi selezionare il corpus **Demo** e attivare le ricerche per lemmi tramite [RICERCHE.. + PER LEMMI].

Digitare **andare** in un selettore quindi effettuare la ricerca con [CONTESTI]: si otterrà direttamente la lista delle 98 occorrenze lemmatizzate con **andare**; ognuno dei contesti ottenuti è caratterizzato dalla presenza del simbolo @ a destra nella riga di testata, cliccando il quale viene mostrato il lemma associato all'occorrenza, ovvero **andare (v.)**.

Esempio 2: ricerca di lemmi terminanti per **ire**.

Riportarsi nella finestra RICERCA PER LEMMI e digitare ***ire** in un selettore, quindi avviare la ricerca con [AVVIO RICERCA]; si otterranno 136 (per sincerarsene usare [N. ELEMENTI]) lemmi, quasi tutti verbi; per vedere più agevolmente quali e quanti dei lemmi trovati sono o non sono verbi impostare **ordina per categorie gramm.** nel *box* di ordinamento a destra e quindi riordinare la griglia con (**A-Z**): si vedrà che il risultato è costituito da **6** sostantivi maschili e **130** verbi.

Selezionando ad esempio i 6 sostantivi maschili e trasferendoli nell'ACCUMULATORE si può continuare la ricerca esattamente come nel caso della ricerca per forme, fino ad arrivare ad ottenere **22** contesti lemmatizzati con uno dei 6 sostantivi maschili scelti (**ardire, avvenire, desire, martire, sire** e **vestire**).

Esempio 3: ricerca di lemmi terminanti per **ire**, incluse le forme riflesse dei verbi.

Riportarsi sulla finestra RICERCA PER LEMMI e digitare ***ire** in un selettore, attivando la corrispondente opzione **Forma pron.**, quindi avviare la ricerca con [AVVIO RICERCA]; questa volta si otterranno 161 lemmi, che terminano per **ire** o per **ire(-si)** (forse anche dei lemmi terminanti per **ire(-se)**), l'esempio è stato provato prima di inserire la relativa modifica in Gatto).

Esempio 4: ricerca del lemma **andare**, con possibilità di selezione delle forme.

Nella finestra di impostazione RICERCA PER LEMMI digitare **andare**, attivando l'opzione **mostra forme**, quindi avviare la ricerca con [AVVIO RICERCA]; vengono localizzate **69** forme diverse aventi occorrenze lemmatizzate con **andare (v.)** (**86** se è attiva l'opzione **Forma pron.**); selezionando uno o più elementi da questa lista e proseguendo la ricerca si possono ottenere i contesti lemmatizzati con **andare** e relativi alle forme scelte.

RICERCHE PIÙ COMPLESSE BASATE SULLE OPZIONI INCLUSE NELL'ACCUMULATORE

L'ACCUMULATORE contiene alcuni strumenti che permettono di modificare e meglio articolare le ricerche in corso. Prima di descriverli è opportuno vedere in dettaglio gli elementi che possono comparire all'interno della griglia inclusa nella finestra; a questo scopo il suo contenuto verrà predisposto mediante le 4 ricerche seguenti, da eseguire nell'ordine.

1. Ricerca *per lemmi* di **palazzo**, senza attivare nessuna opzione; avviare la ricerca con [AVVIO RICERCA] (si ottengono **5** occorrenze), selezionare la forma così trovata, svuotare eventualmente l'ACCUMULATORE con [SVUOTA ACC.] e copiarcela con [COPIA IN ACC.]. Nelle successive ricerche non svuotare l'ACCUMULATORE.
2. Ricerca *per lemmi* di **castello**, attivando l'opzione **mostra forme**; aggiungere i **4** nuovi elementi all'ACCUMULATORE.
3. Ricerca *per forme* di **casa**, senza attivare nessuna opzione (si ottengono **83** occorrenze); aggiungere la forma così trovata all'ACCUMULATORE.
4. Ricerca *per forme* di **torre**, attivando l'opzione **mostra lemmi**; aggiungere all'ACCUMULATORE i **3** elementi trovati.

Al termine la griglia dell'ACCUMULATORE include **9** righe, corrispondenti ad altrettante combinazioni forma più lemma.

Notare che l'ordine delle colonne, prima forma e poi lemma o viceversa, dipende dall'ultima ricerca effettuata: la prima colonna contiene le forme nelle ricerche *per forme*, contiene i lemmi negli altri casi.

Se si osserva il contenuto della griglia si possono notare alcune differenze di contenuto tra le 9 righe, cui corrispondono differenze nelle occorrenze cui si riferiscono:

- a. le 4 righe centrali, da **castella + castello (s.m.)** a **castellu + castello (s.m.)** e le due righe finali, **torre + togliere (v.)** e **torre + torre (s.f.)**, contengono effettivamente coppie forma più lemma. Derivano o da una *ricerca per forme* con **mostra lemmi** attivo o da una *ricerca per lemmi* con **mostra forme** attivo; ogni riga fa riferimento alle occorrenze della forma citata, lemmatizzate col lemma riportato;
- b. la prima riga contiene * nella colonna delle forme e **palazzo (s.m.)** in quella dei lemmi: questa riga si riferisce a tutte le occorrenze lemmatizzate col lemma citato, a prescindere dalle forme di appartenenza;
- c. la seconda riga, al contrario, indica **casa** come forma e * per i lemmi: la riga va riferimento a tutte le occorrenze della forma **casa**, lemmatizzate (con qualunque lemma) o meno;
- d. la terzultima riga contiene solo la forma **torre**, mentre nella colonna dei lemmi non è riportato niente: questa riga si riferisce a tutte le occorrenze non lemmatizzate di quella forma.

Nei casi b., c. e d., ovvero quando non siano esplicitati contemporaneamente sia la forma che il lemma, la selezione di righe di quel tipo implica sempre e comunque la presentazione di tutte e sole le occorrenze che soddisfino i requisiti riportati nelle righe stesse.

Nel caso a. si può invece utilizzare il riquadro **mostra occorrenze** posto in basso a destra. Esso contiene 3 *check box* rispettivamente caratterizzati dalle diciture **con i lemmi selez., non lemmatizzate** e **escl. lemmi selez.** In partenza il primo dei 3 è sempre attivo, gli altri sono sempre disattivati; questa configurazione dice a GATTO di cercare esattamente ciò che è indicato nelle righe selezionate della griglia, ovvero occorrenze di certe forme, lemmatizzate con certi lemmi. Provare ad esempio a selezionare solo la riga contenente **torre più togliere (v.)** e a passare ai contesti con [MOSTRA CONTESTI]: si ottengono esattamente i 2 contesti attesi; si torni all'ACCUMULATORE con [VAI A.. + ACCUMULATORE] e, mantenendo selezionata la stessa riga della griglia, si disattivi il *check box con i lemmi selez.*, si attivi invece **non lemmatizzate** e si passi di nuovo ai contesti: questa volta si ottengono 3 contesti di **torre**, nessuno dei quali lemmatizzato (confrontare la terzultima riga dell'ACCUMULATORE, che indica proprio la presenza di 3 occorrenze non lemmatizzate di **torre**). Infine si ripeta la prova selezionando solo l'ultimo *check box, escl. lemmi selez.*: si otterrà l'unica occorrenza di **torre** lemmatizzata con un altro lemma, in questo caso **torre (s.f.)**; anche in questo caso confrontare il risultato con il contenuto dell'ultima riga dell'ACCUMULATORE.

In conclusione, nel caso di righe dell'ACCUMULATORE includenti in forma esplicita una forma e un lemma, si può *dirottare* la ricerca dall'obiettivo di partenza, includendo anche (o in sua vece) altre occorrenze della forma, non lemmatizzate o lemmatizzate con altri lemmi. I 3 *check box* possono essere usati singolarmente, come fatto finora, o in una qualunque combinazione: non è difficile rendersi conto che l'uso simultaneo dei 3 *check box* corrisponde alla ricerca di tutte le occorrenze della forma, lemmatizzate o meno.

Per terminare verranno mostrati tre esempi di utilizzo di questo meccanismo.

Esempio 5: ricerca delle occorrenze, non ancora lemmatizzate, di forme riferibili al verbo **portare**.

Impostare la *ricerca per lemmi* del lemma **portare**, con **mostra forme** attivo. Svuotare l'ACCUMULATORE e portarvi tutte le 54 coppie ottenute. Selezionarle tutte, disattivare il *check box con i lemmi selez.*, attivare **non lemmatizzate** e passare ai contesti: si ottengono 116 contesti, ovvero tutti quelli relativi alle forme associabili al lemma **portare** ma caratterizzati dal fatto di riferirsi ad occorrenze non ancora lemmatizzate.

Esempio 6: ricerca delle occorrenze di forme riferibili al verbo **portare** ma lemmatizzate con lemmi diversi da **portare**.

Supponendo di aver provato l'esempio precedente e di aver mantenuto il contenuto dell'ACCUMULATORE con tutte le righe selezionate, partendo da quest'ultimo attivare solo il terzo *check box* a destra **escl. lemmi selez.** e passare ai contesti: si ottengono complessivamente 8 occorrenze di forme, omografe di forme riferibili a **portare**, lemmatizzate con altri lemmi.

Notare che in questi due esempi si è partiti da ricerche *per lemmi* mirate formalmente a cercare un certo lemma (**portare**) e si è finito per ottenere due serie di occorrenze che lo escludono. Questo risultato evidenzia la versatilità delle opzioni presenti in GATTO se usate in combinazione fra di loro.

Il ricorso a questa opzione dell'ACCUMULATORE può essere particolarmente utile quando, a differenza dei casi trattati negli esempi precedenti, la ricerca coinvolga forme collegate a più lemmi e si intenda localizzarne le occorrenze lemmatizzate con lemmi diversi da uno, due o più di essi.

Esempio 7: ricerca delle occorrenze, lemmatizzate con altri lemmi, delle forme associate (anche) al lemma **annanti (prep.)**.

Impostare la *ricerca per lemmi* del lemma **annanti**, con **mostra forme** attivo. Svuotare l'ACCUMULATORE e portarvi le 3 coppie ottenute. Solo due di esse si riferiscono a un lemma **annanti (prep.)**. Selezionarle entrambe, disattivare il *check box con i lemmi selez.*, attivare **escl. lemmi selez.** e passare ai contesti: si ottengono 5 contesti, ovvero i 4 relativi alle occorrenze lemmatizzate con il già scartato

lemma **annanti** (avv.) e in più un'occorrenza di **ananti** lemmatizzata col lemma **annantiché** (cong.). Questo era proprio l'obiettivo della ricerca⁷.

LEZIONE 36. COME EFFETTUARE LE ALTRE RICERCHE BASATE SUI LEMMI: PER CATEGORIE GRAMMATICALI, PER DISAMBIGUATORI, CONCORDANZE LEMMATIZZATE. COME EFFETTUARE RICERCHE COMBinate.

RICERCA PER CATEGORIE GRAMMATICALI

L'esempio riportato utilizza l'attributo *categoria grammaticale* nel senso più generale di categoria attribuibile ad un lemma nell'ambito di una classificazione non necessariamente grammaticale. Sia l'attributo *categoria grammaticale* che l'attributo *disambiguatore* possono essere utilizzati con una certa flessibilità dal punto di vista semantico, fermo restando che le due denominazioni non sono modificabili.

Esempio 1: ricerca delle occorrenze lemmatizzate con lemmi di derivazione latina o associabili a festività.

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE], quindi selezionare il corpus **Demo** e attivare le ricerche *per categorie grammaticali* tramite [RICERCHE.. + PER CATEGORIE GRAMMATICALI]. Per provare entrambi i modi disponibili per specificare i dati da cercare, digitare nel primo selettore **lat.**, poi selezionare nella griglia **categ. gramm. presenti** la categoria grammaticale **fest.** e cliccare su (**<= copia**).

Poiché questo tipo di ricerca può portare ad un numero imprevisto di contesti non è disponibile il menu di passaggio diretto ai contesti: avviare quindi la ricerca con [AVVIO RICERCA]. Si ottengono **5** lemmi; selezionandoli e proseguendo nella ricerca si arriva alla visualizzazione dei **6** contesti lemmatizzati con i lemmi ottenuti sulla base delle relative categorie grammaticali.

RICERCA PER DISAMBIGUATORI

Esempio 2: ricerca delle occorrenze lemmatizzate con lemmi caratterizzati dalla presenza di un disambiguatore iniziante per **co**.

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE], quindi selezionare il corpus **Demo** e attivare le ricerche *per disambiguatori* tramite [RICERCHE.. + PER DISAMBIGUATORI]. Digitare in un selettore **co***, attivare l'opzione **mostra forme** e procedere con [AVVIO RICERCA].

Si otterranno **13** coppie forma più lemma nelle quali al lemma è associato un disambiguatore avente la caratteristica indicata. La ricerca può proseguire normalmente fino ai contesti.

Notare che sia in questo caso che in quello delle ricerche *per categorie grammaticali* le occorrenze che il sistema si avvia a localizzare sono tutte quelle lemmatizzate con i lemmi ottenuti sulla base delle indicazioni fornite. Tuttavia, una volta giunti all'ACCUMULATORE, con le tecniche illustrate nella Lezione precedente è sempre possibile allargare la ricerca ad altre tipologie di contesti.

PRODUZIONE DI CONCORDANZE LEMMATIZZATE

Per motivi tecnici questo tipo di ricerca non è combinabile con gli altri tramite l'ACCUMULATORE, per cui ogni volta che viene messa in esecuzione un messaggio avverte che, per proseguire, GATTO dovrà procedere preventivamente alla cancellazione dell'eventuale contenuto di quella finestra.

Per il resto la ricerca si svolge come una normale ricerca per lemmi, salvo il diverso ordinamento finale dei contesti risultanti.

Esempio 3: produzione delle concordanze lemmatizzate con lemmi inizianti per **caro**.

⁷ La coppia **ananti - annanti** (avv.) non è stata selezionata a scopo dimostrativo, per evidenziare che il meccanismo attivato mostra sia le occorrenze relative ai lemmi non presenti nell'accumulatore, sia quelle relative ai lemmi presenti ma non selezionati.

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE], quindi selezionare il corpus **Demo** e attivare le ricerche di *concordanze lemmatizzate* tramite [RICERCHE.. + CONCORDANZE LEMMATIZZATE]. Rispondere (**Si**) al messaggio presentato, quindi digitare in un selettore **caro***, procedere con [AVVIO RICERCA], passare per l'ACCUMULATORE e continuare fino ad arrivare a i contesti.

I **12** contesti ottenuti sono ordinati per lemma (ad esempio i primi 10 sono lemmatizzati con **caro (agg.)**), a parità di lemma per forma (come si può vedere dalle forme **car,..,Char** presentate nell'ordine) e a parità di forma per testo (le 3 occorrenze relative alla forma **cara** derivano da 2 testi diversi, che si susseguono in ordine cronologico).

Questo ordinamento, diverso da quelli previsti nelle altre tipologie di ricerca, è esattamente quello che si intendeva ottenere; esso si mantiene, naturalmente, anche nel caso di stampa o scrittura su file dei contesti ottenuti.

RICERCHE COMBinate

Come accennato nella parte introduttiva del capitolo 14, le ricerche combinate consistono in due o più ricerche, eventualmente di natura diversa, gli elementi risultanti dalle quali vengono riuniti prima di passare a visualizzare i relativi contesti. La combinazione può aver luogo per *somma* dei risultati ottenuti (logica OR - si accettano tutti i risultati presenti in almeno una delle ricerche effettuate) o per *intersezione* degli stessi (logica AND - si accettano solo i risultati presenti in tutte le ricerche fatte).

L'esempio che segue utilizza la logica AND (verrà evidenziata la differenza rispetto alla logica OR).

Esempio 4: ricerca delle occorrenze lemmatizzate con **sostantivi maschili** terminanti per **ore**.

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE], quindi selezionare il corpus **Demo** e attivare le ricerche per categorie grammaticali tramite [RICERCHE.. + PER CATEGORIE GRAMMATICALI]. Utilizzando la barra di scorrimento della griglia a destra portarsi su **s.m.** (sostantivo maschile); cliccare sulla riga relativa e poi su (**<= copia**) per selezionare la categoria grammaticale. Procedere con [AVVIO RICERCA], eventualmente vuotare l'ACCUMULATORE con [SVUOTA ACC.], selezionare tutti i lemmi trovati con [SEL/DESEL.. + SELEZIONA TUTTO] e trasferire il contenuto della griglia nell'ACCUMULATORE con [COPIA IN ACC.]. In questo modo abbiamo 'messo da parte' tutti i lemmi presenti nel corpus e caratterizzati dalla categoria grammaticale **sostantivo maschile**. Si tratta ora di estrarre da questi i soli lemmi terminanti per **ore**.

Impostare una ricerca per lemmi tramite [VAI A.. + MODALITÀ RICERCA] seguito da [RICERCHE.. + PER LEMMI]; nel primo selettore digitare ***ore** e procedere con [AVVIO RICERCA]; anche questa volta selezionare tutti i lemmi trovati con [SEL/DESEL.. + SELEZIONA TUTTO] (non vuotare l'ACCUMULATORE!). Se ora si proseguisse, come prima, con [COPIA IN ACC.] si otterrebbero nell'ACCUMULATORE tutti i lemmi aventi categoria grammaticale **s.m. oppure** terminanti per **ore** (con automatica eliminazione dei doppi); si tratterebbe di una ricerca combinata basata sulla logica OR (nel senso della individuazione di tutti i lemmi soddisfacenti il primo **o** il secondo criterio).

Procedere invece con [COMBINA IN ACC.]: nell'ACCUMULATORE rimangono solo i 122 **sostantivi maschili** terminanti per **ore** presenti nel lemmario del corpus (logica AND - individuazione di tutti i lemmi soddisfacenti il primo **e** il secondo criterio).

CAP. 17 - GENERAZIONE DI INDEX LOCORUM

La possibilità di accedere velocemente alle occorrenze delle forme volute in tutti i testi di un corpus, estraendo in parallelo una serie di informazioni testuali e bibliografiche ad esse relative ha reso meno centrale l'interesse a realizzare degli Index locorum. Per questo motivo le precedenti versioni di GATTO non prevedevano la possibilità di generarli.

Peraltro gli Index locorum possono comunque costituire dei comodi ausili alla consultazione, soprattutto in conseguenza della loro compattezza.

Per tali motivi è stata aggiunta nel programma la possibilità di produrre questo particolare tipo di uscita, pensata esclusivamente per l'esportazione ovvero come documento stampabile su carta o memorizzabile su file, mentre non ne è prevista la presentazione a video.

Tutte le ricerche effettuabili con GATTO, come già detto più volte, possono essere estese all'intero corpus in uso al momento o, al contrario, essere limitate ad un suo sottoinsieme. Ciò è particolarmente

importante nel caso degli Index locorum, che sovente hanno significato in quanto specificamente riferiti a un testo singolo o ad un gruppo di opere. Per questo motivo la Lezione 37 (che mostra come generare degli Index locorum tramite GATTO), anticipando in parte quanto contenuto nel capitolo 20, esemplificherà anche la restrizione dell'operazione ad un sottocorpus costituito da un solo testo.

Gli Index locorum ottenibili sono di due categorie, a seconda che vengano realizzati basandosi o meno sull'eventuale lemmatizzazione effettuata sui testi: si parla nei due casi rispettivamente di *Index locorum per forme* e *Index locorum lemmatizzati*.

Generazione di *index locorum per forme*

L'operazione viene impostata come una normale ricerca per forme, il che permette di restringerla alle parole di interesse o, digitando il carattere jolly *, di estendere l'Index locorum a tutte le forme presenti. Il risultato consisterà in una serie di *loci*, presentati in ordine alfabetico di forma; le occorrenze di ogni forma saranno raggruppate in base ai testi cui appartengono, nell'ordine di apparizione al loro interno.

Per ogni occorrenza potranno essere riportati, a scelta dell'utente:

- il solo riferimento organico,
- il solo riferimento topografico,
- entrambi i riferimenti.

Come detto, le informazioni ottenute potranno essere stampate su carta e/o registrate su file. L'eventuale scelta del testo o dei testi ai quali limitare l'Index locorum verrà fatta a monte della ricerca.

Generazione di *index locorum lemmatizzati*

Questa volta l'operazione viene impostata a partire da una ricerca dei lemmi, anche in questo caso potendola restringere agli elementi di interesse o, digitando il carattere jolly *, estendendola a tutti i lemmi. Il risultato consisterà anche in questo caso in una serie di *loci*, presentati però in ordine alfabetico di lemma; per ogni lemma compariranno le sole occorrenze lemmatizzate con tale lemma, in ordine alfabetico di forma; infine le occorrenze di ogni forma saranno raggruppate in base ai testi cui appartengono, nell'ordine di apparizione al loro interno.

Per ogni occorrenza potranno essere riportati, a scelta dell'utente:

- il solo riferimento organico,
- il solo riferimento topografico,
- entrambi i riferimenti.

Valgono le considerazioni precedenti per il tipo di uscita e per la scelta dei testi.

Per entrambi i tipi di index locorum, le occorrenze relative a corpus nei quali la presentazione di numeri di pagina e di riga è *disabilitata* conterranno, come riferimento topografico, l'indicazione "**pag. 0.0**". Lo stesso comportamento si ha, nei corpus in cui tale presentazione è *facoltativa*, quando le occorrenze sono incluse in pagine denotate in filgato dal codice **\$0\$** (pagina 0) (vedere capitolo 6).

LEZIONE 37. COME GENERARE INDEX LOCORUM PER FORME E LEMMATIZZATI

Come al solito le cose illustrate nella parte iniziale di questo capitolo verranno applicate in pratica facendo ricorso ad una serie di esempi tratti dal corpus **Demo**.

Esempio 1: produzione dell'Index locorum per forme delle forme inizianti per **sanct**.

Una volta avviato GATTO e aperto l'ambiente **Ricerche** sul corpus **Demo**, nella finestra MODALITÀ RICERCA cliccare su **[RICERCHE.. + INDEX LOCORUM PER FORME]**.

Digitare in un selettore la stringa **sanct*** e procedere con **[AVVIO RICERCA]**; vengono localizzate nel corpus **2017** occorrenze di **10** forme; selezionarle tutte e procedere con **[COPIA IN ACC.]**; selezionare ancora l'intero contenuto della griglia e proseguire con **[MOSTRA CONTESTI]**; arrivati a 300 contesti viene richiesto come continuare: deselezionare il *check box* **ripresenta questa richiesta** e cliccare su **(Continua)**. Nella finestra IMPOSTAZIONI DI STAMPA scegliere **file**, mantenere l'indicazione **file nuovo** e procedere con **(OK)**; nella directory voluta indicare come nome file **prova6** e generare il file con **(Salva)**: un messaggio conferma l'avvenuta scrittura del file.

Se ora si apre **prova6.rtf**, ad esempio con Word, si possono leggere, dopo la data, le forme indicate nella ricerca, a partire da **sancta**, con le relative occorrenze ordinate per testo. Ad esempio, la prima

occorrenza di **sancta** in **Giovanni Campulu, 1302/37 (mess.)** (il testo è individuato mediante il contenuto del campo **titolo abbreviato** in bibliografia) è citata tramite **L. 1, cap. 1 - pag. 7.28**, dove **L. 1, cap. 1** è il riferimento organico, **pag. 7.28** è il riferimento topografico.

Allo stesso modo il resto del file riporta i dati relativi a tutte le 2017 occorrenze localizzate. In alcuni casi il riferimento organico sembra non essere stato riportato: in realtà è formalmente presente, salvo che è stato lasciato vuoto nel testo, come è il caso delle prime due occorrenze riportate, tratte da **Doc. venez., 1315 (02)**.

Come accennato in precedenza si può scegliere di riportare nell'Index locorum solo uno dei due riferimenti disponibili (organico e topografico). Per vedere la cosa in pratica, tornare alla finestra INDEX LOCORUM PER FORME e, sempre partendo da **sanct***, deselezionare nel riquadro **dati da estrarre** posto in basso a destra il **check box rif. topografico**; procedere con [CONTESTI] fino a prelevare tutti i 2017 contesti e ad arrivare alla finestra IMPOSTAZIONI DI STAMPA; scegliere **file** (nuovo) e (**OK**); salvare il risultato in **prova7**.

Restando alla finestra INDEX LOCORUM PER FORME selezionare il solo **check box rif. topografico** in basso a destra e procedere come sopra, fino ad ottenere il file **prova8.rtf**.

È facile vedere, dal confronto dei 3 file, che i contenuti, sempre relativi agli stessi contesti, variano però in funzione delle opzioni selezionate.

Esempio 2: produzione dell'Index locorum lemmatizzato delle occorrenze lemmatizzate col lemma **casa**.

Dalla finestra MODALITÀ RICERCA cliccare su [RICERCHE.. + INDEX LOCORUM LEMMATIZZATI]; impostare come lemma **casa** e procedere con [CONTESTI]; salvare il risultato in **prova9**.

Aprendo quest'ultimo si può vedere che il contenuto è molto simile a quello del caso precedente, salvo due differenze:

- prima delle forme riportate, a partire da **casa**, compare il lemma (in questo caso **casa (s.f.)**), evidenziato in corsivo e con caratteri più grandi;
- sono riportate solo le occorrenze lemmatizzate delle forme.

Esempio 3: produzione dell'Index locorum per forme del testo **Libro d'amministrazione dell'eredità di Baldovino Iacopi Riccomanni**.

In questo caso è necessario definire inizialmente un sottocorpus formato dal solo testo considerato, individuato all'interno del corpus **Demo** dalla sigla **P48**. A questo scopo portarsi nella finestra MODALITÀ RICERCA (finestra di avvio dell'ambiente **Ricerche**) e da lì passare alla finestra di definizione dei sottocorpus tramite [SOTTOCORPUS]. Nel riquadro **Sigla** (terzo dall'alto) digitare **p48** e concludere l'operazione con [OPERA SELEZIONE.. + NUOVA]. Nella griglia compaiono i dati del testo scelto, mentre in basso a destra è evidenziata la dicitura **sottocorpus A**, ad indicare che in questo momento all'interno del corpus **Demo** è definito un sottocorpus, identificato dalla lettera **A** e costituito dal solo testo **P48**. Uscire dalla finestra con [CHIUDI FINESTRA].

Passare all'impostazione della ricerca con [RICERCHE.. + INDEX LOCORUM PER FORME]. Digitare * in un selettore per indicare che l'Index locorum dovrà essere relativo a tutte le forme incluse nel testo, poi attivare il **check box sottoc.** A posto nel riquadro **dominio di ricerca**, a destra; contemporaneamente disattivare il **check box corpus**. Avviare la ricerca con [CONTESTI] fino a localizzare tutti gli 8916 contesti quindi come nei casi precedenti salvare il risultato nel file **prova10**.

CAP. 18 - RICERCHE PER IPERLEMMI

LOGICA DELLE RICERCHE BASATE SUGLI IPERLEMMI

Gli elementi sui quali si sono incentrate le ricerche illustrate nei capitoli precedenti appartengono a tre gruppi: lemmi, forme, occorrenze; queste categorie costituiscono una sorta di catena *lemma -> forma -> occorrenza* o, se si preferisce, una struttura a tre livelli nel più *alto* dei quali si collocano i lemmi. Le ricerche permettono di attraversare questa struttura in una sola direzione, ovvero dai lemmi alle forme e dalle forme alle occorrenze, come evidenziato dalle frecce; le ricerche basate sui lemmi consentono anche di effettuare dei *salti*, passando direttamente dai lemmi alle occorrenze.

In conclusione, le ricerche viste finora consentono i passaggi riassunti dallo specchietto che segue:

- *lemma -> forma -> occorrenza*
- *lemma -> occorrenza.*

Gli iperlemmi si inseriscono come un nuovo elemento della catena e si collocano al livello più alto della struttura, la quale diviene *iperlemma - lemma - forma - occorrenza*. A differenza delle altre categorie di dati gli iperlemmi non costituiscono un blocco unico ma presentano internamente una struttura fine determinata dalla possibile presenza di iperlemmi di livelli diversi; in ultima analisi la catena assume l'aspetto *iperlemma (liv. sup.) - iperlemma (liv. inf.) - lemma - forma - occorrenza*.

All'interno della struttura degli iperlemmi la catena è percorribile nei due sensi, passando cioè da un iperlemma di livello qualunque agli iperlemmi di livello superiore o inferiore ad esso.

Per completare queste considerazioni generali resta solo da aggiungere che, finché si rimane all'interno degli iperlemmi, una stessa ricerca può procedere avanti e indietro indefinitamente, utilizzando gli iperlemmi ottenuti in un passaggio come punti di partenza per quello successivo, eventualmente invertendo la *direzione* di spostamento.

PANORAMICA DELLE RICERCHE POSSIBILI A PARTIRE DAGLI IPERLEMMI

Vi sono diversi tipi di ricerche possibili a partire da uno o più iperlemmi. Obiettivo di queste ricerche è la localizzazione di altri iperlemmi, di lemmi, di forme o di specifiche occorrenze.

Tutti i tipi di ricerche hanno in comune la modalità di avvio, che consiste nell'indicazione degli iperlemmi e del livello da cui si vuole partire (si può specificare un singolo livello oppure *tutti i livelli*) e nella successiva localizzazione degli iperlemmi che soddisfano i requisiti fissati. A quel punto tra gli iperlemmi ottenuti si selezionano quelli a partire dai quali si vuole effettuare la ricerca vera e propria, da scegliere tra quelle disponibili e illustrate qui di seguito; in molti casi il risultato di una ricerca può essere usato come punto di partenza per una ricerca ulteriore.

1. Ricerca degli iperlemmi di livello superiore

Vengono mostrati tutti gli iperlemmi, di livello immediatamente superiore, collegati direttamente con quelli di partenza.

Si possono utilizzare gli iperlemmi così ottenuti come nuovo punto di partenza, selezionando quelli desiderati e proseguendo con una nuova ricerca non necessariamente dello stesso tipo (cioè *ricerca di iperlemmi di livello superiore*).

Questo modalità di ricerca non risulta disponibile qualora gli iperlemmi di partenza siano già del massimo livello presente nel corpus o qualora si sia partiti da iperlemmi *di tutti i livelli*.

2. Ricerca degli iperlemmi di livello inferiore

Vengono localizzati tutti gli iperlemmi, di livello immediatamente inferiore, collegati direttamente con quelli di partenza. Come nel caso precedente, dagli iperlemmi ottenuti si può partire per una nuova ricerca.

Questo modalità di ricerca non risulta disponibile qualora gli iperlemmi di partenza siano di livello 1 o qualora si sia partiti da iperlemmi *di tutti i livelli*.

3. Ricerca dei lemmi

Partendo dagli iperlemmi selezionati dall'utente il programma individua quelli di livello immediatamente inferiore direttamente collegati ad essi; l'operazione viene quindi ripetuta più volte, utilizzando ogni volta gli iperlemmi trovati in precedenza come nuovo punto di partenza, percorrendo verso il basso la struttura ad albero degli iperlemmi fino a raggiungere i lemmi collegati, che vengono visualizzati. Da questo momento la ricerca prosegue come una normale *ricerca per lemmi* sulla base delle nuove indicazioni dell'utente.

Se nel corpus è stato precedentemente importato il lemmario di un altro corpus, l'operazione ha probabilmente introdotto negli archivi del corpus ricevente delle coppie forma-lemma che in precedenza non vi comparivano; data una di queste coppie, nel corpus non esistono occorrenze della forma che vi compare, lemmatizzate col lemma incluso in quella coppia. Per questo motivo tali coppie sono dette *mute*. Se nel corpus sono presenti coppie forma-lemma mute, la ricerca dei lemmi a partire dagli iperlemmi prevede la scelta tra due opzioni, **esclusi lemmi muti** e **inclusi lemmi muti**. Una ricerca basata sui lemmi muti ha senso se, nell'ACCUMULATORE, si fa poi uso delle opzioni **non lemmatizzate** e/o **escl. lemmi selez.**

4. Ricerca delle occorrenze lemmatizzate

Questa ricerca è simile alla precedente, nel senso che ne ripercorre tutte le tappe salvo che non si arresta ai lemmi ma prosegue individuando le forme associate a questi ultimi e mostrando le occorrenze di tali forme lemmatizzate con i lemmi trovati.

5. Ricerca di coppie lemma/iperlemma e relative occorrenze

Anche questa ricerca, come le due precedenti, percorre verso il basso la struttura ad albero degli iperlemmi partendo da quelli selezionati dall'utente. In questo caso però la ricerca si arresta ai lemmi e vengono visualizzate tutte le coppie lemma-iperlemma di livello 1 così incontrate.

A questo punto è possibile selezionare una o più di tali coppie e proseguire la ricerca per localizzare tutte e sole le occorrenze cui siano stati associati in fase di lemmatizzazione entrambi gli elementi di una di esse.

6. Ricerca di occorrenze associate a iperlemmi

Questa ricerca è analoga alla precedente, salvo che non prevede la selezione manuale di specifiche coppie lemma-iperlemma di livello 1. Si parte selezionando degli iperlemmi del livello voluto e si cercano direttamente tutte le occorrenze cui siano stati associati in fase di lemmatizzazione entrambi gli elementi di una delle coppie lemma-iperlemma di livello 1 incontrate durante l'attraversamento della struttura ad albero.

Notare che nelle ultime due modalità di ricerca gli iperlemmi associati alle occorrenze non sono necessariamente quelli selezionati in partenza. Tale corrispondenza si ha solo se si parte proprio da iperlemmi di livello 1, altrimenti gli iperlemmi di livello 1 associati alle occorrenze non sono quelli di partenza ma quelli che si incontrano al livello 1 *scendendo* attraverso lo schema ad albero delle connessioni tra iperlemmi a partire da quelli impostati.

LEZIONE 38. COME EFFETTUARE RICERCHE PER IPERLEMMI

Per seguire meglio gli esempi che seguono si può fare riferimento allo schema degli iperlemmi inseriti nel corpus **Demo**, riportato nel capitolo 12.

Esempio 1: passaggio agli iperlemmi collegati di livelli diversi, a partire dall'iperlemma **attività sensoriali**, di livello 1.

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE], quindi selezionare il corpus **Demo** e attivare le ricerche *per iperlemmi* tramite [RICERCHE.. + PER IPERLEMMI].

Lasciare il *box a tendina* in alto a sinistra posizionato su livello 1, poiché si intende partire da un iperlemma di questo livello; digitare **attività sensoriali** in uno dei selettori **iperlemmi da cercare** e avviare la ricerca con (**Mostra =>**). Viene effettivamente localizzato nell'iperlemmario l'elemento cercato.

Come primo passo si risalirà agli iperlemmi di livello superiore, cioè di livello 2, collegati ad **attività sensoriali**, se ve ne sono; per farlo, selezionare l'unico iperlemma trovato cliccando in un punto qualunque della sua riga, quindi utilizzare [TROVA.. + IPERLEMMI DI LIVELLO SUP.]: si trova un solo iperlemma di livello 2, **uomo fisico**. Per risalire ulteriormente al livello 3 selezionare **uomo fisico** e ripetere la ricerca tramite [TROVA.. + IPERLEMMI DI LIVELLO SUP.]; si trova l'iperlemma di livello 3 **uomo**. Se si cercasse di risalire ulteriormente al livello 4 si troverebbe che ciò non è possibile in quanto manca il menu [TROVA.. + IPERLEMMI DI LIVELLO SUP.], a conferma del fatto che nel corpus **Demo** non esistono iperlemmi di livello 4.

Per trovare ora tutti gli iperlemmi di livello 2 collegati a **uomo** selezionare quest'ultimo e usare [TROVA.. + IPERLEMMI DI LIVELLO INF.]: si ritrova, ovviamente, **uomo fisico**, insieme a **uomo sociale**.

Per trovare infine tutti gli iperlemmi di livello 1 collegati a **uomo sociale** selezionare quest'ultimo e usare [TROVA.. + IPERLEMMI DI LIVELLO INF.]: si trovano **linguaggio** e **scrittura**.

Essendo arrivati al livello 1 non è più disponibile il menu [TROVA.. + IPERLEMMI DI LIVELLO INF.].

In questo esempio è stato selezionato ad ogni passaggio un solo iperlemma: niente avrebbe impedito di selezionarne contemporaneamente 2 o più, quando presenti.

Esempio 2: ricerca dei lemmi collegati all'iperlemma di livello 2 **scienza e tecnica** e localizzazione di occorrenze di alcune forme lemmatizzate anche con alcuni di tali lemmi.

Posizionarsi nella finestra RICERCA PER IPERLEMMI, fare un reset della finestra con [CANCELLA TUTTI I DATI], selezionare nel *box a tendina* in alto a sinistra il livello 2, poi digitare in un selettore **scienza e tecnica** e localizzare l'iperlemma con (**Mostra =>**). Selezionarlo e cliccare su [TROVA.. + LEMMI]; vengono trovati i

14 lemmi associati agli iperlemmi di livello 1 **scrittura** e **matematica** a loro volta connessi con l'iperlemma di partenza (vedere lo schema al capitolo 12).

Da questo momento la ricerca prosegue come una normale ricerca *per lemmi*. Per avere un quadro più analitico usare (**Mostra forme**) e ordinare i risultati per lemmi crescenti, tramite il *box a tendina* sulla destra e (**A-Z**). Le coppie lemma-forma così ottenute sono **99**. Selezionare ad esempio le occorrenze delle forme **perchamene**, **perghamena** e **perghamene** (4 occorrenze in tutto) e copiarle nell'ACCUMULATORE. Selezionarle e attivare anche il *check box non lemmatizzate* nel riquadro a destra **mostra occorrenze**; infine procedere con [MOSTRA CONTESTI].

In conclusione si arriva a visualizzare 5 contesti: notare che una delle 5 occorrenze non è lemmatizzata, ma che è stato possibile raggiungerla pur partendo da una ricerca per iperlemmi.

Esempio 3: ricerca di tutte le occorrenze lemmatizzate con i lemmi collegati (indirettamente) all'iperlemma di livello 2 **scienza e tecnica**.

Partendo dalla finestra RICERCA PER IPERLEMMI, selezionare nel *box a tendina* in alto a sinistra il livello 2, poi digitare in un selettore **s*** e localizzare l'unico iperlemma esistente di livello 2 iniziante per **s** con (**Mostra =>**). Selezionarlo e cliccare su [TROVA.. + OCCORRENZE LEMMATIZZATE]. Vengono localizzate **139** occorrenze: si tratta di tutte le occorrenze presenti nel corpus **Demo** lemmatizzate con i lemmi associati agli iperlemmi di livello 1 **scrittura** e **matematica**, a loro volta collegati col lemma di partenza **scienza e tecnica**, di livello 2.

Esempio 4: ricerca delle occorrenze lemmatizzate con il lemma **occhio (s.m.)** e contemporaneamente associate in modo diretto all'iperlemma **attività sensoriali**.

Nella parte relativa alla lemmatizzazione è stato visto che è possibile associare direttamente ad una specifica occorrenza non solo un lemma ma anche un iperlemma di livello 1, magari per meglio specificare l'accezione da attribuirle. In fase di ricerca è possibile sfruttare questa doppia assegnazione per restringere espressamente la ricerca alle occorrenze aventi questa doppia marcatura.

Nel caso in questione la prima cosa da fare potrebbe essere quella di sincerarsi che effettivamente all'iperlemma di livello 1 **attività sensoriali** sia associato il lemma **occhio (s.m.)**; si tratta ovviamente di una condizione necessaria anche se non sufficiente perché occorrenze come quelle cercate siano presenti nel corpus.

Per cercare il suddetto collegamento tra iperlemma e lemma sulla falsariga degli esempi precedenti partire dalla finestra RICERCA PER IPERLEMMI, selezionare nel *box a tendina* in alto a sinistra il livello 1, poi digitare in un selettore **attività sensoriali** e localizzare l'iperlemma con (**Mostra =>**). Selezionarlo e cliccare su [TROVA.. + COPPIE LEMMA/IPERLEMMA]. Vengono visualizzati i 5 lemmi associati all'iperlemma da cui si è partiti. Selezionare la coppia **attività sensoriali/occhio** e proseguire la ricerca con [TROVA.. + OCCORRENZE ASSOCIATE A IPERLEMMI]. Il risultato è costituito da 2 occorrenze, una della forma **ochi** e l'altra della forma **ochu**. Cliccando sul simbolo **#** a destra nelle rispettive righe di testata si può verificare che entrambe sono state lemmatizzate con **occhio (s.m.)** ed associate all'iperlemma **attività sensoriale**. Le 2 occorrenze trovate sono le uniche nel corpus caratterizzate da questa doppia associazione.

Esempio 5: ricerca di tutte le occorrenze cui siano simultaneamente associati lemmi e iperlemmi (di livello 1) ai quali si arrivi a partire dall'iperlemma di livello 3 **uomo**.

Si tratta di una ricerca simile alla precedente nel senso che si prefigge lo stesso tipo di obiettivo, salvo che in questo caso non ci si limita a specifiche coppie lemma - iperlemma ma si cercano tutte le occorrenze lemmatizzate e contemporaneamente collegate ad iperlemmi di livello 1 ai quali si arrivi scendendo attraverso la struttura ad albero degli iperlemmi, a partire da quelli selezionati.

Dalla finestra RICERCA PER IPERLEMMI, selezionare nel *box a tendina* in alto a sinistra il livello 3, poi digitare in un selettore **uomo** e localizzare l'iperlemma con (**Mostra =>**). Selezionarlo e cliccare su [TROVA.. + OCCORRENZE ASSOCIATE A IPERLEMMI] (si tratta dello stesso menu usato nell'esempio precedente, questa volta però utilizzato senza prima passare attraverso [TROVA.. + COPPIE LEMMA/IPERLEMMA]). Cliccando sui simboli **#** associati alle 4 occorrenze localizzate si possono ritrovare i lemmi e gli iperlemmi associati ad ognuna; ad esempio, le 2 occorrenze di **lingua** sono entrambe lemmatizzate con **lingua (s.f.)** ed associate all'iperlemma **linguaggio**.

Questo tipo di ricerca si ripromette di localizzare più parole in sequenza o poste in prossimità tra di loro. Le parole cercate possono essere specificate direttamente sulla base del loro assetto grafico, oppure partendo dai lemmi associati o dalle relative categorie grammaticali.

Si possono cercare anche combinazioni di forme e specifici segni di interpunzione.

Come nelle ricerche di occorrenze si può accedere direttamente ai contesti oppure passare attraverso l'elenco dei testi contenenti le cooccorrenze e da qui filtrare queste ultime sulla base dei testi cui appartengono.

SINTASSI DELLA STRINGA CHE DEFINISCE LA COCCORRENZA

La cooccorrenza cercata viene descritta tramite una stringa di caratteri, la quale include sia le parole da cercare che gli eventuali codici necessari. In questo paragrafo vengono elencate e brevemente descritte tutte le regole di sintassi previste nella scrittura della stringa; la successiva Lezione fornirà una serie di esempi di applicazione di tali regole.

- si possono cercare parole singole e/o *stringhe*, ovvero gruppi di parole in sequenza;
- all'interno delle stringhe possono essere cercati anche alcuni segni di interpunzione;
- l'intervallo massimo di ricerca è di 10 parole, ovvero vengono cercate solo parole separate da 8 parole intermedie al massimo;
- si possono cercare combinazioni di parole e stringhe formate complessivamente da un massimo di 10 parole; le eventuali interpunzioni sono escluse dal conteggio; l'intervallo di ricerca dovrà essere necessariamente maggiore o uguale al numero di parole cercate;
- si intende col termine *qualificatore* il carattere "/";
- si intendono col termine *codici* i caratteri "L", "G", "C", "P", "X", "^", "B";
- qualificatore e codici si applicano sia a parole singole che a parole incluse nelle stringhe;
- se presenti, qualificatore e/o codici precedono immediatamente la parola cui si riferiscono; se sono presenti sia il qualificatore che i codici, inserire prima il qualificatore, poi i codici, infine la parola cui si riferiscono, tutto senza spazi intermedi;
- tutte le parole vanno digitate in caratteri minuscoli, anche se le parole da cercare contengono lettere maiuscole; tutti i *codici* vanno scritti in caratteri minuscoli;
- ogni parola va separata dalla precedente mediante uno o più spazi;
- per definire una stringa, racchiuderne il contenuto tra virgolette doppie (es.: "**quella volta**" implica la ricerca precisa di **...quella volta...** e non, magari, di **...quella certa volta...**);
- per cercare delle parole in prossimità tra di loro, digitarle senza virgolette (es.: **quella volta** indica la ricerca della cooccorrenza delle parole **quella** e **volta** in prossimità tra loro, quindi verrebbero accettate come cooccorrenze valide sia **...quella volta...** che **...quella era la seconda volta...**, purché nel secondo caso l'*intervallo di ricerca* sia almeno di 5 parole);
- il *qualificatore* "/" posto prima di una parola significa che la parola va cercata in alternativa alla precedente (es.: "**dio santo /mio**" implica cercare **...Dio santo...** oppure **...Dio mio...**);
- ogni parola non preceduta da un codice viene cercata semplicemente sulla base della sua forma grafica (es.: **che** cerca semplicemente la forma **che**);
- ogni parola preceduta dal codice "L" è assunta essere un lemma e implica la ricerca di occorrenze lemmatizzate con tale lemma (es.: **Lmangiare**);
- ogni parola preceduta dal codice "G" implica la ricerca di occorrenze, lemmatizzate col lemma indicato dalla parola oppure non lemmatizzate, di una forma che sia stata lemmatizzata anche con tale lemma. Ad esempio **Gessere** trova tutte le occorrenze delle forme lemmatizzate con **essere**, sia quelle direttamente lemmatizzate con **essere** v. sia quelle non lemmatizzate; sono invece escluse dalla ricerca le occorrenze lemmatizzate con altro lemma, come ad esempio quelle di **stato** (che potrebbe essere accettata come participio passato di **essere**) lemmatizzate però con **stato s.m.**. Notare che se la ricerca è fatta in un sottocorpus (vedere capitolo successivo) le cooccorrenze cercate sono solo quelle incluse al suo interno, ma la lemmatizzazione della forma mediante il lemma indicato può aver avuto luogo in un testo qualunque del corpus, anche esterno al sottocorpus considerato;
- ogni parola preceduta dal codice "C" è assunta essere una categoria grammaticale (es.: **Cs.m.** cerca occorrenze lemmatizzate con un sostantivo maschile);

- in luogo di una parola può, all'interno di una cooccorrenza, essere cercato un segno di interpunzione; per farlo digitare i segni di interpunzione cercati preceduti dal codice "P" e non separati da spazi (es.: P;: cerca ";" o ":");
- per cercare il segno di interpunzione *virgolette doppie*, cioè ",", digitare il carattere due volte (es.: P,"" cerca i segni di interpunzione virgola e virgolette doppie);
- se una parola è preceduta dal codice "X" è oggetto di ricerca espansa (es.: Xche cerca **che**, **ché**, **chè**,...);
- per ogni parola preceduta dal codice "^", se iniziante per consonante, verrà cercata anche la stessa parola con iniziale raddoppiata (es.: ^che cerca **che** oppure **cche**);
- le parole **polirematiche** devono essere racchiuse tra parentesi tonde; i codici come "L", "X", se presenti, devono precedere la parentesi tonda; (es.: L(santa caterina) per cercare le occorrenze lemmatizzate col lemma 'santa caterina');
- il codice "B" da solo sta per **fine periodo** (es.: B ma cerca **ma** a inizio frase);
- più codici possono essere usati insieme, purché non siano in conflitto come significato (es.: ^Xche o X^che cercano **che**, **ché**, **chè**, **cche**, **cché**, **cchè**, mentre LGcasa è un errore);
- i caratteri jolly e i diacritici ammessi nella definizione delle parole da cercare sono "?", "*", "(...)", "<...>", "[...]", "[...-...]" e vengono utilizzati con gli stessi significati con cui vengono usati nelle ricerche di occorrenze,
- il carattere jolly "*" da solo sta ad indicare "una parola qualunque". Può essere usato solo all'interno delle stringhe ma non può costituirne la parola iniziale o finale. Quindi **casa * scuola** e **casa di *** sono ricerche non accettate, mentre **con * sentimento** lo è.

INTERVALLO DI RICERCA

Con questo termine si intende l'estensione della porzione di testo nella quale devono comparire contemporaneamente le parole cercate perché la cooccorrenza sia accettata come valida.

L'*intervallo di ricerca*, quando non si includono nella ricerca anche delle interpunzioni, è definito mediante la combinazione di due parametri: limitazione o meno ad un periodo e numero di parole costituenti; se entrano in gioco delle interpunzioni l'intervallo di ricerca è definito solo in numero di parole.

entro 1 periodo

Questo vincolo specifica che la cooccorrenza deve essere formata da parole non separate da segni di punteggiatura forte e quindi comprese entro uno stesso periodo, a prescindere dalla loro distanza. Questa opzione è in conflitto con la ricerca di interpunzioni entro la cooccorrenza.

entro N parole di testo

L'attivazione di questa opzione considera valida la cooccorrenza solo se tutte le parole cercate (o per lo meno quelle indicate dall'opzione **almeno**: se attiva) compaiono entro una zona di testo ampia al massimo N parole. Se ad esempio si cerca la cooccorrenza di 2 parole **entro 6 parole di testo** questo significa che tra le due parole ne possono essere presenti al massimo altre 4.

DEFINIZIONE DI COCCORRENZE

Questo paragrafo contiene una discussione di natura particolarmente *teorica* su quali combinazioni di occorrenze si debbano effettivamente considerare *cooccorrenze valide* a partire da quanto richiesto dall'utente. La sua lettura non è necessaria al fine di imparare ad effettuare ricerche di cooccorrenze con GATTO.

L'argomento trattato può sembrare abbastanza scontato e di scarso interesse poiché dal punto di vista concettuale è piuttosto facile stabilire un legame preciso tra la richiesta *pensata* e le combinazioni di parole da assumersi come cooccorrenze accettabili *rispetto ad essa*.

In realtà la *richiesta pensata* va poi tradotta concretamente in una *stringa digitata*, da fornire al calcolatore; questo deve poi basarsi (solo) sul comando così introdotto per analizzare le combinazioni di parole effettivamente presenti nei testi, combinazioni che possono presentare intrecci non sempre previsti dall'utente. Un esempio per tutti: partendo dalla richiesta **c* *i chi**, che indica la ricerca di cooccorrenze formate da 3 parole, una iniziante per **c**, una coincidente con la forma **chi** e una terminante per **i**, la frase

...**chi vi crede**... costituisce o no cooccorrenza? No, se si considerano le parole nell'ordine con cui compaiono, perché **chi** è in accordo con **c***, **vi** è in accordo con ***i**, ma **crede** non lo è con **chi**; sì se si associa la prima parola (**chi**) con la terza parte della stringa (**chi**), la seconda parola (**vi**) con la seconda parte (***i**) e la terza (**crede**) con la prima (**c***): in questo caso il programma dovrebbe analizzare 27 possibili combinazioni per vedere se almeno una di esse è in accordo con la richiesta.

Sempre riflettendo sul caso precedente è anche possibile che più combinazioni delle parole presenti in uno stesso *intervallo di ricerca* siano in accordo con la richiesta: a quel punto, quali e quante di queste combinazioni vanno assunte come cooccorrenze *distinte*, da mostrare come eventi diversi?

Nel caso di cooccorrenze includenti forme ad alta frequenza è facile che si possano trovare, a distanza ravvicinata, più occorrenze delle stesse forme. Si pone a quel punto il problema di definire con precisione che cosa si intenda esattamente per *cooccorrenza*, al fine di evitare sia la mancata segnalazione di eventi che si intendono comunemente con questo termine sia al contempo la presenza di *rumore*, ovvero la ripresentazione delle stesse associazioni di parole in prossimità tra di loro all'interno di più *cooccorrenze* individuate dal programma come distinte. Ad esempio, nella ricerca di cooccorrenze tra due forme, la presenza nello stesso *intervallo di ricerca* di 2 occorrenze della prima forma e di 3 della seconda potrebbe dare origine, in linea teorica, a 6 cooccorrenze, la cui visualizzazione separata si tradurrebbe nella ripresentazione, per ben 6 volte, dello stesso brano di testo: appare ragionevole in questo caso evidenziare simultaneamente tutte le 5 parole nell'ambito di un'unica visualizzazione, lasciando all'utente il compito di selezionare ciò che ritiene di effettivo interesse; naturalmente in questo caso l'evento complessivo che viene mostrato corrisponde al senso generale di ciò che si stava cercando, ma non soddisfa in senso stretto i vincoli fissati nella stringa di ricerca impostata (che di per sé richiedeva la presenza di 2 sole parole), e alcune delle combinazioni che si possono ottenere con le parole evidenziate potrebbero non avere nulla a che vedere con le cooccorrenze cui si sta pensando. Problemi più complessi si possono avere nel caso di presentazione ripetuta delle diverse forme cooccorrenti in un ambito di lunghezza complessiva di poco superiore all'*intervallo di ricerca* (es. nell'arco di 16 parole, per cooccorrenze definite entro il limite di 10 parole).

Per garantire l'estrazione di tutta l'informazione, eliminando nel contempo il rumore, sono stati adottati i seguenti criteri:

1. tutte le occorrenze di parole incluse tra quelle cercate, che rientrino nell'*intervallo di ricerca* della cooccorrenza e che soddisfino i vincoli posti, sono evidenziate come appartenenti alla cooccorrenza (ad esempio, la ricerca della cooccorrenza di **che** e **la** potrà fornire tra le cooccorrenze trovate la cooccorrenza ...**che...che...la...**);
2. due cooccorrenze diverse devono differire per almeno una parola evidenziata;
3. nessuna cooccorrenza può risultare *inclusa* in un'altra, ciò significa che alcune ma non tutte le occorrenze componenti possono comparire simultaneamente in un'altra cooccorrenza.

RAGGIUNGIMENTO DEL NUMERO LIMITE DI COCCORRENZE LOCALIZZATE

Nella Lezione 26 era stato fatto riferimento al numero massimo di occorrenze che GATTO localizza, che per default è 300, prima di chiedere se procedere o meno nella ricerca in corso.

Lo stesso meccanismo, con lo stesso valore di default, è presente nel caso di ricerche di cooccorrenze, però con una leggera differenza.

La ricerca di cooccorrenze è eseguita *per testi*, considerando questi ultimi, in questo caso, come elementi indivisibili: pertanto se la cooccorrenza n. 300 verrà localizzata all'interno di un certo testo, la ricerca verrà fermata non subito ma alla fine della scansione di tale testo, producendo quindi l'arresto della ricerca in corrispondenza di un numero di cooccorrenze maggiore o uguale al limite impostato (300).

LEZIONE 39. COME IMPOSTARE ED EFFETTUARE RICERCHE DI COCCORRENZE

L'esecuzione di una ricerca di cooccorrenze è molto semplice, per cui ne verranno dati solo un paio di esempi. Verranno invece presentati molti esempi di impostazione di ricerche di cooccorrenze, per mostrare come utilizzare codici, qualificatori ed opzioni disponibili.

Esempio 1: ricerca della stringa **la casa di**.

Avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [RICERCHE], quindi selezionare il corpus **Demo** e attivare le ricerche di cooccorrenze tramite [RICERCHE.. + COCCORRENZE / INTERPUNZIONI].

Digitare la stringa "**la casa di**" nel riquadro a fondo bianco e avviare la ricerca con [CONTESTI].

Si trovano le **4** occorrenze della stringa **la casa di**, due delle quali, per compattezza, sono evidenziate nello stesso contesto in quanto incluse in uno stesso intervallo di ricerca.

Esempio 2: ricerca delle cooccorrenze di **signore** e **padre**.

Tornare alla finestra di ricerca tramite [VAI A.. + IMPOSTAZIONE RICERCA] e cancellare le impostazioni tramite [CANCELLA I DATI.. + INTERA FINESTRA] o [CANCELLA I DATI.. + SOLO STRINGA DI RICERCA].

Digitare **signore padre** e avviare la ricerca con [CONTESTI]. Si ottengono **3** contesti.

Tornare alla finestra di ricerca, disattivare l'opzione **entro 1 periodo** e ripetere la ricerca. Questa volta si ottengono **5** contesti; nel quarto **signore** e **padre** sono separati da un segno di due punti, nel quinto da un punto e virgola: questi contesti non erano stati localizzati in precedenza in quanto le due parole appartengono a frasi diverse. Per lo stesso motivo nel primo contesto risulta ora evidenziata una seconda occorrenza della forma **padre**, che nell'esempio con limite ad **1 periodo** compariva solo come appartenente al contesto in quanto separata dalle altre da segni di interpunzione forte.

Segue ora una serie di esempi di semplice impostazione di ricerche di cooccorrenze.

la casa in

Cerca tutte le cooccorrenze di **la**, **casa** e **in** comunque disposte entro l'*intervallo di ricerca*. Se si effettua la ricerca si trovano **6** cooccorrenze, la prima delle quali si estende su 9 parole (...**in chilla alegricza, subitamenti, standu tucta la casa in**....)

la casa in (intervallo di ricerca = 6)

(Digitare il numero **6** nel riquadro bianco compreso tra **entro** e **parole di testo**).

Cerca tutte le cooccorrenze di **la**, **casa** e **in** comunque disposte entro l'*intervallo di ricerca*. Se si effettua la ricerca, analoga alla precedente ma con l'intervallo di ricerca di lunghezza ridotta, si trovano non più 6 ma **4** cooccorrenze (mancano la terza e la quinta).

"la /in casa"

Cerca tutte le stringhe **la casa** e **in casa**.

la /in casa

Cerca tutte le cooccorrenze di **la** o di **in** cooccorrenti con **casa** entro l'*intervallo di ricerca*.

"in casa /chasa"

Cerca le stringhe **in casa** e **in chasa**.

la /in casa /chasa /chaxa

Cerca le cooccorrenze di **la** o di **in** con **casa** o **chasa** oppure **chaxa** entro l'*intervallo di ricerca*.

"cas? d*"

Cerca stringhe come **casa di**, **case dei**, **casa della**, **casi da**,...

donna va*<a,e>

Cerca cooccorrenze di **donna** e di una qualunque parola iniziante per **va** e terminante per **a** o **e**, entro l'*intervallo di ricerca*.

tavol<a,e> Lmangiare (entro 1 periodo disattivata)

Cerca cooccorrenze di **tavola** o **tavole** con una occorrenza lemmatizzata con **mangiare**, a prescindere dall'appartenenza alla stessa frase. Se non si precisa il numero di parole per l'intervallo di ricerca, selezionando **1 periodo** l'intervallo è impostato al valore massimo di 10 parole, fermo restando comunque il limite del periodo.

vi<c,t>tori<a,e> Lbattaglia /battaglia

Cerca cooccorrenze di **vittoria**, **vittorie**, **victoria** o **victorie** con occorrenze lemmatizzate con il lemma **battaglia** oppure con occorrenze (lemmatizzate o meno) della forma **battaglia**.

Se si esegue la ricerca si troveranno **4** contesti: tutti meno il terzo sono caratterizzati dalla presenza del simbolo **@**; se si vanno a guardare i lemmi, si troverà **battaglia (s.f.)** per i primi due contesti e **vittoria (s.f.)** per l'ultimo. Il simbolo **@** nel caso delle cooccorrenze è meno significativo rispetto al caso della ricerca di occorrenze in quanto l'informazione che esso fornisce (presenza o assenza di lemmatizzazione e, in caso di presenza, identità del lemma usato) si riferisce esclusivamente alla prima parola evidenziata in rosso nella cooccorrenza, ovvero al primo elemento della cooccorrenza in ordine di comparizione nel testo.

Dal momento che le parole cooccorrenti potrebbero essere numerose sarebbe materialmente impossibile presentare sempre tutti i simboli **@** riferiti alle occorrenze lemmatizzate: anziché sopprimerlo completamente nel caso di cooccorrenze è stato deciso di mantenerlo, con la limitazione di significato suesposta.

Ldonna amorosa

Cerca cooccorrenze di forme lemmatizzate col lemma **donna** cooccorrenti con la forma **amorosa** nell'*intervallo di ricerca*. Se si esegue la ricerca si trovano due cooccorrenze valide (nella seconda l'occorrenza di **donne** è stata evidentemente lemmatizzata con **donna**, ma la cosa non è evidenziabile in quanto l'eventuale simbolo **@** si potrebbe riferire solo alla prima occorrenza presentata, cioè **amorosa**).

Gdonna amorosa

Cerca occorrenze di forme lemmatizzate col lemma **donna**, od anche occorrenze delle stesse forme ma non lemmatizzate, in presenza della forma **amorosa** nell'*intervallo di ricerca*. Se si esegue la ricerca questa volta si trovano **8** cooccorrenze valide, ovvero le precedenti più altre **6** in cui **donna** non è lemmatizzata.

"Cprep. casa"

Cerca **casa** preceduta immediatamente da una **preposizione**, o meglio da una occorrenza esplicitamente lemmatizzata con una preposizione. Eseguendo la ricerca si trova una sola cooccorrenza **...in casa...**

"in casa"

Cerca la stringa **in casa**; eseguendo la ricerca si trovano **12** cooccorrenze valide, una delle quali (riconoscibile dalla presenza di **@**) coincide con quella trovata nella ricerca precedente.

"che lla"

Cerca la stringa **che lla**. Se eseguita, la ricerca trova **3** cooccorrenze.

"Xche lla"

Cerca le stringhe **che lla**, **ché lla**, **chè lla**,.... Se eseguita, la ricerca trova **4** cooccorrenze, 3 delle quali coincidono con le precedenti.

"casa di"

Cerca le stringhe **casa di** (ne risultano presenti **9** in 8 contesti).

"^casa di"

Cerca le stringhe **casa di** e **ccasa di** (ne risultano presenti **10**, 9 coincidenti con le precedenti più una cooccorrenza di **ccasa di**).

"uomo P!" (entro 1 periodo disattivata se ! è punteggiatura forte)

Cerca le occorrenze di **uomo!**. Questo è un esempio di ricerca di cooccorrenze con interpunzioni. In questo caso anzi la parola cooccorrente è una sola, in quanto l'altro elemento è un segno di punteggiatura.

"B e" (entro 1 periodo disattivata)

Cerca la parola **e** a inizio periodo, ovvero immediatamente preceduta da punto, punto e virgola o altro segno di punteggiatura forte.

La Lezione viene conclusa con due esempi di ricerche di cooccorrenze originariamente effettuate su corpus di dimensioni molto più rilevanti di **Demo**; la loro applicazione a quest'ultimo non fornisce risultati particolarmente significativi, ma le relative stringhe di impostazione, le cui strutture nascono da precise finalità di ricerca nel corpus maggiore, danno un'idea delle potenzialità offerte dai codici disponibili nelle ricerche di cooccorrenze.

Esempio 3: ricerca di stringhe formate da 2 parole con un ampio spettro di varianti grafiche:

"<c,ch,cc,cch,k><a,o><p,v>* <c,k,cc,ch,cch,><o,au,u,ol,uo,on,io,os,ou><s,ss,ls,x>*"

Se effettuata, la ricerca fornisce le 3 cooccorrenze ...cavallo cosa..., ...cavaliere così... e ...cavalieri cose...

Esempio 4: ricerca della cooccorrenza, entro l'intervallo di ricerca, delle parole precedenti:

<c,ch,cc,cch,k><a,o><p,v>* <c,k,cc,ch,cch,><o,au,u,ol,uo,on,io,os,ou><s,ss,ls,x>*

Questa volta i contesti ottenuti sono 30 se limitati ad un periodo, 48 in caso contrario.

CAP. 20 - RICERCHE LIMITATE A SOTTOCORPUS. DOMINIO DI RICERCA

Le ricerche descritte nei capitoli precedenti sono state sempre riferite all'intero corpus.

In certi casi è però utile poter limitare la ricerca a uno o più testi del corpus, caratterizzati dall'appartenenza ad uno stesso periodo, area linguistica, tipologia, o magari dall'essere stati scritti dallo stesso autore. Per consentire questa operazione GATTO consente di definire in maniera dinamica dei sottocorpus, ovvero dei sottoinsiemi di testi selezionati tra quelli appartenenti al corpus. Una volta operata questa definizione e fino a che non viene chiuso GATTO o non si passa ad un altro corpus ogni ricerca può essere effettuata, a scelta, sull'intero corpus oppure su uno o più dei sottocorpus definiti in quel momento; poiché uno stesso testo può essere incluso in più sottocorpus GATTO provvede automaticamente ad evitare che nei risultati compaiano dei contesti duplicati.

I sottocorpus diversi che si possono definire contemporaneamente sono al massimo 6. È sempre possibile modificarne la composizione aggiungendo o togliendo testi o costruire nuovi sottocorpus, sempre entro il predetto limite di 6 contemporanei.

L'insieme di testi sui quali opera una ricerca, sia esso costituito dall'intero corpus o da una combinazione di sottocorpus, è indicato col termine *dominio di ricerca*.

Ogni sottocorpus è indicato con una lettera maiuscola, da **A** a **F**. I titoli delle finestre di ricerca indicano sempre qual'è la struttura del *dominio di ricerca* corrente, con diciture come (**intero corpus**), (**sottoc. A, B, C**) o (**testo XXX**).

I testi da includere in un sottocorpus vengono individuati sulla base di una o più condizioni applicate ai dati bibliografici relativi, compresi la sigla e l'anno o il periodo cronologico associato al testo.

Un sottocorpus, inteso come lista di testi, può essere salvato su file e richiamato successivamente, magari in una successiva sessione di lavoro, senza doverlo ridefinire di nuovo. Concettualmente niente impedisce di richiamare in seguito un sottocorpus mentre si sta operando con un corpus diverso da quello di partenza: in questo caso il sottocorpus ricostruito sarà formato dai soli testi previsti al suo interno ma effettivamente presenti nel nuovo corpus.

LEZIONE 40. COME DEFINIRE DEI SOTTOCORPUS

Verrà ora visto concretamente come costruire e modificare la struttura dei sottocorpus, come ottenere informazioni statistiche sul loro contenuto, come salvarli e richiamarli, come esportare le relative informazioni bibliografiche.

Per cominciare, avviare GATTO e portarsi nell'ambiente **Ricerche** mediante [**RICERCHE**], quindi selezionare il corpus **Demo** e accedere alla finestra di gestione dei sottocorpus con [**SOTTOCORPUS**].

La finestra è divisa in 4 parti.

- a sinistra si trova il quadro con cui assegnare i valori dei dati bibliografici in base ai quali costruire i sottocorpus,
- la griglia **Selezione** in alto a destra, inizialmente vuota, mostra il contenuto attuale del sottocorpus che si sta definendo,
- in basso al centro si trova il quadro **Sel. cronologica** per assegnare i testi ai sottocorpus sulla base di criteri cronologici,
- in basso a destra si trova il quadro **Sottocorpus** che specifica qual'è il sottocorpus sul quale si sta operando al momento. I sottocorpus ammessi sono 6, indicati dalle lettere A,...,F; quello evidenziato in rosso è quello corrente, in partenza A: per cambiarlo basta cliccare sulla riga corrispondente alla nuova lettera.

COME COSTRUIRE E MODIFICARE DEI SOTTOCORPUS⁸

Esempio 1: sottocorpus A formato dai testi di autore **anonimo**.

Digitare **anonimo** nella casella **Autore** in alto a sinistra. Il sottocorpus A non va specificato perché è quello attivo di default (vedere riquadro **Sottocorpus** in basso a destra). Costruire il sottocorpus tramite [OPERA SELEZIONE.. + NUOVA]. Il sottocorpus A è stato così ottenuto e il suo contenuto è mostrato dalla griglia: si tratta dei 5 testi individuati dalle sigle R3, IW, OD, D3 e DDT; i testi sono presentati nell'ordine con cui i relativi contesti verrebbero mostrati a seguito di eventuali ricerche. Per ampliare la griglia a tutto lo schermo cliccare in un suo punto qualunque col tasto destra del mouse; per riportarla alle dimensioni originali ripetere l'operazione.

Esempio 2: sottocorpus B formato dai testi di autore **anonimo**, escluso il testo **IW** (primo modo)

Cliccare sull'*option box* **sottocorpus B** nel riquadro **Sottocorpus** in basso a destra: la seconda riga verrà presentata in rosso, evidenziando che il sottocorpus in corso di definizione è il B. Se si clicca sull'*option box* **sottocorpus A** si vedrà ricomparire nella griglia il suo contenuto attuale. Cliccare di nuovo su **sottocorpus B**.

Per rimuovere i contenuti di tutte le altre parti della finestra cliccare su [AZZERA IMPOSTAZIONI].

Per costruire il nuovo sottocorpus digitare come prima **anonimo** nella casella **Autore**, **iw** nella casella **Sigla** e a destra di quest'ultima attivare il *check box* **no**. In questo modo si stanno imponendo 2 condizioni simultanee, ovvero si sta dicendo che il sottocorpus deve essere formato da tutti i testi di autore **anonimo** ma **non** da quello identificato dalla sigla **iw**.

Costruire il sottocorpus tramite [OPERA SELEZIONE.. + NUOVA]. Il sottocorpus B è formato da 4 testi, uno meno del sottocorpus A: il testo mancante è proprio IW.

Esempio 3: sottocorpus B formato dai testi di autore **anonimo**, escluso il testo **IW** (secondo modo).

Per vuotare la griglia cliccare sul riquadro grigio del primo testo (a sinistra della sigla R3) e, tenendo pigiato il tasto <Maiusc>, cliccare sul corrispondente riquadro dell'ultimo testo (DDT): tutte le righe risulteranno selezionate; premere <Canc> per rimuovere completamente il contenuto della griglia (e del sottocorpus B). Rimuovere i contenuti di tutte le altre parti della finestra cliccando su [AZZERA IMPOSTAZIONI].

Digitare come prima **anonimo** nella casella **Autore** e premere [OPERA SELEZIONE.. + NUOVA]; si otterrà l'iniziale sottocorpus di 5 testi; rimuovere il contenuto della casella **Autore** con [AZZERA IMPOSTAZIONI]; nella casella **Sigla** impostare **iw** e attivare il *check box* **no**. Premere [OPERA SELEZIONE.. + RESTRITTIVA]: il sottocorpus viene modificato rimuovendo il testo iw. La logica del menu utilizzato (*restrittiva*) è che i testi appartenenti al sottocorpus per continuare a farne parte devono soddisfare anche la nuova condizione, ovvero di non avere sigla IW (operatore AND). Come ulteriore esempio si azzerino le impostazioni con [AZZERA IMPOSTAZIONI] e si inserisca una terza condizione, digitando **storie*** nella casella **Titolo**, per indicare che i testi devono essere non solo di autore **anonimo** e **non** essere definiti dalla sigla **iw**, ma che devono anche avere un titolo che cominci con la parola **storie**. Premendo nuovamente [OPERA SELEZIONE.. + RESTRITTIVA] si vedrà il contenuto del sottocorpus B ridursi all'unico testo che soddisfa tutte le 3 condizioni.

Esempio 4: sottocorpus B formato dai testi di autore **anonimo**, escluso il testo **IW** (terzo modo)

Azzerare le impostazioni, digitare al solito **anonimo** nella casella **Autore** e premere [OPERA SELEZIONE.. + NUOVA]; si otterrà il solito sottocorpus di 5 testi (notare che per sostituire il contenuto di un sottocorpus non è necessario cancellarlo dalla griglia: è sufficiente sostituirlo con uno nuovo tramite [OPERA SELEZIONE.. + NUOVA]); cliccare nella griglia sul quadratino grigio a sinistra della sigla IW: la riga corrispondente risulterà selezionata. Premere <Canc> per rimuovere il testo dal sottocorpus B.

Esempio 5: sottocorpus C formato dai testi di autore **anonimo** e da quelli il cui titolo inizi con **iscrizione**.

⁸ Si suggerisce di seguire gli esempi che seguono passo passo, eseguendo esattamente le operazioni suggerite. In particolare, non rimuovere i contenuti dei sottocorpus via via prodotti, se non indicato esplicitamente.

Attivare il sottocorpus C cliccando sul corrispondente *option box* in basso a destra. Azzerare le impostazioni, digitare **anonimo** nella casella **Autore** e premere [OPERA SELEZIONE.. + NUOVA]. Azzerare nuovamente le impostazioni. Digitare **iscrizione*** nella casella **Titolo**. Premere [OPERA SELEZIONE.. + AGGIUNTIVA]: il contenuto del sottocorpus viene aumentato tramite l'aggiunta dell'unico testo che soddisfa questa seconda condizione. In questo caso la logica del menu (*aggiuntiva*) è che i testi per appartenere al sottocorpus devono soddisfare almeno l'una o l'altra delle condizioni imposte (operatore *OR*).

Esempio 6: sottocorpus **D** formato dai testi di un anno compreso (o di un periodo almeno parzialmente compreso) tra gli anni **1280** e **1300**.

Attivare il sottocorpus D nel solito modo ed azzerare tutte le impostazioni. Nel riquadro **Sel. cronologica** in basso al centro digitare **1280** nella casella **Anno inizio** e **1300** nella **Anno fine**. Premere [OPERA SELEZIONE.. + NUOVA]. Scorrendo i dati bibliografici dei **6** testi trovati si noterà che al primo è associato un anno che rientra nell'intervallo indicato, mentre agli altri sono associati periodi che vi rientrano almeno parzialmente (anche per un solo anno, come nel caso dell'ultimo testo).

Esempio 7: sottocorpus **D** formato dai testi dell'anno **1315**.

Azzerare tutte le impostazioni. Nel riquadro **Sel. cronologica** in basso al centro tramite il *box a tendina* impostare **Anno esatto**, poi digitare **1315** nella casella **Anno** e premere [OPERA SELEZIONE.. + NUOVA]. Si può vedere che il testo incluso questa volta nel sottocorpus è (l'unico) dell'anno **1315** presente in **Demo**.

Per paragonare questi dati con i dati cronologici di tutti i testi del corpus si può chiudere la finestra con [CHIUDI FINESTRA] e accedere alla bibliografia completa di **Demo** con [BIBLIOGRAFIA]; a quel punto per tornare alla finestra di gestione dei sottocorpus usare [CHIUDI FINESTRA] e [SOTTOCORPUS].

Esempio 8: sottocorpus **D** formato dai testi cui è associato un periodo che comprende l'anno **1274**.

Azzerare tutte le impostazioni. Controllare che sia attivo il sottocorpus D. Nel riquadro **Sel. cronologica** in basso al centro tramite il *box a tendina* impostare **Anno compreso**, poi digitare **1274** nella casella **Anno** e premere [OPERA SELEZIONE.. + NUOVA]. Ai **2** testi ottenuti sono effettivamente associati periodi che includono l'anno **1274**.

COME OTTENERE INFORMAZIONI STATISTICHE CIRCA IL CONTENUTO DEI SOTTOCORPUS

In conclusione, se sono stati provati i vari esempi illustrati fin qui, al momento sono stati popolati i sottocorpus A, B, C e D.

Si possono avere informazioni statistiche sul loro contenuto. Ad esempio, cliccare su [STATISTICHE .. + SOLO OCCORRENZE + SOTTOCORPUS B]: si può vedere che il sottocorpus **B** è formato da **4** testi, per un totale di **29.599** occorrenze delle quali **5.149** lemmatizzate.

Si possono ottenere informazioni più complete con un comando leggermente diverso, tenuto separato dal primo perché nel caso di corpus e sottocorpus di dimensioni elevate a differenza del comando precedente può richiedere un certo tempo per la risposta. Cliccare ad esempio su [STATISTICHE .. + COMPLETE + SOTTOCORPUS C]: si vede che il sottocorpus **C** è formato da **6** testi, per complessive **33.696** occorrenze di **6.169** forme diverse; **6.270** occorrenze sono state lemmatizzate con **2.565** lemmi diversi..

COME SALVARE SU DISCO LA STRUTTURA DI UN SOTTOCORPUS E COME RICHIAMARLA

Come già detto, la struttura di un sottocorpus, ovvero la lista dei testi che lo costituiscono, viene perduta al momento dell'abbandono del corpus corrente per passaggio ad altro corpus o per uscita da GATTO. Volendo quindi riutilizzarla in una sessione di lavoro successiva sarà necessario definirla nuovamente.

È possibile evitare questo passaggio salvando su file la struttura del sottocorpus. Ad esempio cliccare sull'*option box* **sottocorpus A** nel riquadro **Sottocorpus** in basso a destra, quindi cliccare su [FILE .. + SALVA] e salvare il file **prova11**, in una directory a piacere (non usare la directory di GATTO per includervi dei file, in quanto si rischia di modificare inavvertitamente il contenuto della directory o di perdere i file reinstallando GATTO). In questo modo in **prova11** è stata salvata la configurazione del sottocorpus A.

Adesso, immaginando di essere in un'altra sessione di lavoro e di voler ricaricare quella struttura come sottocorpus **D**, cliccare sull'*option box* **sottocorpus D** nel riquadro **Sottocorpus**, quindi cliccare su [FILE .. +

CARICA], selezionare il file **prova11** e aprirlo con [APRI]. Si può vedere che la struttura del sottocorpus salvato su file ha ora sostituito quella attuale del sottocorpus **D**, che ora è quindi uguale all'**A**.

COME STAMPARE O RIPORTARE SU FILE (IN TUTTO O IN PARTE) I DATI BIBLIOGRAFICI DI UN SOTTOCORPUS

Cliccare sull'*option box* **sottocorpus A** nel riquadro **Sottocorpus** in basso a destra, cliccare su [STAMPA.. + TUTTI I TESTI], quindi scegliere il *check box* **stampante** e dare il via con (**OK**). Si ottiene la stampa (meglio se trasversale) su più fogli dei dati bibliografici relativi ai 5 testi del sottocorpus A. In alternativa si sarebbero potuti selezionare preventivamente alcuni di questi testi cliccando nei corrispondenti riquadri grigi a sinistra della sigla (con l'eventuale ausilio dei tasti <Ctrl> e <Maiusc>) limitando ad essi i dati da stampare tramite [STAMPA.. + TESTI SELEZIONATI] e proseguendo come prima. Per ottenere i dati su file, anziché su stampante, fare tutto come in precedenza salvo scegliere il *check box* **file** nella finestra IMPOSTAZIONI DI STAMPA e assegnare il nome **prova12** (come **file nuovo**), cui verrà aggiunta automaticamente l'estensione **txt**.

Per vedere il contenuto del file sarà sufficiente aprirlo con un editore o con un programma di videoscrittura. Aprendolo ad esempio con Word, selezionando tutta la parte successiva alle prime righe che riportano nome del corpus, nome del sottocorpus e data, e usando il menu [TABELLA + CONVERTI IL TESTO IN TABELLA..]⁹ è possibile vedere i dati in forma tabellare.

Se si vogliono escludere uno o più campi delle griglia dalla stampa o dalla scrittura su file, prima di effettuare l'operazione, *trascinare* per ciascuno di essi il separatore destro del campo (ovvero la linea nera verticale che separa la sua intestazione da quella del campo immediatamente alla sua destra) completamente a sinistra restringendo così la larghezza del campo non voluto che sparirà quasi del tutto: il campo non verrà stampato o scritto su file.

LEZIONE 41. COME EFFETTUARE RICERCHE LIMITATE A UNO O PIÙ SOTTOCORPUS

L'uso dei sottocorpus per limitare le ricerche al loro interno è molto semplice e, soprattutto, è uguale per tutti i tipi di ricerche.

Innanzitutto bisogna avere effettivamente definito uno o più sottocorpus usando le tecniche descritte nella Lezione precedente. Per brevità qui si farà conto che la Lezione precedente sia stata seguita integralmente, quindi a questo punto saranno attivi i sottocorpus A, B, C e D (in caso contrario ridefinire il sottocorpus A come formato da tutti i testi di autore **anonimo** e il B uguale all'A, salvo che non contiene il testo **iw**).

Ad esempio si vogliono localizzare tutte le occorrenze delle forme che iniziano per **cas**, limitatamente ai testi inclusi nei sottocorpus **A** e **B**.

Se si è nella finestra SOTTOCORPUS si esca con [CHIUDI FINESTRA]. In tutti i casi, partendo dalla finestra MODALITÀ RICERCA, si scelga [RICERCHE.. + PER FORME] e si digiti **cas*** in un selettore.

Si osservi il riquadro a destra **dominio di ricerca**: a differenza degli esempi precedenti questa volta non solo il *check box* **corpus** è selezionato, ma sono abilitati anche i *check box* relativi ai sottocorpus definiti al momento (A e B se si è seguita l'indicazione precedente oppure A, B, C e D se è stata svolta tutta la Lezione 40); i *check box* relativi a sottocorpus non definiti sono disabilitati.

Lasciando le cose come stanno si inizi la ricerca con [AVVIO RICERCA]: si troveranno **27** forme diverse, con occorrenze distribuite in tutto il corpus, come al solito.

Si torni ora alla finestra RICERCA PER FORME con [VAI A.. + IMPOSTAZIONE RICERCA]; in **dominio di ricerca** si disabiliti il *check box* **corpus** e si abilitino **sottoc. A** e **sottoc. B**, quindi si esegua nuovamente [AVVIO RICERCA]. Questa volta il risultato è diverso: le forme localizzate sono solo **9** e la testata della finestra (così come quelle delle finestre successive se si prosegue la ricerca) riporta la dicitura (**sottoc. A, B**), per ricordare qual'è il *dominio di ricerca* attuale. Se si porta in fondo la ricerca si possono notare due cose:

1. tutte le occorrenze trovate appartengono esclusivamente ai testi inclusi nei due sottocorpus,
2. tutti gli eventuali doppioni sono stati eliminati (i testi inclusi nel sottocorpus B appartengono anche ad A).

Con lo stesso meccanismo si possono limitare ai sottocorpus tutti i tipi di ricerche.

⁹ Impostare le Tabulazioni come separatori tra le colonne.

Per terminare l'argomento vanno date solo due avvertenze, una di uso molto generale e l'altra specifica di certe ricerche di cooccorrenze.

1. Supponiamo che sia stata fatta una ricerca estesa ad un certo *dominio di ricerca* (corpus o combinazione di sottocorpus) passando per l'ACCUMULATORE. Una delle cose che può essere fatta è di tornare ad impostare una nuova ricerca (o ripetere la stessa) cambiando il *dominio di ricerca*: una volta tornati all'ACCUMULATORE si vedrà che il *dominio di ricerca* corrente, riportato in testata, è l'ultimo usato in ordine di tempo, ma che tutti gli elementi (forme, lemmi, coppie forma-lemma) precedentemente inseriti nell'ACCUMULATORE a partire da un *dominio di ricerca* diverso sono ora caratterizzati dalla presenza della dicitura **indef.** nel campo **occ.**, che normalmente riporta il numero di occorrenze presenti. Questo è dovuto al fatto che tali numeri di occorrenze, riferiti ad un altro *dominio di ricerca*, potrebbero non essere più esatti, anzi gli elementi corrispondenti potrebbero non essere nemmeno presenti nel *dominio di ricerca* attuale. GATTO consente comunque di effettuare la ricerca dei contesti di tali elementi, fermo restando che potrebbe anche non venirne localizzato nessuno.
2. Nel caso delle ricerche di cooccorrenze con codice **G** (che consente di localizzare occorrenze lemmatizzate col lemma indicato, o anche non lemmatizzate), l'associazione col lemma può anche essere stata fatta al di fuori del *dominio di ricerca* attuale: in quel caso le occorrenze non lemmatizzate e interne al *dominio di ricerca* vengono comunque localizzate anche se poi all'interno di quest'ultimo non risultano associazioni tra le forme corrispondenti e il lemma indicato.
3. La stessa situazione si verifica eseguendo ricerche di occorrenze basate sui lemmi e limitate a sottocorpus, quando si fa uso del **check box non lemmatizzate** incluso nel riquadro **mostra occorrenze** dell'ACCUMULATORE: anche in questo caso potranno comparire contesti relativi a forme la cui associazione ai lemmi trovati si è verificata al di fuori del sottocorpus considerato.

CAP. 21 - ANALISI DELLE FREQUENZE

GATTO offre la possibilità di conoscere e mettere a confronto le frequenze con cui specifiche forme compaiono in testi diversi; in alternativa il confronto può riguardare le frequenze relative a gruppi di testi (sottocorpus o intero corpus).

I confronti possono riguardare tre valori correlati tra loro, ovvero **frequenze assolute**, **frequenze relative**, **periodicità**. Per chiarezza vengono qui riportate le esatte definizioni adottate (riferite ad una forma):

frequenza assoluta: numero di occorrenze della forma nell'ambito testuale considerato (testo o gruppo di testi);

frequenza relativa: numero di occorrenze della forma nell'ambito testuale considerato, espresso in percentuale del numero complessivo di occorrenze di cui quest'ultimo è costituito;

periodicità: è l'inverso del dato precedente: indica il numero di parole del testo che intercorrono in media tra due successive occorrenze della forma in oggetto: ad esempio, se una certa forma compare 8 volte in un testo di 200 parole, i suoi dati statistici (relativi a quel testo) saranno: frequenza assoluta 8, frequenza relativa 4%, periodicità 25 (cioè ricorre una volta ogni 25 parole).

L'analisi delle frequenze viene avviata partendo da una ricerca *per forme* o *per forme mai lemmatizzate*. Se il dominio di ricerca è costituito dal corpus o da un solo sottocorpus, cioè da un solo insieme di testi, per le forme selezionate vengono mostrate le frequenze complessive e quelle relative ad ogni testo dell'insieme; se invece il dominio di ricerca è formato da due o più insiemi di testi, tra corpus e sottocorpus, automaticamente il confronto viene fatto tra i valori di frequenza relativi a ciascun insieme di testi considerato come un tutt'uno.

LEZIONE 42. COME EFFETTUARE ANALISI DELLE FREQUENZE

Esempio 1: frequenze assolute, frequenze percentuali e periodicità di alcune forme inizianti per **cas** nei testi del corpus **Demo**.

Nel corpus **Demo** si imposti la *ricerca per forme* delle forme che cominciano per **cas**. Avviare la ricerca con [AVVIO RICERCA] senza attivare l'opzione **mostra lemmi**. Si ottengono **27** forme diverse; selezionare ad

esempio le prime 22, da **casa** a **castellu** e avviare l'analisi delle frequenze con [FREQUENZE.. + FREQUENZE ASSOLUTE].

I risultati sono collocati all'interno di una griglia. La prima colonna (preceduta a sinistra dai riquadri grigi di selezione delle righe), contiene le parole selezionate; la seconda, con dati in rosso su fondo giallo, riporta il numero di occorrenze complessive di ogni parola nell'insieme dei testi; le colonne successive riportano le frequenze ripartite per testi. La prima riga, con dati in rosso su fondo giallo, riporta il totale di occorrenze nel complesso dei testi (136, colonna 2) e separatamente per ogni testo. Per accedere ai testi non visibili utilizzare la barra di scorrimento orizzontale; notare che le prime due colonne rimangono bloccate e visibili, per comodità.

Ad esempio, si può notare che il testo **IN** contiene in tutto **102** occorrenze delle forme selezionate, 65 delle quali relative alla forma **casa**.

Per passare alla visualizzazione delle frequenze percentuali usare [FREQUENZE.. + FREQUENZE PERCENTUALI]. Si nota ora che le occorrenze della forma **casa** rappresentano lo **0,07325%** delle occorrenze presenti nel testo **IN** e lo **0,05698%** delle occorrenze totali del corpus. Le occorrenze complessive delle 22 forme scelte rappresentano lo **0,11495%** delle occorrenze totali del testo **IN** e lo **0,09337%** delle occorrenze del corpus.

Per passare alla terza modalità di rappresentazione, utilizzare [FREQUENZE.. + PERIODICITÀ]. Si vede che la forma **casa** compare nel testo **IN** una volta ogni **1365** parole e nel corpus **Demo** una volta ogni **1755**.

Nel testo **OD** compare un'occorrenza di una delle 22 forme ogni **928** parole; nel corpus **Demo** l'evento si verifica una volta ogni **1071** parole.

Per tornare alla finestra FORME LOCALIZZATE utilizzare [CHIUDI FINESTRA].

Esempio 2: frequenze assolute delle forme inizianti per **ho** nei testi di autore **anonimo** del corpus **Demo**.

Dalla finestra MODALITA RICERCA passare alla finestra SOTTOCORPUS con [SOTTOCORPUS], definire il sottocorpus **A** come formato dai (5) testi di autore **anonimo**, quindi tornare alla finestra MODALITA RICERCA con [CHIUDI FINESTRA] e cliccare su [RICERCHE.. + PER FORME]. Digitare **ho*** in un selettore, definire il *dominio di ricerca* come formato dal sottocorpus **A** e procedere con [AVVIO RICERCA]. Selezionare tutte le 7 forme ottenute e avviare l'analisi delle frequenze con [FREQUENZE.. + FREQUENZE ASSOLUTE]. Questa volta la griglia è molto più piccola, in quanto sia le forme localizzate che i testi interessati sono in numero minore; il significato dei dati riportati è ovviamente lo stesso.

Scambio degli assi e limiti della tabella

Gli assi orizzontale e verticale della finestra si possono scambiare tramite [FREQUENZE.. + SCAMBIO ASSI X - Y]; ripetendo il comando si ripristina la configurazione iniziale. Notare che il menu [N. ELEMENTI] riporta ogni volta il numero delle *righe* mostrate al momento, che possono variare di numero a seconda dell'orientazione degli assi.

La tabella consente di mostrare al massimo 100 colonne di dati (oltre alle 2 iniziali). Questo significa che l'inversione degli assi è possibile solo se sia il numero delle righe che quello delle colonne non supera questo valore; in caso contrario, infatti, il menu [FREQUENZE.. + SCAMBIO ASSI X - Y] non è abilitato.

Quando si avvia l'analisi delle frequenze e il numero dei testi coinvolti non supera 100 questi sono riportati tramite le loro sigle sull'asse orizzontale mentre le forme sono elencate nella prima colonna (asse verticale). In caso contrario, purché il numero delle forme non superi 100, vengono riportate queste sull'asse orizzontale e i testi lungo quello verticale. Se sia i testi che le forme superano 100 la tabella delle frequenze non viene aperta e compare un messaggio di avvertimento.

Scorrimento dei dati

La griglia è dotata di due barre di scorrimento, orizzontale e verticale, per visualizzare i dati che non rientrano nella schermata.

Un modo per avvicinare colonne lontane tra loro è quello di *trascinare* il separatore destro di una colonna (ovvero la linea nera verticale che separa la sua intestazione da quella della colonna immediatamente successiva) verso sinistra fino a chiudere la colonna non voluta.

Ricordare che le colonne così *minimizzate* sono **escluse dalla stampa**.

Accesso ai dati bibliografici

Per accedere ai dati bibliografici e statistici di uno dei testi cui fa riferimento la griglia, selezionare la colonna o la riga relativa (cliccando sulla testata se si tratta di una colonna o sul riquadro grigio a sinistra se si tratta di una riga) e poi usare il menu [DATI BIB.].

Ordinamento della griglia

È possibile riordinare l'intera griglia in ordine crescente o decrescente del contenuto di una qualunque delle sue colonne, comprese le prime due. Per farlo cliccare sulla testata della colonna desiderata (che verrà evidenziata), quindi utilizzare [ORDINA.. + A-Z] per ordinarla in senso crescente o [ORDINA.. + Z-A] per ordinarla in senso opposto.

La possibilità di scambiare le righe con le colonne consente di fatto di operare questo ordinamento anche in funzione del contenuto di una riga, semplicemente scambiando preventivamente righe e colonne.

Stampa e scrittura su file del contenuto della griglia

Si può stampare su carta o trascrivere su file il contenuto della griglia delle frequenze. Per farlo utilizzare i menu [STAMPA.. + TUTTE LE RIGHE] o [STAMPA.. + RIGHE SELEZIONATE]. Per la selezione multipla delle righe di una griglia vedere il capitolo 29; per quanto riguarda le colonne, queste normalmente sono esportate tutte: è però possibile escludere dall'esportazione quelle indesiderate *minimizzandole*, ovvero spostando completamente a sinistra il separatore di campo successivo (cioè la linea nera verticale che separa la sua intestazione da quella della colonna alla sua destra); per la selezione utilizzare anche i menu [SEL/DESEL.. + SELEZIONA TUTTO] e [SEL/DESEL.. + DESELEZIONA TUTTO]; se si effettua la scrittura su file al nome scelto per quest'ultimo viene automaticamente associata l'estensione **rtf**; inoltre, sempre nel caso di scrittura su file, il suo contenuto può essere trasformato in forma tabellare (per questo argomento vedere anche il capitolo 29 - *Esportazione del contenuto di una griglia*).

Esempio 3: confronto tra le frequenze percentuali delle forme inizianti per **h** nei testi di autore **anonimo** e nel **Libru de lu dialagu de sanctu Gregoriu** di **Giovanni Campulu**, all'interno del corpus **Demo**.

Utilizzando la finestra SOTTOCORPUS si assegnino al sottocorpus **A** i 5 testi di autore **anonimo** e al sottocorpus **B** il solo testo individuato dalla sigla **in**.

Nella finestra RICERCA PER FORME si digiti **h*** in un selettore e si definisca il *dominio di ricerca* come formato dai soli sottocorpus **A** e **B**, escludendo quindi il corpus completo. Si avvii la ricerca con [AVVIO RICERCA], si selezionino tutte le forme trovate con [SEL/DESEL.. + SELEZIONA TUTTO] e si visualizzino le frequenze cercate con [FREQUENZE.. + FREQUENZE PERCENTUALI].

Al di là dei valori numerici ottenuti si potrà notare che questa volta le frequenze non sono più riferite ai testi considerati singolarmente ma ai sottocorpus A e B; questo è avvenuto perché sono stati scelti almeno 2 insiemi di testi nel definire il *dominio di ricerca*, in questo caso il sottocorpus A e il sottocorpus B: ciò è stato interpretato automaticamente da GATTO come una richiesta implicita di mettere a confronto in blocco i due sottocorpus. Se si fossero volute invece confrontare tra di loro le frequenze relative a ciascuno dei 6 testi inclusi nei sottocorpus considerati (5 nel primo e 1 nel secondo) sarebbe stato sufficiente costruire un unico sottocorpus che li contenesse, avviando poi l'analisi delle frequenze solo su di esso; la presenza di un solo insieme di testi avrebbe indicato a GATTO di procedere all'analisi separata delle frequenze per ognuno dei testi componenti.

Tutte le altre operazioni effettuabili sono le stesse dell'esempio precedente.

Esempio 4: stesso confronto del caso precedente, esteso però anche all'intero corpus **Demo**.

Tornare alla finestra RICERCA PER FORME, attivare nel riquadro *dominio di ricerca* anche il *check box corpus* (portando così a 3 gli insiemi di testi considerati) e riportarsi come prima nella finestra FREQUENZE con [FREQUENZE.. + FREQUENZE ASSOLUTE]. Come si può vedere dal risultato il corpus è trattato alla stessa stregua di un sottocorpus, permettendo così il confronto tra le frequenze relative all'intero corpus e a un suo sottoinsieme.

In casi del genere (quando cioè uno o più testi appartengono a più sottoinsiemi) fare attenzione al contenuto della colonna Totali, che è realizzata semplicemente sommando i contenuti delle altre colonne; essa può contenere dei dati duplicati, come è immediatamente evidente guardando ad esempio la prima riga, relativa alla forma **ha**: sono dichiarate in totale **69** occorrenze di questa forma mentre il corpus completo ne contiene solo **49** (il 69 è semplicemente la somma di 49 + 20 + 0)!

CAP. 22 - LETTURA DEI TESTI. PRODUZIONE DI FORMARI E LEMMARI CON RELATIVI CONTESTI. GENERAZIONE DI INCIPITARI

Esistono informazioni ottenibili tramite GATTO, non immediatamente correlate con le occorrenze. Esse sono.

- visualizzazione di brani di testo, scelti sulla base dei riferimenti organici;
- lista completa (*formario*) delle forme presenti in un testo, in un sottocorpus o in un gruppo di sottocorpus (e frequenze relative);
- lista completa (*formario esclusivo*) delle forme presenti in un testo, in un sottocorpus o in un gruppo di sottocorpus (e frequenze relative) ma non negli altri testi del corpus;
- lista completa (*lemmario*) dei lemmi utilizzati per lemmatizzare un testo, un sottocorpus o un gruppo di sottocorpus (e frequenze relative);
- lista completa (*lemmario esclusivo*) dei lemmi utilizzati per lemmatizzare un testo, un sottocorpus o un gruppo di sottocorpus (e frequenze relative) e mai usati per lemmatizzare gli altri testi inclusi nel corpus;
- incipitario di un testo, un sottocorpus o un gruppo di sottocorpus.

A queste informazioni si accede dalla finestra MODALITÀ RICERCA, eventualmente previa definizione dei sottocorpus da considerare.

LEZIONE 43. COME VISUALIZZARE BRANI DI TESTO SULLA BASE DEI RIFERIMENTI ORGANICI

Supponiamo di voler prendere visione del **capitolo 13 del libro secondo** del già utilizzato **Libru de lu dialagu de sanctu Gregoriu di Giovanni Campulu** (sigla **in**).

Dalla finestra MODALITÀ RICERCA si sceglia [**RICERCHE.. + OPERAZIONI SU SINGOLI TESTI + RIFERIMENTI ORGANICI**], si selezioni la sigla **IN** e si clicchi su (**OK**). Appare una finestra che riporta sulla sinistra i titoli dei riferimenti organici presenti nel testo, il primo dei quali è evidenziato; a destra vi è un secondo riquadro che contiene la dicitura **1 - 154**, anch'essa evidenziata. Il riquadro di destra chiarisce che quello di sinistra sta riportando i primi (e complessivi) 154 riferimenti organici presenti nel testo: se ve ne fossero stati più di 200, il riquadro di destra li avrebbe indicati in gruppi separati di 200 ciascuno (es.: 1 - 200 nella prima riga, 201 - 400 nella seconda e così via); in tal caso andrebbe scelto prima di tutto il gruppo voluto, cliccandovi sopra: il riquadro di sinistra elenca sempre i riferimenti organici (al massimo 200) individuati dalla riga evidenziata a destra. Così definito il gruppo di riferimenti organici che include (o probabilmente include) quello voluto, scorrere verticalmente il riquadro di sinistra fino a raggiungerlo (o fino a decidere che va selezionato un altro gruppo nel riquadro a destra) e cliccarvi sopra per selezionarlo. Nel caso dell'esempio si selezionerà **L. 2, cap. 13**. Cliccando su (**Mostra**) sarà possibile prendere visione della parte di testo individuata dal riferimento organico scelto.

Se la finestra non è sufficiente a visualizzare tutto il testo associato al riferimento organico è sufficiente utilizzare la barra di scorrimento a destra per vederne le altre parti; se il testo è troppo lungo anche per le capacità della barra di scorrimento risulteranno automaticamente attivati i bottoni (**Indietro**) e (**Avanti**), che permetteranno di spostarsi a blocchi entro il testo interessato, scorrendo poi il singolo blocco con la barra di scorrimento.

Il bottone (**Dati bib.**) mostra informazioni bibliografiche e statistiche circa il testo in esame mentre il menu [**BIBLIOGRAFIA**] permette di accedere ai dati bibliografici di tutti i testi del corpus. Per chiudere la finestra secondaria usare i soliti comandi. A quel punto è possibile leggere un altro riferimento organico, ripetere l'operazione su un altro testo col menu [**TESTO**] o cambiare operazione con [**CHIUDI FINESTRA**].

LEZIONE 44. COME OTTENERE FORMARIO, FORMARIO ESCLUSIVO, LEMMARIO, LEMMARIO ESCLUSIVO O INCIPITARIO DI UN SINGOLO TESTO

Si supponga di voler analizzare il contenuto del solito testo contraddistinto dalla sigla **in**, ovvero di **Libru de lu dialagu de sanctu Gregoriu di Giovanni Campulu**, nel senso di avere un quadro completo delle forme presenti nel testo e dei lemmi utilizzati per lemmatizzarlo e di disporre di un incipitario dell'opera; verranno ora passate in rassegna le operazioni eseguibili a questo scopo (a titolo di esempio nel caso del formario esclusivo si procederà anche alla visualizzazione di alcuni contesti).

Presentazione del formario del testo IN

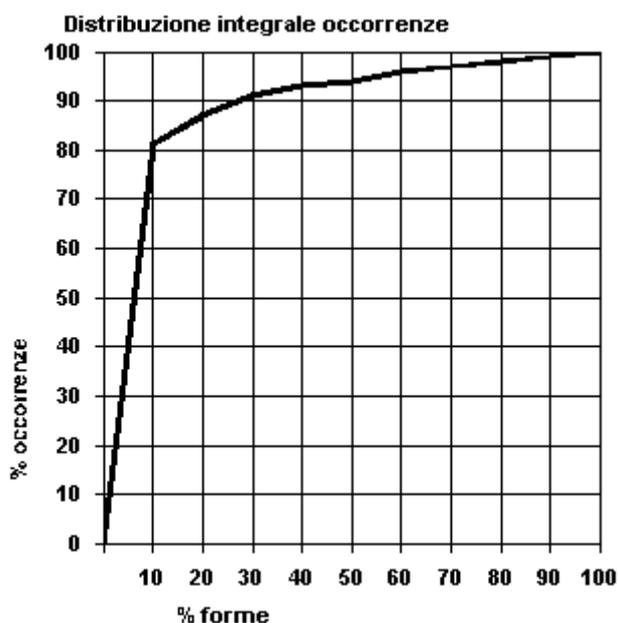
Dalla finestra MODALITÀ RICERCA si scelga [RICERCHE.. + OPERAZIONI SU SINGOLI TESTI + FORMARI], si selezioni la sigla IN e si clicchi su (OK).

In risposta al comando viene presentata una griglia che riporta in ordine alfabetico tutte le forme presenti nel testo **in**, con il numero di occorrenze di ciascuna. Viene inoltre documentato che il testo contiene **88.733** occorrenze di **9.875** forme diverse. Per avere informazioni bibliografiche e altre informazioni statistiche su **in** utilizzare (Dati bib.).

La griglia può essere riordinata in senso crescente o decrescente sulla base del contenuto di una colonna. Volendo ad esempio visualizzarla in ordine di frequenza delle forme, a partire dalla più frequente, cliccare sulla testata della colonna **occ.**, quindi cliccare su (Z-A) nel riquadro **ordina**: la forma più frequente risulta **de**, con **3.356** occorrenze.

Per stampare su carta o scrivere su file il contenuto della griglia (completo o relativo alle sole righe selezionate), usare il menu [STAMPA..] (vedere anche il capitolo 29 - *Esportazione del contenuto di una griglia*). Per deselezionare tutte le righe eventualmente selezionate usare [SEL/DESEL.. + DESELEZIONA TUTTO]. Per agevolare una selezione multipla usare il *checkbox* **blocco selezionati** (vedere capitolo 29).

Un altro modo per interpretare il contenuto della griglia è quello basato sul menu [GRAFICA], il cui effetto è riportato qui di seguito, relativamente al testo **in**.



Per spiegare il senso di questa rappresentazione, converrà rifarsi all'esempio mostrato in figura.

Il grafico si sviluppa dal punto iniziale, in basso a sinistra, caratterizzato dalle coordinate (0,0); il punto successivo, di coordinata orizzontale 10 e coordinata verticale circa 80, sta ad indicare che il 10% delle forme del *formario* (quelle a più alta frequenza) è responsabile dell'80% di tutte le occorrenze del testo; se si estende l'analisi al 20% di forme più frequenti nel testo si trova che ad esse risale circa l'87% delle occorrenze, e così via. Si tratta di un grafico di natura *integrale*, nel quale i risultati intermedi via via si sommano ai precedenti: procedendo fino a considerare il 100% delle forme del testo si troverà che sono ovviamente responsabili del 100% delle occorrenze (ultimo punto del grafico in alto a destra).

Più il grafico si innalza rapidamente nella parte sinistra più il lessico utilizzato è concentrato in un numero limitato di forme, e viceversa.

La finestra che presenta questo grafico consente alcune operazioni accessorie:

- modifica delle scritte riportate,
- variazione del numero di punti di cui è costituito il grafico,
- stampa o copia negli Appunti del grafico, per un successivo trasferimento in un altro documento

Per i dettagli su come eseguire queste operazioni vedere l'esempio 4 nella Lezione 32.

Per tornare alla finestra precedente utilizzare [CHIUDI FINESTRA].

Il menu [BIBLIOGRAFIA] permette di accedere ai dati bibliografici di tutti i testi del corpus.

Il menu [TESTO] permette di visualizzare il formario di un altro testo, senza risalire alla finestra MODALITÀ RICERCA; al contrario il menu [CHIUDI FINESTRA] riporta proprio a tale finestra.

Presentazione del formario esclusivo del testo IN e visualizzazione di alcuni contesti

L'operazione è identica alla precedente, salvo che parte da [RICERCHE.. + OPERAZIONI SU SINGOLI TESTI + FORME ESCLUSIVE].

Scegliendo **IN** questa volta il numero di forme si riduce a **8.770**, per un totale di **50.498** occorrenze: queste forme non compaiono negli altri testi del corpus.

Una volta ottenuta la lista delle forme presenti solo nel testo **IN**, selezionare la forma **abbate** (82 occorrenze), procedere con [COPIA IN ACC.] e dall'ACCUMULATORE, tramite [MOSTRA CONTESTI], si può procedere alla visualizzazione dei relativi contesti.

Presentazione del lemmario del testo IN

Questa operazione ha luogo in maniera del tutto simile alla precedente, partendo da [RICERCHE.. + OPERAZIONI SU SINGOLI TESTI + LEMMARI] e selezionando il testo desiderato.

Scegliendo **IN** si trova, tra l'altro, che il testo in oggetto contiene **11.545** occorrenze lemmatizzate con **3.191** lemmi diversi.

Tutte le altre operazioni effettuabili sono identiche a quelle descritte in relazione alla presentazione del formario, salvo l'uscita grafica che non è disponibile. Il riordinamento della griglia può essere fatto sulla base di una qualunque delle sue colonne.

Presentazione del lemmario esclusivo del testo IN

L'operazione è identica alla precedente, salvo che parte da [RICERCHE.. + OPERAZIONI SU SINGOLI TESTI + LEMMI ESCLUSIVI].

Scegliendo **IN** questa volta il numero di occorrenze si riduce a **4.051**, lemmatizzate con **1.896** lemmi diversi, mai utilizzati in altri testi del corpus.

Generazione dell'incipitario del testo IN

Per ottenere su file l'incipitario del testo **IN**, partire con [RICERCHE.. + OPERAZIONI SU SINGOLI TESTI + INCIPITARI]. Scegliere **IN** e poi decidere se creare un nuovo file oppure utilizzarne uno già esistente, aggiungendo il nuovo incipitario al suo contenuto attuale o sostituendolo ad esso. Il file sarà in tutti i casi un file **RTF**. A titolo di esempio, scegliere **file nuovo** e **clickare su (OK)**; scegliere come nome **prova13**: nella directory scelta verrà creato il file **prova13.incipit.rtf**.

Aperto il file così ottenuto è possibile leggere una serie di record, ognuno dei quali contiene la sigla del testo (**in** nell'esempio), il titolo abbreviato (vedere bibliografia), un riferimento organico (ovviamente nell'ordine in cui compaiono nel testo) e il relativo incipit.

Usando Word è possibile selezionare tutto il contenuto del file, esclusa la riga iniziale che contiene la data, e trasformarlo in una tabella tramite [TABELLA.. + CONVERTI IL TESTO IN TABELLA..] (definire come separatori le tabulazioni); sempre in Word, selezionando poi tutte le colonne della tabella così ottenuta e usando [TABELLA.. + ALTEZZA E LARGHEZZA CELLE] seguito da (**Adatta**) si ottiene una migliore formattazione della tabella.

Sono inoltre disponibili i menu [TESTO] e [BIBLIOGRAFIA] con gli stessi significati dei casi precedenti.

LEZIONE 45. COME OTTENERE FORMARIO, FORMARIO ESCLUSIVO, LEMMARIO, LEMMARIO ESCLUSIVO O INCIPITARIO DI UNO O PIÙ SOTTOCORPUS

Queste operazioni si svolgono in maniera identica alle corrispondenti azioni riferite a singoli testi, salvo che in questo caso si opera su uno o più sottocorpus da selezionare preventivamente.

Inizialmente tramite la finestra SOTTOCORPUS si costruiscano nel corpus **Demo**:

1. sottocorpus **A**, contenente per brevità il solo testo **IN**,
2. sottocorpus **B**, contenente il solo testo **OD**,
3. sottocorpus **C**, contenente il testo **P48**.

Poi, dalla finestra MODALITÀ RICERCA, si selezioni [RICERCHE.. + OPERAZIONI SU SOTTOCORPUS + FORMARI]. Comparirà una nuova finestra, SELEZIONE SOTTOCORPUS, nella quale vengono evidenziati i sottocorpus disponibili al momento; a titolo di esempio selezionare i sottocorpus **A** e **B** e procedere con

(OK) (l'operazione può richiedere un certo tempo). Il risultato è formalmente identico a quello ottenuto nella presentazione del formario di un singolo testo, salvo che tutti i dati sono riferiti all'unione dei sottocorpus considerati. Eventuali doppioni nei calcoli di frequenze, numero di forme, numero di lemmi o altro vengono automaticamente rimossi dal programma.

In modo assolutamente identico si procede per ottenere formari esclusivi, lemmari, lemmari esclusivi e incipitari di uno o più sottocorpus.

Anche in questi casi si può passare alla visualizzazione dei contesti relativi a forme o lemmi selezionati, con meccanismi del tutto analoghi a quelli illustrati nei capitoli precedenti.

CAP. 23 - ACCESSO ALLE INFORMAZIONI BIBLIOGRAFICHE E STATISTICHE

Come è stato mostrato nei precedenti capitoli, in ogni situazione è possibile accedere ai dati bibliografici e statistici del testo sul quale si sta operando, con menu o pulsanti riportanti la dicitura **Dati bib.**. I dati statistici che appaiono sono sempre aggiornati in tempo reale e tengono conto di ogni eventuale modifica apportata al testo. La finestra aperta mediante **Dati bib.** può essere stampata cliccando sull'icona che rappresenta una stampante, posta nella finestra stessa in alto a destra.

L'accesso alla bibliografia complessiva del corpus è possibile tramite i menu [**BIBLIOGRAFIA**] presenti nelle finestre di avvio dei quattro ambienti di GATTO e in alcune altre finestre. La modalità di accesso alla bibliografia complessiva cambia però a seconda dell'ambiente in cui ci si trova.

Prima di illustrare queste differenze va fatta una premessa. A ciascun testo incluso nel corpus in uso deve corrispondere un record di informazioni nell'archivio bibliografico associato; questo però non significa che tutti i record inclusi in quell'archivio debbano riferirsi a testi presenti in quel corpus. Basta pensare al caso di un corpus di 30 testi con relativa bibliografia e a un secondo corpus, costruito per effettuare delle prove, riferito allo stesso archivio bibliografico e nel quale sia stato incluso uno solo dei testi: in questo caso aprendo il secondo corpus avremo che la bibliografia associata contiene 29 record (su 30) non riferiti al corpus in uso.

Tornando ora ai menu di accesso alla bibliografia, valgono le seguenti differenze in conseguenza dell'ambiente in cui ci si trova:

- se si usano i menu [**BIBLIOGRAFIA**] trovandosi all'interno dell'ambiente **Ricerche** o degli ambienti **Lemmatizzazione** cliccandoli viene visualizzata una griglia che mostra le informazioni tratte dall'archivio bibliografico e relative ai soli testi inclusi nel corpus in uso al momento (nel caso dell'esempio ora citato mostrerebbe quindi un solo record). Se sono definiti dei sottocorpus è possibile selezionarne uno o più e limitare la visualizzazione ai soli dati appartenenti ai testi inclusi in essi. La finestra è aperta *in sola lettura*, ovvero i dati possono essere letti ma non modificati. I record sono ordinabili in base al contenuto di una colonna qualunque;
- se si usa il menu [**BIBLIOGRAFIE..**] nell'ambiente **Gestione base dati** si deve selezionare l'archivio bibliografico tra quelli presenti nei dischi del calcolatore; l'eventuale corpus in uso al momento viene chiuso e la finestra dà accesso *in lettura e scrittura* all'intero contenuto dell'archivio bibliografico selezionato (sempre nel caso dell'esempio citato in questo capitolo verrebbero mostrati 30 record). I record sono ordinabili in base al contenuto di una colonna qualunque. L'uso del menu [**BIBLIOGRAFIE..**] verrà esaminato nella parte V del manuale (capitolo 28).

LEZIONE 46. COME PRENDERE VISIONE DEI DATI BIBLIOGRAFICI RELATIVI A TESTI INCLUSI IN CORPUS O SOTTOCORPUS

L'accesso ai dati bibliografici in lettura è possibile nelle finestre di avvio e in alcune finestre secondarie degli ambienti **Ricerche** e **Lemmatizzazione**, tramite i menu [**BIBLIOGRAFIA**].

Per una dimostrazione sull'uso di questi menu, nel corpus **Demo** costruire 3 sottocorpus:

1. sottocorpus **A**, contenente i testi **IN, IW, KAA**,
2. sottocorpus **B**, contenente i testi **IN, OD**,
3. sottocorpus **C**, contenente i testi **OD, P48**.

(Utilizzare a questo scopo, nella finestra SOTTOCORPUS, il riquadro **Sigla** e i menu [**OPERA SELEZIONE.. + NUOVA**] e [**OPERA SELEZIONE.. + AGGIUNTIVA**]).

Fatto questo, dalla finestra MODALITÀ RICERCA cliccare su [**BIBLIOGRAFIA**].

La griglia mostra le colonne relative ai primi dati bibliografici dei testi che costituiscono il corpus **Demo**. Per vedere gli altri dati utilizzare la barra di scorrimento orizzontale; in alternativa, volendo ad esempio vedere tutti i dati del testo **p48**, cliccare sulla riga di quest'ultimo che apparirà evidenziata e usare il menu [SCHEDA]. Il pacchetto completo dei dati apparirà nelle caselle della scheda. Volendo vedere in questo modo i dati di un altro testo cliccare sulla riga relativa senza bisogno di richiudere la scheda: provare ad esempio a cliccare su un punto della riga relativa al testo **btb**. Per ripristinare la visualizzazione a colonne usare [NO SCHEDA]. Volendo ordinare i testi in ordine alfabetico di titolo, scegliere titolo tra le voci offerte dal *box a tendina* **ordina** posto in basso a sinistra e cliccare su **(A-Z)** (il bottone utilizzato rimane evidenziato).

Supponiamo ora di voler vedere solo i dati relativi ai testi inclusi nei 3 sottocorpus definiti. Nel riquadro in basso a destra disattivare il *check box* **corpus** e attivare invece quelli relativi ai 3 sottocorpus **A**, **B** e **C**; per rendere effettiva la nuova impostazione utilizzare [AGGIORNA]: il nuovo contenuto della griglia soddisfa la richiesta mantenendo l'ordinamento impostato e rimuovendo gli eventuali doppioni (in questo caso **in** e **od**).

PARTE V - OPERAZIONI DI GESTIONE DEI DATI

Questa parte del manuale descrive in maniera completa tutte le operazioni effettuabili nell'ambiente **Gestione base dati**. Alcune di queste sono state già incontrate nel corso delle descrizioni relative alle modalità di costruzione di un corpus o di inserimento di un testo al suo interno.

CAP. 24 - GESTIONE DEI CORPUS

Una volta che un corpus è stato costruito e sono stati inseriti al suo interno i testi desiderati sono possibili operazioni quali la modifica di alcune sue caratteristiche, il compattamento o la cancellazione del corpus.

Modifica dei dati di un corpus

Con questo termine non si intende l'inserimento di nuovi testi in un corpus (già illustrato nella parte II), la loro lemmatizzazione (parte III) o cancellazione (Lezione 48), ma la modifica di alcuni degli attributi specifici del corpus, indipendenti dai testi contenuti. Gli attributi di un corpus modificabili successivamente alla sua costruzione sono:

- **descrizione** del corpus, ovvero la stringa descrittiva che compare assieme all'elenco dei corpus localizzati utilizzando il menu [CORPUS] nelle finestre di avvio di tutti gli ambienti;
- **nome e localizzazione dell'archivio bibliografico** associato al corpus;
- **nome della directory contenente i testi** del corpus.

Compattamento di un corpus

Il motore di ricerca del database sottostante GATTO utilizza una serie di parametri che consentono di ottimizzarne le prestazioni, riducendo i tempi di risposta; questi parametri sono legati al numero e alle caratteristiche dei dati contenuti negli archivi. Effettuando inserimenti di testi o modifiche alla lemmatizzazione o alla iperlemmatizzazione i parametri di ottimizzazione risultano via via sempre meno allineati con il contenuto della base di dati e ciò può introdurre rallentamenti nelle risposte. Per ripristinare le prestazioni originarie può essere talvolta opportuno procedere all'operazione di *compattamento* del corpus la quale, oltre a ridurre le dimensioni, ottimizza i parametri di controllo sulla base degli effettivi contenuti degli archivi.

Copia di un corpus

Questa operazione consente di ottenere un duplicato di un corpus esistente. Il nuovo corpus potrà essere collocato nello stesso disco di quello di partenza ma con un nome diverso, poiché in uno stesso disco non sono ammessi due corpus con lo stesso nome. In alternativa è possibile inserire il corpus duplicato in un altro disco, se disponibile, e in questo caso il nome può essere uguale o meno a quello del corpus di partenza.

Cancellazione di un corpus

Il significato di questa operazione è evidente. C'è solo da chiarire che essa non tocca assolutamente i contenuti della directory dei testi (quindi file filgat, ecc.) né l'archivio bibliografico associato al corpus rimosso.

Attenzione però al suggerimento seguente: non inserire manualmente nessun file (archivio bibliografico compreso), **nella directory contenente un corpus** (nel caso di **Demo** la directory **demo.gat.**) perché l'operazione di *cancellazione del corpus* produce la rimozione di tale directory con tutto il suo contenuto e delle relative sottodirectory.

Attivazione dell'ordinamento dei contesti per sintagmi

Una delle modalità di ordinamento nella presentazione dei contesti ottenuti nel corso delle ricerche è quella detta per *sintagmi*. Significato preciso e modalità di utilizzo di questa funzione sono descritte nella Lezione 29. Qui basta chiarire che all'atto della creazione di un nuovo corpus questa funzione è formalmente attiva, ma viene disabilitata ogni volta che nel corpus stesso vengono inseriti o sostituiti uno o più testi. L'operazione descritta nella Lezione 47 riattiva questa funzione. La riattivazione è valida fino alla prima successiva operazione di inserimento o sostituzione di testi dopo di che, volendolo, va ripetuta (l'esecuzione

è effettuata solo a richiesta dato che su corpus di dimensioni elevate richiede un certo tempo). Per conoscere la situazione del corpus è sufficiente vedere se il menu che attiva l'ordinamento è presente: se non lo è vuol dire che l'ordinamento per sintagmi è attivo, e viceversa.

LEZIONE 47. COME MODIFICARE, COMPATTARE, COPIARE, CANCELLARE UN CORPUS. COME ATTIVARE L'ORDINAMENTO DEI CONTESTI PER SINTAGMI.

Le operazioni descritte nel seguito utilizzeranno il corpus **Prova**, la cui creazione e popolazione con 3 testi è stata descritta nella Parte II del manuale. Prima di procedere si suggerisce di copiare la relativa directory, **prova.gat**, con tutte le sottodirectory in un'altra cartella o in un floppy disk, in modo da ripristinarla una volta modificato e cancellato per esercizio il corpus.

Preliminarmente si crei una directory con nome a piacere, per esempio **C:\pippo**, e si copi al suo interno l'intero contenuto della directory **C:\gattesti**, a suo tempo costruita introducendovi l'archivio bibliografico di **Prova**, ovvero **esebib.mdb** e i 5 file relativi ai 3 testi **es1**, **es2** e **es3**; si rinomini poi il nuovo archivio **C:\pippo\esebib.mdb** in **C:\pippo\provabib.mdb**.

Modifica dei dati di un corpus

Si utilizzeranno ora la directory **C:\pippo** e l'archivio bibliografico **C:\pippo\provabib.mdb** come nuovi attributi del corpus **Prova**, la cui stringa descrittiva **corpus di prova contenente 3 testi** verrà cambiata in **corpus modificato**.

Dalla finestra di avvio di GATTO scegliere [GEST. BASE DATI], quindi [CORPUS]. Cliccare sulla riga del corpus **Prova**, che risulta così evidenziato (non occorre aprirlo con [APRI]), quindi usare [GESTIONE.. + MODIFICA]. All'interno della finestra MODIFICA DATI CORPUS, nella casella **descrizione del corpus**: digitare **corpus modificato**; in **archivio bibliografico**: digitare **c:\pippo\provabib.mdb** e in **directory contenente i testi**: scrivere **c:\pippo**. Scegliere (OK) per registrare le modifiche.

Per verificare che le nuove assegnazioni sono corrette cliccare di nuovo sulla riga del corpus **Prova** e questa volta aprirlo con [APRI]. Se non vengono segnalate difficoltà vuol dire che tutte le modifiche apportate al corpus sono andate a buon fine; in caso contrario analizzare i messaggi forniti e cercare di correggere gli eventuali errori (nomi di file o directory digitati male o altro).

Compattamento di un corpus

Sempre restando nella finestra APERTURA E GESTIONE GENERALE DEI CORPUS cliccare sulla riga del corpus **Prova**, che risulterà evidenziata, poi cliccare su [GESTIONE.. + ORDINA E COMPATTA]; durante l'operazione appare un riquadro a fondo giallo nella parte inferiore della finestra e il compactamento è concluso quando esso scompare.

Se un archivio risulta già ordinato la cosa viene segnalata e il programma passa alla fase successiva dell'operazione.

In caso di dubbi il compactamento può essere ripetuto quante volte si vuole senza che questo crei problemi.

Copia di un corpus

Dalla finestra APERTURA E GESTIONE GENERALE DEI CORPUS cliccare sulla riga del corpus **Demo**, che risulterà evidenziata, poi cliccare su [GESTIONE.. + COPIA]; apparirà una finestra per selezionare nome e collocazione del corpus duplicato. Nella casella *corpus* digitare **Demo2**; nel sottostante box a tendina se il disco proposto è diverso da **C** cliccare sulla freccia a destra per visualizzare i dischi disponibili, quindi cliccare sul disco **C**. Per effettuare la duplicazione del corpus cliccare su (OK) (o fermare tutto con (Annulla)).

Da questo momento, GATTO presenterà tra i corpus disponibili sia **Demo** che **Demo2**.

Nota: Qualora, per qualunque motivo, la costruzione del corpus duplicato non andasse a buon fine, procedere alla rimozione manuale della relativa directory (ad esempio **C:\Demo2.gat** per il corpus **Demo2**) e delle relative sottodirectory con Gestione risorse di Windows o altro strumento analogo.

Cancellazione di un corpus

Dalla finestra APERTURA E GESTIONE GENERALE DEI CORPUS cliccare sulla riga del corpus **Prova**, che risulterà evidenziata, poi cliccare su [GESTIONE.. + ELIMINA]: appare un messaggio che chiede conferma della cancellazione; rispondere (Sì) volendo procedere (garantirsi di disporre di un'altra copia del corpus **prova**).

Un ulteriore messaggio conferma l'avvenuta cancellazione. Una volta risposto a quest'ultimo avviso la riga relativa al corpus eliminato viene rimossa dalla griglia.

Nota: Qualora, per qualunque motivo, la cancellazione del corpus non andasse a buon fine, procedere alla rimozione manuale della relativa directory (ad esempio **C:\prova.gat** per il corpus **prova**) e delle relative sottodirectory con Gestione risorse di Windows o altro strumento analogo.

Attivazione dell'ordinamento per sintagmi dei contesti ottenuti nelle ricerche

Ancora dalla finestra APERTURA E GESTIONE GENERALE DEI CORPUS aprire il corpus **Prova** con [APRI], quindi cliccare su [GESTIONE.. + ATTIVA ORDINAMENTO PER SINTAGMI]; durante l'operazione appare un riquadro a fondo giallo nella parte inferiore della finestra e l'operazione è conclusa quando esso scompare.

CAP. 25 - GESTIONE DEI TESTI



Nella parte introduttiva del manuale è stato fatto un cenno al ciclo di elaborazione dei testi. Lo schema mostrato nel disegno ne fornisce il quadro completo (a titolo di esempio si fa riferimento ad un ipotetico testo identificato dalla sigla 'XXX')¹⁰.

Inizialmente il testo, supposto non lemmatizzato, è costituito dal solo file **filgat.XXX.txt**, posto nella *cartella* (o *directory*) *dei testi* del corpus; se il testo contiene un apparato di note sarà presente anche il file **note.XXX.txt**; se vi è anche una traduzione in altra lingua la cartella conterrà anche il file **trad.XXX.txt**. Tutti questi file saranno stati preparati secondo quanto descritto nei capitoli 3 e 4.

Con riferimento allo schema, i file ora citati sono quelli raffigurati schematicamente come *gruppo di documenti* a sinistra in alto, sopra il riquadro **cartella dei testi**.

La fase successiva consiste nell'inserire il testo nel corpus, facendo uso di questi file. L'operazione è stata descritta nel capitolo 8.

Da questo momento GATTO non fa più uso dei file, che però vanno lasciati immodificati nella *cartella dei testi* per rendere possibile una eventuale successiva estrazione del testo lemmatizzato dal corpus.

¹⁰ Tutte le cose dette in questo paragrafo sono ripetute in altre parti del manuale: sono state qui ribadite e raccolte insieme per chiarezza.

Si può ora procedere a lemmatizzare o interrogare il testo. L'interrogazione fornirà in risposta anche i dati di lemmatizzazione inseriti fino a quel momento.

Se si desidera trasferire il testo lemmatizzato in un altro corpus o se si intende apportargli delle modifiche (si intende al testo, non alla lemmatizzazione), l'operazione da eseguire è l'estrazione del testo, completo di lemmatizzazione, dal corpus. L'operazione è descritta nel capitolo 25 e fornisce come risultato, nella sottocartella **out** del corpus, i file mostrati nello schema. Di questi:

- **filgat.XXX.txt** è tratto dall'omonimo file presente nella cartella dei testi (dal quale è stata rimossa l'eventuale lemmatizzazione preesistente), con aggiunta della lemmatizzazione così come risulta al momento;
- **codlemmi.XXX** è un file riservato di GATTO e non deve essere aperto con un editore per non danneggiarlo; contiene i dati di lemmatizzazione;
- **lemmi.XXX.txt** non è mai utilizzato da GATTO e può essere consultato liberamente: contiene in chiaro i dati di lemmi e iperlemmi citati all'interno di **filgat** mediante il loro codici numerici;
- **note.XXX.txt** e **trad.XXX.txt**, se presenti, sono una copia esatta degli omonimi file inseriti nella *cartella dei testi*.

Per trasferire il testo lemmatizzato in un altro corpus basta copiare questi file (escluso **lemmi.XXX.txt**, non usato), nella *cartella dei testi* del corpus di destinazione.

Per modificare il testo, apportare i cambiamenti al **filgat.XXX.txt** appena estratto; naturalmente eventuali modifiche possono coinvolgere anche **note.XXX.txt** o **trad.XXX.txt**. terminate le correzioni, sostituire nella *cartella dei testi* del corpus i file esistenti con i nuovi prelevati dalla sottocartella **out** e modificati. Sostituire tutti i file, quindi anche **codlemmi.XXX**. Per inserire nel corpus il testo corretto procedere infine all'operazione di *sostituzione del testo*, secondo quanto descritto nella Lezione 51.

A questo punto il ciclo di elaborazione è completato e il contenuto della sottocartella **out** può essere cancellato. Naturalmente è possibile ripetere l'intero pacchetto di operazioni tutte le volte che risultino necessarie ulteriori correzioni.

OPERAZIONI PREVISTE E LEZIONI CHE LE DESCRIVONO

Le operazioni previste sui testi di un corpus si possono raggruppare secondo l'elenco che segue. Con esclusione della numero 3 (lemmatizzazione e iperlemmatizzazione) tutte le altre vengono eseguite nell'ambiente **Gestione base dati**.

Operazioni sui testi base		
N.	Lezione	operazione
1	13	controllo di correttezza
2	13	inserimento nel corpus
3	15-24	lemmatizzazione e iperlemmatizzazione
4	49	estrazione dal corpus per modifica o trasferimento in un altro corpus
5	51	reinserimento nel corpus (sostituzione)
6	48	cancellazione dal corpus

Operazioni sui testi associati		
N.	Lezione	operazione
7	14	controllo di correttezza
8	14	verifica dell'allineamento tra testo base e testo associato di tipo Trad
9	14	inserimento nel corpus
10	52	reinserimento (sostituzione) nel corpus
11	53	verifica della presenza dei testi associati previsti
12	48	cancellazione dal corpus

Altre operazioni		
N.	Lezione	operazione
13	53	generazione di copie dei file testo prive dei codici relativi alla lemmatizzazione
14	53	ricerca dei testi del corpus erroneamente non inclusi in bibliografia
15	53	visualizzazione e stampa dei dati bibliografici e statistici di un testo

Operazioni su singoli testi o su gruppi di testi

In linea di massima le operazioni elencate in precedenza possono essere effettuate sia su un solo testo per volta che su più testi. Fanno eccezione la n. 3 in quanto lemmatizzazione e iperlemmatizzazione, già descritte nella parte III, sono effettuate in un altro ambiente, la n. 8 in quanto va seguita passo passo dall'utente, le operazioni di verifica n. 11 e 14 applicate automaticamente al corpus considerato un tutt'uno e la n. 15 che ha lo scopo di fornire informazioni su singoli testi.

Nel caso di operazione eseguita su un singolo testo il programma chiederà all'utente di selezionare la sigla corrispondente nella lista di quelle presenti, avviando l'operazione col pulsante **(OK)**; se l'operazione o le operazioni riguardano un gruppo di testi dovrà invece essere indicato il file contenente la lista delle sigle corrispondenti.

Ogni volta che GATTO esegue una delle operazioni che possono essere compiute su un singolo testo o su un gruppo di testi, come può essere *l'inserimento di testi*, esso esegue all'inizio e alla fine un pacchetto di operazioni di allineamento del corpus che incidono inevitabilmente sulla durata dell'operazione. Il tempo necessario per eseguire le operazioni di allineamento è indipendente dal fatto che si sia agito su pochi testi, su molti o su uno solo.

Da quanto detto consegue che, dovendo eseguire una stessa operazione su N testi, è normalmente preferibile eseguirla in un colpo solo operando sull'intero gruppo di testi, così da eseguire una sola volta, anziché N volte, le operazioni di allineamento iniziali e finali

Operazioni già illustrate in precedenti parti del manuale

Il controllo di correttezza di un testo base e il suo inserimento in un corpus sono stati trattati nella Lezione 13. Le corrispondenti operazioni relative ai testi associati, insieme al controllo di allineamento tra testo base e testo associato di tipo Trad sono stati descritti nella Lezione 14. Lemmatizzazione e iperlemmatizzazione sono stati l'oggetto della III parte del manuale. Per questi argomenti si rimanda pertanto alle parti citate.

NOTA SULLA MODIFICA DEI TESTI LEMMATIZZATI DOPO L'INSERIMENTO IN UN CORPUS

Una volta che un testo è stato inserito in un corpus con l'intenzione di lemmatizzarlo o di modificarne la lemmatizzazione, **non si deve apportare nessuna modifica, anche minima, al corrispondente file filgat.<sigla>.txt posto nella directory dei testi di quel corpus.**

Il motivo deriva dal fatto che il file **filgat.<sigla>.txt** costituisce la "*chiave*" di riconoscimento del testo stesso, che GATTO utilizza per **autorizzarne l'estrazione** dal corpus; l'estrazione tentata sulla base di un file filgat modificato darebbe luogo alla segnalazione di una serie di errori.

Volendo apportare cambiamenti al file filgat di un testo inserito in un corpus e al quale sono state effettuate successivamente modifiche alla lemmatizzazione, attenersi esclusivamente alle procedure descritte nelle Lezioni 49, 50 e 51.

NOTA SULLA MODIFICA DELLA SIGLA DI UN TESTO GIÀ INTRODOTTI NEL CORPUS

Una volta che un testo è stato introdotto in un corpus, l'operazione di modifica della sigla ad esso associata è particolarmente delicata: se eseguita secondo modalità diverse da quelle indicate in questo paragrafo può produrre un danneggiamento anche grave del corpus.

Per procedere alla modifica della sigla si deve procedere come segue.

A. Testo la cui lemmatizzazione non è stata modificata dopo l'inserimento nel corpus. Nell'ordine:

1. cancellare il testo dal corpus (Lezione 48);
2. modificare la sigla nel record corrispondente di bibliografia (Lezione 62);
3. modificare le sigle dei file relativi al testo (filgat e Note, Trad e codlemmi se presenti) nella cartella dei testi;

4. reinserire nel corpus il testo con la sigla modificata (Lezione 13).
- B.** Testo la cui lemmatizzazione è stata modificata dopo l'inserimento nel corpus. Nell'ordine:
1. estrarre il testo dal corpus (Lezione 49);
 2. cancellare il testo dal corpus (Lezione 48);
 3. modificare la sigla nel record corrispondente di bibliografia (Lezione 62);
 4. copiare nella cartella dei testi i file estratti (filgat, e codlemmi se presente), sostituendo quelli preesistenti;
 5. modificare le sigle dei file relativi al testo (filgat e Note, Trad e codlemmi se presenti) così inseriti nella cartella dei testi;
 6. reinserire nel corpus il testo con la sigla modificata (Lezione 13).

ARGOMENTI DELLE LEZIONI DI QUESTO CAPITOLO

L'ordine con cui gli argomenti vengono presentati nelle Lezioni che seguono è determinato principalmente da criteri di comodità espositiva.

Lezione 48: Come cancellare un testo dal corpus, rimuovendone automaticamente anche gli eventuali testi associati.

Lezione 49: Come estrarre un testo lemmatizzato dal corpus per apportarvi modifiche (se il testo che si vuole modificare non è stato lemmatizzato, oppure se la sua lemmatizzazione non è stata modificata dopo che è stato inserito nel corpus questa operazione è inutile: in tal caso le modifiche possono essere apportate direttamente al file filgat già utilizzato a suo tempo per introdurre il testo nel corpus: una volta modificato, tale file potrà essere utilizzato, insieme a codlemmi se il testo è lemmatizzato, per reinserirvelo).

Lezione 50: Come modificare con un editore un file testo già lemmatizzato, senza alterare la lemmatizzazione presente al suo interno (la modifica di un testo non lemmatizzato non richiede invece spiegazioni, in quanto è sufficiente modificarne il file testo con un programma opportuno, utilizzando le stesse avvertenze con le quali era stata effettuata la sua prima stesura).

Lezione 51: Come sostituire un testo attualmente inserito nel corpus con una versione modificata.

Lezione 52: Come sostituire o reinserire un testo associato cui siano state apportate delle modifiche o dopo che sia stato sostituito il corrispondente testo base.

Lezione 53: Come verificare l'effettiva presenza nel corpus dei testi associati previsti nei testi base corrispondenti; come generare delle copie dei file testo prive però dei codici relativi alla lemmatizzazione e iperlemmatizzazione; come verificare l'effettiva presenza in bibliografia dei record relativi a tutti i testi del corpus; come visualizzare e stampare i dati bibliografici e statistici di un qualunque testo del corpus.

Lezione 54: Come effettuare simultaneamente la stessa operazione su un gruppo di testi.

OPZIONI DISPONIBILI NELLE OPERAZIONI SUI TESTI

Alcune di queste opzioni sono state già illustrate nel capitolo 8 in quanto specifiche della fase di inserimento dei testi, base o associati; si tratta di **ins. testi associati**, **uso memoria** e **limita avvisi**, per le quali si rimanda al capitolo suddetto.

Opzione 'testo singolo'

Precisa che l'operazione verrà eseguita su un solo testo; pertanto cliccando su **(OK)** prima di effettuarla verrà presentata una finestra di selezione della sigla corrispondente, da scegliere tra quelle disponibili.

Questa opzione è evidentemente alternativa alla seguente e non prevede la presenza delle opzioni **compatta** e **chiudi GATTO**.

Opzione 'gruppo di testi'

Precisa che l'operazione verrà eseguita su un gruppo di testi; pertanto cliccando su **(OK)** prima di effettuarla verrà presentata una finestra per indicare nome e collocazione del file in cui dovranno essere state introdotte le sigle relative.

Questa opzione è evidentemente alternativa alla precedente e selezionandola si attivano le opzioni **compatta** e **chiudi GATTO**.

Opzione 'compatta'

Opzione presente solo nel caso di operazioni da eseguire su un gruppo di testi. Prevede che al termine venga eseguito automaticamente un compattamento del corpus. Se il tipo di operazioni da effettuare comporta modifiche di rilievo al corpus può essere opportuno farne uso.

Opzione 'chiudi GATTO'

Opzione presente solo nel caso di operazioni da eseguire su un gruppo di testi. Prevede che al termine GATTO venga chiuso automaticamente. Qualora le operazioni programmate richiedano un certo tempo e non si voglia seguirne lo svolgimento di persona questa opzione garantisce la chiusura degli archivi e del programma non appena esse si siano concluse.

FILE COINVOLTI NELLE OPERAZIONI SUI TESTI

Coll'indicazione <sigla> si intende una parte del nome del file che nella realtà viene sostituita dalla sua sigla. Ad esempio nel caso specifico del testo **p48** il file qui indicato genericamente come **filgat.<sigla>.txt** si chiamerà in realtà **filgat.p48.txt**.

filgat.<sigla>.txt

Si tratta del file (*file testo*) che contiene il testo vero e proprio, ovvero il *testo base*. Insieme al testo dell'opera include tutti i codici necessari a GATTO per interpretarne il contenuto. Per poter essere utilizzato nell'inserimento del testo in un corpus questo file si deve trovare nella corrispondente *directory dei testi*. Qualora si proceda all'estrazione del testo dal corpus, ovvero alla generazione di un nuovo file testo (completo di tutte le modifiche eventualmente apportate a lemmi e iperlemmi successivamente all'inserimento nel corpus) quest'ultimo file, anch'esso denominato **filgat.<sigla>.txt**, non verrà collocato nella directory dei testi in luogo di quello esistente ma bensì nella sottodirectory **out** del corpus. Può essere letto e modificato con un editore o un programma di videoscrittura. A questo riguardo, leggere le avvertenze riportate in **Dati e strumenti accessori di cui dotarsi nel capitolo 2**.

codlemmi.<sigla>

File contenente i dati di lemmatizzazione e iperlemmatizzazione del testo. Se il testo non è lemmatizzato questo file non è presente. Per poter essere utilizzato nell'inserimento del testo in un corpus si deve trovare nella corrispondente *directory dei testi*. Qualora si proceda all'estrazione del testo dal corpus, ovvero alla generazione di un nuovo file testo (completo di tutte le modifiche eventualmente apportate a lemmi e iperlemmi successivamente all'inserimento nel corpus) quest'ultimo file, anch'esso denominato **codlemmi.<sigla>**, non verrà collocato nella directory dei testi in luogo di quello esistente ma bensì nella sottodirectory **out** del corpus. Questo file non deve essere aperto con editor o programmi di videoscrittura; qualora lo si facesse inavvertitamente, uscire dal programma senza salvare il file, ma abortendo l'operazione o dichiarando comunque di non registrare le modifiche, al fine di non danneggiarlo irreparabilmente.

lemmi.<sigla>.txt

Non è utilizzato da GATTO. Qualora si proceda all'estrazione del testo dal corpus, ovvero alla generazione di un nuovo file testo (completo di tutte le modifiche eventualmente apportate a lemmi e iperlemmi successivamente all'inserimento nel corpus) quest'ultimo file, anch'esso denominato **lemmi.<sigla>.txt**, non verrà collocato nella directory dei testi in luogo di quello esistente ma bensì nella sottodirectory **out** del corpus. Questo file contiene in chiaro tutti i lemmi e gli iperlemmi utilizzati all'interno del testo (in questo senso è equivalente a **codlemmi.<sigla>**) ed è a disposizione dell'utente che lo può liberamente leggere ed eventualmente modificare con un programma qualunque. Al suo interno ogni lemma e iperlemma è individuato da un numero intero, lo stesso numero riportato, racchiuso tra una coppia di codici "=" all'interno di **filgat.<sigla>.txt**. Gli eventuali iperlemmi presenti nel file ma non richiamati direttamente in **filgat.<sigla>.txt** non sono associati direttamente alle occorrenze nel testo, ma sono concatenati, direttamente o indirettamente, ai lemmi usati: anch'essi verranno trasferiti negli iperlemmi dei corpus nei quali si inserisca il testo in oggetto.

note.<sigla>.txt

Contiene il testo associato di tipo **Note**. Insieme alle note include tutti i codici necessari a GATTO per interpretarne il contenuto. Per poter essere utilizzato nell'inserimento del testo in un corpus si deve trovare nella corrispondente *directory dei testi*. Può essere letto e modificato con un editore o un programma di videoscrittura. A questo riguardo, leggere le avvertenze riportate in **Dati e strumenti accessori di cui dotarsi nel capitolo 2**. Qualora si proceda all'estrazione del testo dal corpus, ovvero alla generazione di un nuovo file testo (completo di tutte le modifiche eventualmente apportate a lemmi e iperlemmi successivamente all'inserimento nel corpus), nella sottodirectory **out** del corpus verrà generata anche una copia del file note.<sigla>.txt.

trad.<sigla>.txt

Contiene il testo associato di tipo **Trad**. Insieme alla traduzione o alla diversa edizione del testo base include tutti i codici necessari a GATTO per interpretarne il contenuto. Per poter essere utilizzato nell'inserimento del testo in un corpus si deve trovare nella corrispondente *directory dei testi*. Può essere letto Dalla finestra modificato con un editore o un programma di videoscrittura. A questo riguardo, leggere le avvertenze riportate in **Dati e strumenti accessori di cui dotarsi nel capitolo 2**. Qualora si proceda all'estrazione del testo dal corpus, ovvero alla generazione di un nuovo file testo (completo di tutte le modifiche eventualmente apportate a lemmi e iperlemmi successivamente all'inserimento nel corpus), nella sottodirectory **out** del corpus verrà generata anche una copia del file trad.<sigla>.txt.

nolemmi.<sigla>.txt

Qualora si decida di estrarre dal corpus una copia del file testo filgat.<sigla>.txt priva però di tutti i codici relativi sia a lemmi e iperlemmi che a note, traduzioni o diverse edizioni associate, nella sottodirectory **out** del corpus verrà generato il file nolemmi.<sigla>.txt, avente i requisiti richiesti. Questo file è a disposizione dell'utente che lo può liberamente leggere ed eventualmente modificare con un programma qualunque, o utilizzare come base per inserire in un corpus il testo in oggetto, privo però di lemmatizzazione (a condizione di riassegnarli il nome filgat.<sigla>.txt)..

err.<sigla>.txt

Viene generato qualora durante il controllo di correttezza o l'inserimento di un testo in un corpus vengano rilevati degli errori. Questo file, generato nella sottodirectory **out** del corpus, conterrà a quel punto le descrizioni degli errori incontrati. Una volta utilizzate le informazioni contenute può essere cancellato liberamente.

<corpus>.log.txt

(<corpus> sta per il nome del corpus). Se vengono segnalati errori nel corso di operazioni eseguite su un gruppo di testi, viene generato questo file nella sottodirectory **out** del corpus. Esso conterrà l'indicazione dei testi affetti da errori e informazioni riguardo a questi ultimi. Nel caso di controllo di correttezza o inserimento di testi in un corpus gli specifici file **err.<sigla>.txt**, generati a loro volta, conterranno i dettagli sugli errori riscontrati per ogni testo. Una volta utilizzate le informazioni contenute può essere cancellato liberamente.

Nella sottodirectory **out** vengono generati automaticamente anche altri file, in funzione delle operazioni eseguite. Si tratta di file temporanei che vengono rimossi, sempre automaticamente, al termine delle operazioni.

Se per qualunque caso (arresto anomalo del programma, interruzione nel funzionamento del calcolatore o altro) tali file rimanessero nella cartella **out** anche riavviando GATTO, significa che non sono più utili e che possono essere cancellati manualmente con Gestione risorse di Windows o altri strumenti analoghi.

LEZIONE 48. COME CANCELLARE UN TESTO DA UN CORPUS

Cancellare un testo dal corpus è molto semplice. È importante però avere ben chiari implicazioni e limiti di questa operazione.

Quando si effettua la cancellazione, questa comporta la rimozione di forme, lemmi e iperlemmi legati solo a quel testo: qualora uno di questi elementi dovesse risultare non presente nei testi rimanenti verrà

rimosso dall'archivio corrispondente, in caso contrario verrà semplicemente ridotto il numero delle sue occorrenze nel corpus.

L'eventuale lemmatizzazione o iperlemmatizzazione del testo operata dopo il suo caricamento nel corpus e dopo la sua ultima eventuale *estrazione* da quest'ultimo andrà ovviamente perduta.

Nessuno dei file contenuti nella directory dei testi e nemmeno nella sottodirectory **out** del corpus viene cancellato, così come non viene toccato il contenuto dell'archivio bibliografico.

Se sono presenti dei testi associati collegati al testo cancellato anch'essi verranno automaticamente rimossi dal corpus.

Si supponga di voler cancellare dal corpus **Prova** creato a suo tempo il testo **es3**, al quale sono associati sia un testo di tipo Note che uno di tipo Trad.

Dall'ambiente **Gestione base dati** aprire il corpus **Prova**; poi dalla finestra GESTIONE BASE DATI cliccare su [TESTI.. + CANCELLAZIONE TESTI DAL CORPUS]; mantenere l'impostazione **testo singolo** e cliccare su (**OK**); nella lista di 3 sigle presentata cliccare su **es3** e procedere con (**OK**); viene presentata la richiesta di conferma dell'operazione: per proseguire scegliere (**Si**); un ulteriore messaggio avverte che l'operazione si è conclusa con la cancellazione del testo **es3**; cliccare su (**OK**). Se ora si prova a ripetere l'operazione con [TESTI.. + CANCELLAZIONE TESTI DAL CORPUS] e successivo (**OK**) si vede che la lista delle sigle relative ai testi inclusi nel corpus si è ridotta a 2, con la scomparsa di **es3**, a conferma dell'avvenuta cancellazione di **es3** da **Prova**. Annullare la nuova operazione con (**Annulla**).

LEZIONE 49. COME ESTRARRE UN TESTO LEMMATIZZATO DAL CORPUS PER MODIFICHE O COPIA IN ALTRO CORPUS

Per apportare modifiche ad un testo si deve operare sul relativo file `filgat.<sigla>.txt`; però l'operazione di per sé non altera anche la copia del testo inserita in un corpus: per cambiare quest'ultima è necessario sostituirla sulla base di una copia corrispondentemente modificata del file `filgat.<sigla>.txt`.

Sono previste due diverse procedure, a seconda che dopo l'inserimento del testo nel corpus siano state apportate o meno modifiche al suo stato di lemmatizzazione; infatti nel primo caso il file `filgat.<sigla>.txt`, o la coppia di file `filgat.<sigla>.txt` più `codlemmi.<sigla>`, presente nella directory dei testi non descrive più esattamente il contenuto del testo, mentre nel secondo tali file sono ancora pienamente validi ed utilizzabili.

L'operazione di *estrazione* di un testo lemmatizzato da un corpus, descritta in questa lezione, serve anche nel caso in cui si voglia trasferire il testo stesso all'interno di un altro corpus.

Caso di un testo la cui lemmatizzazione non è stata modificata dopo l'inserimento nel corpus

In questo caso è inutile procedere all'estrazione e, se necessario, si possono apportare modifiche direttamente su una copia di `filgat.<sigla>.txt`, mediante un editore o un programma di videoscrittura, ferme restando le considerazioni fatte in **Dati e strumenti accessori di cui dotarsi nel capitolo 2.**

Caso di un testo la cui lemmatizzazione è stata modificata dopo l'inserimento nel corpus

Tramite l'operazione di *Estrazione file testo* si generano nella sottodirectory `out` del corpus copie aggiornate di `filgat.<sigla>.txt`, `codlemmi.<sigla>` e `lemmi.<sigla>.txt`. *Aggiornate* significa includenti le modifiche apportate alla lemmatizzazione dopo l'inserimento nel corpus. Anche i file `note.<sigla>.txt` e `trad.<sigla>.txt`, se presenti nella directory dei testi, vengono copiati nella sottodirectory `out`.

Qualora l'estrazione del testo fosse stata motivata dalla necessità di apportarvi delle modifiche, si dovrà procedere ad inserire i cambiamenti corrispondenti nella versione così ottenuta di `filgat.<sigla>.txt`.

In entrambi i casi, una volta apportate le correzioni alla opportuna copia di `filgat.<sigla>.txt`, si sarà pronti per reinserire il testo nel corpus di partenza o per introdurlo in un nuovo corpus.

In questa Lezione verrà visto come effettuare il primo passo, ovvero come estrarre il testo dal corpus.

Si supponga a questo riguardo di voler modificare il testo lemmatizzato **btb**, inserito nel corpus **Demo**.

Dall'ambiente **Gestione base dati** aprire il corpus **Demo**; poi dalla finestra GESTIONE BASE DATI cliccare su [TESTI.. + ESTRAZIONE FILE TESTO + COMPLETI]; mantenere l'impostazione **testo singolo** e cliccare su (**OK**); nella lista di sigle presentata cliccare su **btb** e procedere con (**OK**); viene eseguita l'estrazione del testo e al termine un messaggio avverte che il testo estratto si trova nel file `C:\demo.gat\out\filgat.btb.txt`. In altre parole il nuovo file si trova nella sottodirectory **out** del corpus. In realtà, poiché il testo è lemmatizzato, i file

prodotti saranno 3, ovvero **filgat.btb.txt**, **codlemmi.btb** e **lemmi.btb.txt**; se invece non lo fosse stato sarebbe stato prodotto il solo **filgat.btb.txt**.

In questo secondo caso, se si fosse cioè trattato di un testo non lemmatizzato nemmeno prima di inserirlo nel corpus, l'estrazione sarebbe stata un'operazione inutile, poiché il nuovo filgat.btb.txt sarebbe stato del tutto identico a quello presente nella *directory dei testi*.

Il contenuto dei 3 tipi di file è descritto in **File coinvolti nelle operazioni sui testi** nella parte iniziale di questo stesso capitolo.

Come detto in precedenza per modificare il testo è sufficiente, a questo punto, intervenire con un editore sul file filgat.<sigla>.txt così ottenuto.

Una volta che il file editato e modificato è stato prelevato dalla sottodirectory out e trasferito altrove la copia inclusa in quella sottodirectory può essere cancellata, in quanto non più necessaria.

Ricordarsi che l'operazione di estrazione di un testo da un corpus è accettata solo se nella *directory dei testi del corpus* è presente il file testo del documento in questione.

LEZIONE 50. COME MODIFICARE CON UN EDITORE UN TESTO LEMMATIZZATO

Un testo lemmatizzato, o meglio il file testo corrispondente filgat.<sigla>.txt differisce dal corrispondente file di un testo non lemmatizzato solo perché vi compaiono i codici di lemmatizzazione ed eventualmente di iperlemmatizzazione già presentati nel corso della Lezione 5.

Ad esempio, è possibile incontrare all'interno del file una frase come **...parla de cose multo =47=suctile...** o **...parla de cose multo =47,112=suctile...**. Nel primo caso l'occorrenza della forma **suctile** è stata lemmatizzata col lemma definito nel file col numero **47**, mentre nel secondo a questo si è aggiunta l'associazione all'iperlemma individuato dal numero **112**.

Più in generale le occorrenze lemmatizzate sono precedute da sequenze del tipo **=nnn=** o **=nnn,mmm=** dove **nnn** e **mmm** sono numeri interi, codici dei lemmi e iperlemmi associati alle occorrenze.

Il file lemmi.<sigla>.txt riporta le associazioni tra codici numerici e lemmi o iperlemmi utilizzati all'interno del corrispondente file filgat.<sigla>.txt.

A titolo di esempio, si apra, con Notepad o altro, il file **filgat.nw.txt**, posto nella *directory dei testi* del corpus **Demo**. Si può vedere che molte occorrenze sono precedute da un numero intero racchiuso tra simboli "="; si tratta di occorrenze lemmatizzate e il numero intero identifica il lemma utilizzato. Si può notare che sia la parola **saccio** (seconda riga del riferimento organico individuato da % 1) che la parola **savesse** (prima riga del successivo riferimento organico individuato da % 2) sono precedute da **=6=**. Per sapere di quale lemma si tratti, aprire nella stessa directory il file **lemmi.nw.txt**: nella riga numero **6** si trova il lemma che corrisponde a quel numero, ovvero **sapere (v.)**.

Da quanto detto è facile capire che il file testo di un documento lemmatizzato può essere tranquillamente editato, anche ampiamente, senza con questo danneggiare in alcun modo la lemmatizzazione associata. È soltanto necessario tenere presenti alcune semplici considerazioni:

- un'occorrenza lemmatizzata può essere liberamente spostata in altre zone del testo; se si vuole mantenere la lemmatizzazione nello spostamento l'occorrenza va sempre trattata come un blocco unico col codice numerico che la precede e spostata insieme a questo; se invece la lemmatizzazione non interessa più una volta spostata l'occorrenza si deve cancellare il codice numerico e la coppia di "=" rimasti: infatti, spostando solo l'occorrenza e lasciando dov'era il corrispondente codice di lemma questo verrà associato automaticamente alla parola immediatamente seguente, ovvero a quella successiva alla parola spostata, introducendo ovviamente un errore;
- se non si desidera mantenere una certa lemmatizzazione questa può essere tranquillamente cancellata, rimuovendo il relativo codice **=nnn=** o **=nnn,mmm=**. Se un codice presente nel file lemmi.<sigla>.txt non risulta più usato all'interno di filgat.<sigla>.txt ciò non ingenera errori: semplicemente quel lemma non verrà utilizzato;
- inserendo prima di un'occorrenza un codice di lemma presente in altre parti del testo si produrrà, sia pure in maniera un po' impropria, la lemmatizzazione dell'occorrenza immediatamente successiva con il lemma individuato da quel codice, senza generare errori; ad esempio, se nel caso appena discusso del testo **nw** venisse inserita la codifica **=6=** davanti a un'altra occorrenza qualunque nel file **filgat.nw.txt**, una volta reinserito il testo in un corpus troveremmo anche tale occorrenza lemmatizzata con **sapere (v.)**.

- non introdurre codici di lemma inesistenti in `lemmi.<sigla>.txt`: ciò genererebbe errori in fase di reinserimento del testo: *aggiungere* i corrispondenti codici e lemmi anche nel file `lemmi.<sigla>.txt` non servirebbe a niente in quanto tale file non è utilizzato da GATTO; il file usato in realtà è `codlemmi.<sigla>`, non accessibile con un editore o un programma di videoscrittura.

LEZIONE 51. COME (RE)INSERIRE UN TESTO MODIFICATO, LEMMATIZZATO O MENO

Una volta apportate delle modifiche al file `filgat` relativo ad un testo già incluso in un corpus, perché tali modifiche divengano operative e utilizzabili ai fini delle ricerche è necessario reinserire nel corpus il testo modificato. Se invece si vuole inserire il testo lemmatizzato e modificato in un altro corpus basta seguire la procedura già vista nel capitolo 8, avendo cura che i file necessari (`filgat`, `codlemmi` ed eventualmente `Note` e/o `Trad`) si trovino nella *directory dei testi* del nuovo corpus.

Esistono due possibili procedure per reinserire un testo in un corpus, a prescindere dal fatto che esso sia lemmatizzato o meno e che ad esso siano o meno associati testi di tipo `Note` o `Trad`.

La prima procedura si articola in due fasi, ovvero *cancellazione* del testo e successivo *inserimento*, identico a quello di un nuovo testo e già descritto nella Lezione 13.

La seconda utilizza l'operazione di *sostituzione*, che riunisce cancellazione e successivo inserimento.

Non esistono quasi differenze operative tra le due strade, che sono pertanto pressoché equivalenti.

Nel *sostituire* un testo del corpus vanno tenute presenti alcune considerazioni:

- la cancellazione della precedente versione viene effettuata solo se il controllo di correttezza preliminarmente eseguito sulla nuova ha dato esito favorevole,
- se alla versione precedente erano associati testi di tipo `Note` o `Trad` questi verranno rimossi insieme al testo base; se anche la nuova versione li include potranno essere inseriti aggiungendoli manualmente in un secondo tempo mediante l'uso della procedura descritta nella Lezione 14 e ripresa nella Lezione 52, oppure in maniera automatica, contestualmente alla sostituzione, attivando l'opzione **ins. testi associati**,
- se il testo è lemmatizzato, come nel caso dell'inserimento deve essere obbligatoriamente presente nella *directory dei testi* anche il file `codlemmi.<sigla>`.

A titolo di esempio si procederà (immaginando di averlo modificato) a reinserire due volte il testo **r3** all'interno del corpus **Demo**, la prima volta eseguendo l'operazione in due tempi e la seconda volta in un unico passaggio. **r3** è caratterizzato dall'essere lemmatizzato e dall'avere entrambi i tipi di testo associato previsti, ossia `Note` e `Trad`.

Esempio 1: reinserimento del testo **r3** nel corpus **Demo** tramite cancellazione e successivo reinserimento.

Dall'ambiente **Gestione base dati** aprire il corpus **Demo**; poi dalla finestra `GESTIONE BASE DATI` cliccare su [`TESTI.. + CANCELLAZIONE TESTI DAL CORPUS`]; mantenere l'impostazione **testo singolo** e cliccare su (**OK**); nella lista di sigle presentata cliccare su **r3** e procedere con (**OK**); viene chiesta conferma della cancellazione; per procedere rispondere con (**Si**); al termine un messaggio darà conferma dell'avvenuta cancellazione. Anche i testi associati sono stati rimossi. Per verificare che la cancellazione ha avuto luogo si può, ad esempio, cliccare su [`TESTI.. + VISUALIZZAZIONE DATI BIBLIOGRAFICI E STAT.`]: tra le sigle dei testi presenti nel corpus **r3** non compare più.

Sempre dalla finestra di `GESTIONE BASE DATI` cliccare su [`TESTI.. + INSERIMENTO TESTI`]; mantenere le opzioni **testo singolo**, **uso memoria** e **limita avvisi** e attivare **ins. testi associati**. Procedere con (**OK**). Selezionare **r3** e scegliere (**OK**). Un primo messaggio avvertirà che è stato inserito nel corpus il testo **r3**; un secondo messaggio informerà che è stato caricato anche il testo **Note** associato a **r3**; un terzo messaggio darà lo stesso avvertimento relativo al testo associato **Trad**; in tutti i casi rispondere con (**OK**). Anche in questo caso si può avere conferma del fatto che l'operazione è andata a buon fine tramite [`TESTI.. + VISUALIZZAZIONE DATI BIBLIOGRAFICI E STAT.`]: questa volta **r3** compare di nuovo nella lista dei testi del corpus **Demo**.

Esempio 2: reinserimento del testo **r3** nel corpus **Demo** tramite sostituzione.

Dalla finestra `GESTIONE BASE DATI` cliccare su [`TESTI.. + SOSTITUZIONE TESTI`]; mantenere le impostazioni **testo singolo** e **limita avvisi** e attivare, se non lo è già, **ins. testi associati** (l'assenza dell'opzione **uso memoria** è l'unica differenza rispetto al caso precedente); cliccare su (**OK**); nella lista di sigle presentata selezionare **r3** e procedere con (**OK**); viene chiesta conferma della sostituzione; successivi

messaggi confermeranno l'avvenuto reinserimento del testo base, del testo **Note** e del testo **Trad** associati a **r3**; in tutti i casi rispondere con **(OK)**.

LEZIONE 52. COME MODIFICARE E REINSERIRE UN TESTO ASSOCIATO

Il meccanismo di modifica e reinserimento di un testo associato, di tipo Note o Trad, è molto più semplice rispetto a quello di un testo base.

Innanzitutto per questi testi non sono previste forme di lemmatizzazione, quindi volendoli modificare non c'è necessità di estrarli dal corpus con operazioni corrispondenti a quella descritta nella Lezione 49 in relazione ai testi base. Le modifiche si possono apportare direttamente ai file `note.<sigla>.txt` e `trad.<sigla>.txt` collocati nella *directory dei testi*.

Il secondo elemento di semplificazione consiste nel fatto che il reinserimento di una nuova versione di un testo associato già inserito in un corpus non richiede la preventiva cancellazione della versione precedente: infatti l'inserimento di un testo associato comporta l'automatica cancellazione di una eventuale versione precedente. A dimostrazione di questo sta il fatto che si usa lo stesso menu per l'inserimento e la sostituzione dei testi associati.

A titolo di esempio, si immagini di aver introdotto qualche modifica nel file **trad.r3.txt** associato al testo **r3** del corpus **Demo**. La *precedente versione* di **trad.r3.txt** è correntemente inserita nel corpus e la si voglia sostituire con la *nuova versione*.

Dalla finestra GESTIONE BASE DATI cliccare su [TESTI.. + INSERIM./SOSTITUZ. TESTI ASSOCIATI + TIPO TRAD]; mantenere le impostazioni **testo singolo** e **limita avvisi** e procedere con **(OK)**; compare come possibile scelta solo **r3**, unico testo avente un file di tipo Trad nella *directory dei testi*; selezionarlo e procedere con **(OK)**; viene chiesta conferma della sostituzione; rispondendo **(Sì)** la sostituzione ha luogo e un successivo messaggio comunica la positiva conclusione dell'operazione.

LEZIONE 53. OPERAZIONI DI NATURA ACCESSORIA ESEGUIBILI SUI TESTI

In GATTO sono previste alcune altre operazioni sui testi, di natura accessoria rispetto a quelle illustrate finora; esse sono:

1. **verifica della presenza nel corpus dei testi associati previsti,**
2. **verifica della presenza in bibliografia dei record relativi a tutti i testi del corpus,**
3. **estrazione di file testo privi dei codici relativi a lemmi e iperlemmi,**
4. **visualizzazione e stampa dei dati bibliografici e statistici di un testo.**

VERIFICA DELLA PRESENZA NEL CORPUS DEI TESTI ASSOCIATI PREVISTI

Quando ad un testo base va associato un testo di tipo Note o di tipo Trad, o entrambi, il file `filgat.<sigla>.txt` contiene come prima riga un codice il quale avverte che è prevista l'esistenza dell'uno, dell'altro o di entrambi i tipi di testi associati (vedere a questo riguardo il capitolo 4). Questo codice avverte l'interprete che i successivi caratteri diacritici inseriti nel testo e relativi ai testi associati, che in altri casi verrebbero rifiutati come erronei, vanno considerati codici validi.

Quando si procede all'inserimento del testo base nel corpus può capitare di dimenticarsi di attivare l'opzione **ins. testi associati** o di decidere di inserire i testi associati in un secondo tempo. Se si passa ad interrogare un testo privo dei previsti testi associati, GATTO non è in grado di accorgersi in partenza di questa mancanza e segnalerà, tramite un messaggio di errore, l'assenza di eventuali note o di brani associati in corrispondenza dei contesti che li prevederebbero, ma solo al momento in cui l'utente tentasse di visualizzarli.

Per prevenire queste situazioni, una volta completato il caricamento di uno o più testi nel corpus è possibile utilizzare un comando per verificare l'effettiva presenza di tutti i testi associati previsti nel corpus e segnalare quelli che risultino assenti.

A titolo di esempio il testo **r3** del corpus **Demo**, cui sono associati sia un testo Note che un testo Trad, verrà cancellato e reinserito, tramite l'operazione di *sostituzione*, non attivando però l'opzione **ins. testi associati**. In questo modo durante la fase di cancellazione del testo esistente verranno rimossi dal corpus anche i due file associati, mentre durante la successiva fase di inserimento verrà reintrodotta nel corpus il

solo testo base. Al termine si sarà ottenuta una situazione di errore potenziale, in quanto **r3** prevede la presenza di testi associati che in realtà a quel punto mancano.

Dalla finestra GESTIONE BASE DATI cliccare su [TESTI.. + SOSTITUZIONE TESTI]; mantenere le impostazioni **testo singolo** e **limita avvisi** e disattivare, se fosse attiva, l'opzione **ins. testi associati**; cliccare su **(OK)**; nella lista di sigle presentata selezionare **r3** e procedere con **(OK)**; viene chiesta conferma della sostituzione; un secondo messaggio informerà che la sostituzione ha avuto luogo regolarmente.

Se a questo punto si passasse all'ambiente **Ricerche**, magari definendo per comodità un sottocorpus formato dal solo testo **r3**, e si cercassero occorrenze di forme qualunque, cliccando sul simbolo **A** presente nelle testate dei contesti ottenuti si riceverebbero in risposta messaggi di errore derivanti dall'assenza del testo associato di tipo Trad.

Volendo sincerarsi dell'eventuale presenza di problemi di questo genere prima di procedere con le ricerche, sempre dalla finestra GESTIONE BASE DATI cliccare su [TESTI.. + VERIFICA PRESENZA TESTI ASSOCIATI]; un messaggio segnala, come prevedibile, che sono state riscontrate situazioni di errore (cioè file associati mancanti), descritte nel file di nome fisso **C:\demo.gat\out\demolog.txt.**; cliccare su **(OK)** per procedere.

Aprendo con un editore il file suddetto si potrà vedere che contiene i due avvisi seguenti:

*il testo r3 prevede un testo associato di tipo 'note' non inserito
il testo r3 prevede un testo associato di tipo 'trad' non inserito*

che avvertono dell'assenza dei due file associati a **r3**.

Il modo più veloce per ripristinare l'integrità del corpus **Demo** è quello di ripetere la *sostituzione* del testo **r3**, questa volta stando però attenti ad attivare l'opzione **ins. testi associati**.

Facendolo e ripetendo subito dopo l'operazione attivata dal menu [TESTI.. + VERIFICA PRESENZA TESTI ASSOCIATI] si ottiene il messaggio **nessuna situazione irregolare rilevata**, a conferma che ora i testi associati sono nuovamente tutti presenti.

VERIFICA DELLA PRESENZA IN BIBLIOGRAFIA DEI RECORD RELATIVI A TUTTI I TESTI DEL CORPUS

Un'altra fonte di errori durante l'interrogazione di un corpus può derivare dalla cancellazione dall'archivio bibliografico, magari per errore, di uno o più record relativi a testi inclusi nel corpus. Gli errori scattano al momento di localizzare delle occorrenze all'interno di quei testi in quanto GATTO non può disporre di alcuni dati necessari, dati collocati per l'appunto nell'archivio bibliografico. Questo tipo di situazione è di per sé più difficile da rilevare e può derivare per esempio dalla cancellazione di un certo testo sia da un corpus che dal relativo archivio bibliografico, magari dimenticando che lo stesso testo è ancora incluso in un altro corpus che fa riferimento alla stessa bibliografia.

Anche in questo caso è disponibile un'operazione di verifica che permette di evidenziare il problema prima che insorga. Per sperimentarla va preventivamente introdotta una situazione di errore nella bibliografia.

Dalla finestra GESTIONE BASE DATI cliccare su [BIBLIOGRAFIE.. + APRI] e selezionare l'archivio bibliografico **C:\demobiblio\biblio.mdb**.

Cliccare in un punto della riga relativa al testo **IN** per selezionarla, quindi cliccare su [MODIF. DATI.. + MODIFICA]; nella casella **sigla** sostituire **IN** con **INN** e registrare la modifica con [REGISTRA] e [CHIUDI FINESTRA]. Riaprire il corpus **Demo** (l'accesso in scrittura agli archivi bibliografici chiude sempre l'eventuale corpus in uso) e dalla finestra GESTIONE BASE DATI cliccare su [TESTI.. + RICERCA TESTI MANCANTI IN BIBLIOGRAFIA]. Si ottiene in risposta il messaggio:

i seguenti testi non sono presenti nella bibliografia: IN

I testi sono individuati tramite la sigla: nel corpus è presente il testo caratterizzato dalla sigla **IN**, mentre lo stesso testo non compare più nella bibliografia (la contemporanea presenza del testo con sigla **INN** è ovviamente irrilevante).

Ripristinare il corretto contenuto della bibliografia ripetendo al contrario la precedente operazione di modifica nell'archivio bibliografico. Se ora si ripete il controllo [TESTI.. + RICERCA TESTI MANCANTI IN BIBLIOGRAFIA] si ottiene in risposta

tutti i testi del corpus sono presenti nella bibliografia.

ESTRAZIONE DI FILE TESTO PRIVI DEI CODICI RELATIVI AI LEMMI E AGLI IPERLEMMI

Talvolta può essere utile disporre dei file testo completi delle altre codifiche ma privi dei codici relativi a lemmi e iperlemmi e di quelli relativi ai testi associati.

Una delle operazioni previste nella gestione dei testi consente di ottenere esattamente questo risultato.

Come esempio, si voglia ottenere questa speciale forma del file testo relativamente al testo lemmatizzato **p48**.

Dalla finestra GESTIONE BASE DATI cliccare su [TESTI.. + ESTRAZIONE FILE TESTO + SENZA CODICI DI LEMMA]; mantenere **testo singolo** e cliccare su **(OK)**; selezionare **p48** e procedere con **(OK)**; l'operazione viene eseguita e al termine un messaggio avverte che è stato creato il file **C:\demo.gat\out\nolemmi.p48.txt**.

Se si apre questo file con un editore e lo si paragona con **filgat.p48.txt** si può notare che il testo è identico, salvo che sono stati soppressi tutti i codici di lemma.

VISUALIZZAZIONE E STAMPA DEI DATI BIBLIOGRAFICI E STATISTICI DI UN TESTO

Questa funzione è disponibile in tutti gli ambienti di GATTO, quindi lo è a maggior ragione nell'ambito delle operazioni relative alla gestione dei testi.

Per attivare la finestra contenente i dati bibliografici e statistici di un testo è sufficiente, dalla finestra GESTIONE BASE DATI, cliccare su [TESTI.. + VISUALIZZAZIONE DATI BIBLIOGRAFICI E STAT.], selezionare la sigla del testo voluto e procedere con **(OK)**. Cliccando l'icona presente in alto a destra che rappresenta una stampante è possibile stampare i dati contenuti nella finestra su carta o su file.

Uscendo dalla finestra viene offerta la possibilità di selezionare immediatamente un altro testo per vederne i dati, oppure di tornare definitivamente alla finestra GESTIONE BASE DATI.

LEZIONE 54. COME EFFETTUARE UNA STESSA OPERAZIONE SU PIÙ TESTI

In certi casi è necessario effettuare la stessa operazione su più testi; un caso tipico è quello della costruzione di un corpus. Niente impedisce di effettuare l'operazione caricando un testo per volta, però è senz'altro più conveniente poterla fare con un unico comando.

Come accennato precedentemente il meccanismo si basa sulla preventiva preparazione di un file di tipo **txt** nel quale vanno inserite le sigle dei testi cui si intende applicare l'operazione; i testi verranno elaborati nell'ordine in cui le sigle compaiono nel file.

Per sperimentare questo meccanismo si supponga di voler costruire una copia del corpus **Demo** a partire dai file testo con i quali è stato costruito a suo tempo. Il nuovo corpus si chiamerà **Demobis** e, per semplicità, utilizzerà lo stesso archivio bibliografico e la stessa *directory dei testi* di **Demo**.

Prima di tutto, con le tecniche già discusse, creare il nuovo corpus, nel disco che si vuole (**C**, **D**,...) attribuendogli il nome **Demobis** e assegnandogli archivio bibliografico (probabilmente **C:\Demobiblio\biblio.mdb**) e *directory dei testi* (probabilmente **C:\demotesti**) di **Demo** (per fare questo basta seguire la procedura descritta nella Lezione 11).

Nella *directory dei testi* di **Demobis** è presente il file **Listdemo.txt** che contiene, una per riga, le sigle dei testi 12 costituenti il corpus **Demo**, da **btb** a **vad**; per verificarne il contenuto aprirlo ad esempio con Notepad, senza modificarne il contenuto.

A questo punto dalla finestra GESTIONE BASE DATI aprire il nuovo corpus **Demobis**, poi selezionare [TESTI.. + INSERIMENTO TESTI], attivare l'*option box* **gruppo di testi** e il *check box* **ins. testi associati**, lasciando attivi **uso memoria** e **limita avvisi**; attivare anche **compatta** e, per prova, **chiudi GATTO**; per quanto riguarda **ins. testi associati** va tenuto presente che le assegnazioni delle opzioni sono necessariamente uguali per tutti i testi cui si applica l'operazione e vanno impostate in quest'ottica. Procedere con **(OK)**. Selezionare il file **Listademo.txt** nella *directory dei testi* di **Demo** e avviare definitivamente l'operazione con **(Apri)**. I testi vengono caricati in sequenza; al termine il corpus viene compattato e GATTO chiuso.

Il risultato dell'operazione è la costruzione di un corpus in tutto e per tutto uguale a **Demo**.

Durante l'inserimento dei testi compare il menu [INTERROMPI]: utilizzandolo GATTO completerà l'operazione in corso fermandosi prima di passare al testo successivo.

CAP. 26 - GESTIONE DEI LEMMI

Come è stato spiegato nel capitolo 9 tutti i meccanismi legati all'inserimento e alla rimozione di lemmi dal lemmario di un corpus hanno luogo in maniera automatica. La prima volta che si fa uso di un nuovo lemma, lemmatizzando un'occorrenza di una forma del corpus, il lemma, corredato dei suoi attributi, ovvero categoria grammaticale ed eventualmente disambiguatore e commento, viene inserito nel lemmario. Viceversa quando viene rimossa l'ultima lemmatizzazione presente nel corpus nella quale si fa uso di un certo lemma (per attribuzione di un altro lemma a tale occorrenza, per rimozione pura e semplice di tale lemmatizzazione, per eliminazione di quel testo dal corpus), quest'ultimo viene cancellato dal lemmario. Non è previsto alcuno strumento che permetta all'utente di inserirsi in questo meccanismo di gestione automatica del lemmario.

Peraltro va considerata la possibilità che in fase di inserimento degli attributi di un lemma (entrata lessicale, categoria grammaticale, ecc.) venga commesso un errore. Quest'ultimo può sfuggire anche in seguito e il lemma errato può venire utilizzato per lemmatizzare occorrenze di forme diverse in testi diversi.

Analogamente si possono commettere errori nell'associare un lemma ad una forma, o meglio a diverse occorrenze di una certa forma, magari per un'errata interpretazione del significato di quest'ultima. Anche in questo caso ci si può trovare di fronte, nel tempo, ad una serie di errori interessanti forme e testi diversi, la cui correzione tramite sostituzione ad una ad una delle lemmatizzazioni errate con altre corrette potrebbe rivelarsi un compito assai pesante.

GATTO contiene tre funzioni di correzione di questi tipi di errori, le quali consentono di segnalare al programma quali sono gli elementi errati e qual'è la correzione da applicare, dopo di che l'applicazione di quest'ultima a tutte le occorrenze interessate viene effettuata automaticamente.

Le funzioni in questione sono:

- **modifica dei dati di un lemma,**
- **fusione di due lemmi,**
- **sostituzione con altro lemma di un certo lemma associato a più occorrenze di una stessa forma.**

Inoltre si può desiderare di rimuovere dal corpus i lemmi *muti* o le coppie forma-lemma *mute* presenti e magari non più necessari; si tratta, come spiegato nel capitolo 13, di elementi cui non corrispondono specifiche occorrenze lemmatizzate all'interno del corpus.

L'operazione corrispondente è indicata come:

- **rimozione di lemmi e coppie forma-lemma muti**

Modifica dei dati di un lemma

Consente la correzione di tutte le caratteristiche attribuite ad un lemma, ovvero entrata lessicale, categoria grammaticale, disambiguatore e commento. È ovviamente utile quando ci si rende conto che uno o più dei dati suddetti risultano errati. Tutte le occorrenze lemmatizzate col lemma in questione risulteranno automaticamente lemmatizzate col lemma corretto.

Qualora le modifiche portassero il lemma a coincidere con uno già esistente la correzione verrebbe rifiutata (in questo caso utilizzare la funzione seguente).

Fusione di due lemmi

Può capitare, ad esempio, di generare nel lemmario due lemmi coincidenti dal punto di vista semantico ma diversi per aspetti puramente formali, come potrebbero essere **uomo (s.m.)** e **uomo (s.m)**, in cui il secondo lemma differisce dal primo per l'assenza del punto dopo la **m**. Il secondo lemma potrebbe essere stato digitato perché l'utente non si è accorto della presenza del primo, dopo di che l'introduzione dell'errore ha fatto il resto, inducendo GATTO ad accettare e inserire nel lemmario anche il secondo lemma. Nel corso del tempo, se questo errore non viene rivelato, può capitare di lemmatizzare alcuni gruppi di occorrenze facendo uso del primo lemma e altri facendo uso del secondo.

Per rimediare a questa situazione è possibile specificare quali sono i due lemmi coinvolti, precisando quale è quello corretto e quale quello errato; GATTO provvederà a far *convergere* sul lemma corretto tutte le occorrenze lemmatizzate con l'altro, rimuovendo quest'ultimo dal lemmario.

Se i lemmi errati fossero due o più, basterà ripetere l'operazione più volte, facendo convergere sul lemma corretto ogni volta un lemma diverso.

Se nessuno dei lemmi da *fondere* risultasse esatto, basta correggerne uno mediante la funzione precedente (*Modifica dei dati di un lemma*) e poi procedere a far convergere su di lui gli altri.

Sostituzione con un altro di un certo lemma associato a più occorrenze di una stessa forma

Può darsi il caso che un lemma venga utilizzato per lemmatizzare un certo numero di occorrenze di una certa forma (se il fatto si fosse verificato per una sola occorrenza basterebbe modificare quella specifica lemmatizzazione utilizzando l'ambiente omonimo). Supponiamo che in un secondo tempo ci si renda conto che il lemma da utilizzare sarebbe stato un altro, presente nel lemmario. Per risolvere il problema esiste in GATTO una terza funzione che, a partire dall'indicazione della coppia forma - lemma errata e del lemma che si sarebbe invece dovuto utilizzare, provvede a scambiare il lemma errato con quello giusto in tutte le occorrenze della suddetta coppia forma - lemma (se il nuovo lemma non esistesse nel lemmario basterebbe associarlo ad una singola occorrenza così da generarlo e inserirlo nel lemmario, usandolo poi come nuovo lemma nell'operazione di correzione qui descritta).

LEZIONE 55. COME MODIFICARE I DATI DI UN LEMMA

In **Demo** esistono alcuni lemmi costituiti dalla combinazione di più parole separate, quali **avantiché**, **innanziché**, **giassiaticché**; a tali lemmi è stata attribuita la categoria grammaticale *congiunzioni separate*, abbreviata nel lemmario in **cong. s.**. In questi casi tra le due abbreviazioni costituenti la categoria grammaticale viene generalmente inserito uno spazio per chiarezza di lettura; questo è stato fatto per 8 dei 9 lemmi di questo tipo presenti nel lemmario di **Demo**. Per errore al lemma **sicundukì** è stata associata la categoria grammaticale **cong.s.** digitata senza spazio; come detto in precedenza questo tipo di errore, di per sé di ben scarsa rilevanza, potrebbe generare nel tempo una indesiderata duplicazione del lemma nel lemmario e per questo motivo può essere opportuno rimuoverlo.

A questo scopo, dalla finestra GESTIONE BASE DATI scegliere [**LEMMARIO.. + CORREZIONE LEMMI**]. La finestra che si apre presenta nella parte superiore una griglia che mostra il contenuto del lemmario e in basso un riquadro utilizzabile per modificare i dati del lemma selezionato.

Per visualizzare **sicundukì (cong.s.)** si può scorrere il lemmario usando la barra di scorrimento verticale oppure digitarne le lettere iniziali nella casella a sfondo bianco posta sulla sinistra immediatamente al di sotto della griglia. Seguendo questa seconda strada, non appena digitato in tale casella **sic**, **sicundukì** appare nella griglia. Per avviare la modifica selezionare il lemma voluto cliccando su un punto della sua riga, che come risultato apparirà evidenziata (sfondo blu); volendo conoscere le forme lemmatizzate con quel lemma cliccare su un punto qualunque della griglia col tasto destro: appare una finestra più piccola contenente la lista delle forme associate a **sicundukì**, lista formata dall'unica forma **sicundu**. Chiudere la finestra e, per impostare la modifica, usare [**ESTRAI LEMMA**] (se cliccando su [**ESTRAI LEMMA**] si copia per errore un lemma sbagliato, basta ripetere l'operazione con quello giusto: i dati del nuovo lemma sostituiranno i precedenti nel riquadro **Dati da modificare**). Questa operazione copia tutti i dati del lemma nel riquadro **Dati da modificare**; nella casella contenente **cong.s.** inserire uno spazio dopo il primo punto e salvare la modifica con [**REGISTRA MODIFICHE**] (o annullarla con [**ANNULLA**]). L'operazione di modifica si è conclusa e da questo momento tutte le occorrenze interessate risulteranno lemmatizzate con **sicundukì (cong. s.)**, completo di spazio separatore.

La modifica apportata non ha introdotto cambiamenti nell'ordinamento alfabetico dei lemmi nel lemmario; se invece, per ipotesi, la correzione fosse consistita nel modificare **sicundukì** in **secundukì** il risultato avrebbe implicato uno spostamento del lemma modificato. Per tenerne conto nella griglia si sarebbe dovuta effettuare una riscrittura di quest'ultima, con automatica rappresentazione del lemmario a partire dai suoi primi elementi e perdita del precedente posizionamento sugli elementi inizianti per **sic**. Talvolta può essere comodo operare le modifiche al lemmario in serie, agendo magari su più lemmi alfabeticamente vicini tra di loro: in tal caso questa sorta di rimescolamento della visualizzazione dopo ogni correzione potrebbe risultare fastidiosa.

Per aggirare il problema GATTO non effettua automaticamente il riordinamento della griglia, anche se così il lemma modificato si presenta fuori posizione. Provare ad esempio a selezionare di nuovo **sicundukì** e a copiarlo nel riquadro **Dati da modificare** tramite [**ESTRAI LEMMA**], poi modificarlo effettivamente in **secundukì** e registrare la modifica con [**REGISTRA MODIFICHE**]: come si può vedere il lemma rimane dov'è, anche se il posizionamento alfabetico di **secundukì** tra **sicunduchì** e **sicuramente** è chiaramente errato; in questo modo si può però procedere ad operare eventuali modifiche sui lemmi circostanti senza difficoltà. Se invece si vuole reimpostare la griglia per tener conto della modifica effettuata usare [**ORDINA**]: dopo qualche

secondo la griglia viene riposizionata all'inizio del lemmario ma, se ci si porta nella zona dei lemmi iniziati per **se**, si troverà **secundukì (cong. s.)** correttamente posizionato tra **secundukì (cong.)** e **sedere (v.)**. Deve comunque essere chiaro che queste considerazioni sull'ordinamento riguardano solo la griglia di visualizzazione, mentre il lemmario vero e proprio, inteso come archivio di dati, è sempre internamente ordinato senza bisogno di ricorrere al pulsante [ORDINA] o ad altri comandi.

Se non vi sono ulteriori modifiche o correzioni da apportare si può chiudere la finestra con [ESCI].

LEZIONE 56. COME FONDERE DUE LEMMI INTRODOTTI ERRONEAMENTE COME DIVERSI

La seconda situazione di errore precedentemente descritta in questo capitolo è ben esemplificata da un altro errore presente nel lemmario di **Demo**, ovvero dalla compresenza dei lemmi **alcuno (indef.)** e **alcuno (iindef.)**, il secondo dei quali originato da un evidente errore di battitura nella categoria grammaticale.

Per rimuoverlo, dalla finestra GESTIONE BASE DATI scegliere [LEMMARIO.. + FUSIONE DI 2 LEMMI IN 1]. La finestra che appare è molto simile alla precedente, salvo che presenta in basso due riquadri nei quali copiare i lemmi coinvolti nella fusione.

Per cominciare, digitare **alc** nella casella a sfondo bianco per localizzare più rapidamente i due lemmi. Cliccare su un punto della riga del primo, ovvero quello con la categoria grammaticale errata **iindef.** (la riga apparirà evidenziata) e poi cliccare col tasto destro in un punto qualunque della griglia: apparirà una finestra la quale mostrerà che l'unica forma associata a questo lemma è **alchun**; ripetendo l'operazione per il lemma corretto si vedrà che in questo caso le forme associate sono decisamente più numerose.

Notare ora che i due riquadri in basso sono denominati **Lemma da mantenere** e **Lemma da sopprimere**: queste due diciture chiariscono quale è il lemma che va copiato nell'uno e quale nell'altro al fine di effettuare la fusione nella direzione corretta, ovvero quella di far sparire **alcuno (iindef.)** dal lemmario convertendo tutte le occorrenze, di qualunque forma e in qualunque testo, lemmatizzate con tale lemma in occorrenze lemmatizzate con **alcuno (indef.)**.

I due riquadri sono anche caratterizzati dalla presenza di due diversi segnali stradali (*direzione obbligata*, a sfondo giallo e *divieto di accesso*): cliccando su uno dei due i segnali stradali si invertono: il senso è quello di indicare in ogni momento in quale dei due riquadri verrà copiato il lemma selezionato facendo clic su [ESTRAI LEMMA].

In pratica, per effettuare la correzione dell'esempio:

1. cliccare su **alcuno (indef.)**, ovvero sul lemma corretto, per selezionarlo (si poteva indifferentemente avviare l'operazione partendo dal lemma errato),
2. controllare che il cartello stradale con la freccia e il fondo giallo si trovi nel riquadro **Lemma da mantenere** (se non è così cliccare su uno dei due segnali stradali per scambiarli),
3. cliccare su [ESTRAI LEMMA] per copiare i dati del lemma corretto nel riquadro **Lemma da mantenere** (se cliccando su [ESTRAI LEMMA] si copia per errore un lemma sbagliato, basta ripetere l'operazione con quello giusto: i dati del nuovo lemma sostituiranno i precedenti nel riquadro di arrivo),
4. selezionare nella griglia il lemma errato, ovvero **alcuno (iindef.)**, cliccandovi sopra,
5. scambiare i segnali stradali in modo che la freccia a sfondo giallo si trovi in **Lemma da sopprimere**,
6. riutilizzare [ESTRAI LEMMA] per estrarre anche i dati del secondo lemma,
7. verificare di avere **alcuno (indef.)** come **Lemma da mantenere** e **alcuno (iindef.)** come **Lemma da sopprimere** (in caso contrario ricominciare daccapo),
8. effettuare l'operazione con [FONDI I DUE LEMMI] (o annullarla con [ANNULLA]).

Questa volta la griglia viene riscritta automaticamente e si può vedere che il lemma errato non è più presente. Anche le occorrenze lemmatizzate con **alcuno (iindef.)** sono ora lemmatizzate con **alcuno (indef.)**, in tutto il corpus.

Se i lemmi errati, da far convergere su uno stesso lemma, fossero stati più di uno, sarebbe bastato ripetere l'intera operazione per ognuno di essi, definendo ogni volta come **Lemma da mantenere** il solito lemma corretto e come **Lemma da sopprimere**, di volta in volta, uno di quelli errati. Per accelerare l'operazione si sarebbe potuto attivare (se non era già attivo) il *checkbox* **congela il lemma da mantenere** nel corso della prima fusione, prima di usare il pulsante [FONDI I DUE LEMMI]: come risultato il contenuto del riquadro **Lemma da mantenere** non sarebbe stato cancellato al termine dell'operazione, permettendo di ridefinire ogni volta solo il **Lemma da sopprimere**.

Al termine usare [ESCI] per chiudere la finestra.

LEZIONE 57. COME SOSTITUIRE CON ALTRO LEMMA UN CERTO LEMMA ASSOCIATO A PIÙ OCCORRENZE DI UNA STESSA FORMA

Per mostrare con un esempio concreto questo tipo di intervento di modifica del lemmario si consideri, nel corpus **Demo**, il lemma **acconciare(-si) (v.)**, che è stato utilizzato tra l'altro per lemmatizzare occorrenze della forma **acconcia**; si supponga di voler sostituire in queste ultime lemmatizzazioni **acconciare(-si) (v.)** con **acconciare (v.)**¹¹. Non ha alcuna importanza se già esistono o meno altre occorrenze di **acconcia** lemmatizzate con **acconciare**; importante è invece che questo nuovo lemma esista nel lemmario (come già detto, in caso contrario basterebbe usarlo per lemmatizzare un'occorrenza qualunque, magari proprio di **acconcia**, per inserirlo).

Dalla finestra GESTIONE BASE DATI scegliere [LEMMARIO.. + SOSTITUZIONE LEMMI IN COPPIE FORMA-LEMMA]. La nuova finestra è quasi identica a quella usata nella **Fusione di 2 lemmi in 1**; le differenze consistono in alcune diciture diverse, nella presenza del bottone [LISTA FORME] e della casella **forma** nel riquadro **Coppia forma-lemma da sciogliere**.

Anche in questo caso si selezionano i lemmi cliccando sulle righe relative, si possono vedere le forme associate al lemma selezionato cliccando col tasto destro su un punto qualunque della griglia, si può restringere il contenuto di quest'ultima facendo uso della casella a sfondo bianco posta subito al di sotto di essa.

Per effettuare la sostituzione dell'esempio, procedere come segue:

1. visualizzare in griglia **acconciare** e **acconciare(-si)**, eventualmente utilizzando la casella a sfondo bianco posta sotto di essa,
2. cliccare su **acconciare(-si) (v.)**, ovvero sul lemma da eliminare dalle lemmatizzazioni di **acconcia**, per selezionarlo (si poteva indifferentemente avviare l'operazione partendo dall'altro lemma),
3. controllare che il cartello stradale con la freccia e il fondo giallo si trovi nel riquadro **Coppia forma-lemma da sciogliere** (se non è così cliccare su uno dei due segnali stradali per scambiarli),
4. cliccare su [ESTRAI LEMMA] per copiare i dati del lemma nel riquadro **Coppia forma-lemma da sciogliere** (se cliccando su [ESTRAI LEMMA] si copia per errore un lemma sbagliato, basta ripetere l'operazione con quello giusto: i dati del nuovo lemma sostituiranno i precedenti nel riquadro di arrivo),
5. cliccare su [LISTA FORME] per accedere all'elenco delle forme lemmatizzate con **acconciare(-si)**, tra le quali deve trovarsi anche **acconcia**,
6. selezionare dalla lista che compare la forma **acconcia** cliccandovi sopra, così da copiare anch'essa nel riquadro **Coppia forma-lemma da sciogliere**,
7. selezionare nella griglia l'altro lemma coinvolto, ovvero **acconciare (v.)**, cliccandovi sopra,
8. scambiare i segnali stradali (cliccando su uno di essi) in modo che la freccia a sfondo giallo si trovi in **Nuovo lemma da associare alla forma**,
9. riutilizzare [ESTRAI LEMMA] per estrarre anche i dati del secondo lemma,
10. verificare di avere **acconcia** e **acconciare(-si) (v.)** come **Coppia forma-lemma da sciogliere** e **acconciare (v.)** come **Nuovo lemma da associare alla forma** (in caso contrario ricominciare daccapo),
11. effettuare l'operazione con [SOSTITUISCI IL LEMMA] (o annullarla con [ANNULLA]).

La griglia viene riscritta automaticamente. Se il lemma sostituito non risulta più utilizzato nel corpus, nel senso che era associato solo ad **acconcia**, viene automaticamente cancellato dal lemmario.

Se il lemma da rimuovere, nell'esempio **acconciare(-si)**, avesse dovuto essere rimosso da più coppie forma - lemma, ovvero se fosse stato erroneamente associato a più di una forma diversa, anziché ripetere completamente l'operazione per ognuna di queste forme si sarebbe potuto attivare il *check box* **congela il lemma da rimuovere** nel corso della prima sostituzione, prima di usare il pulsante [SOSTITUISCI IL LEMMA]: come risultato il lemma incluso nel riquadro **Coppia forma-lemma da sciogliere** non sarebbe stato cancellato al termine dell'operazione, permettendo di ridefinire ogni volta solo la forma in oggetto e il nuovo lemma da associarle.

Al termine usare [ESCI] per chiudere la finestra.

¹¹ L'attuale versione di Gatto ricerca automaticamente anche i lemmi terminanti con **(-se)** quindi, in questo caso, anche **acconciare(-se)**.

LEZIONE 58. COME RIMUOVERE COPPIE FORMA-LEMMA MUTE E LEMMI MUTI

Si ricorda che un elemento (lemma o coppia forma-lemma) è detto *muto* quando ad esso non sono associate occorrenze nel corpus.

Per provare ad eseguire queste operazioni è necessario disporre di un corpus che includa lemmi e coppie forma-lemma muti. A tal fine, eseguire preventivamente sul corpus **Demo2** la cancellazione del testo **od**, seguendo esattamente l'esempio descritto nella Lezione 48.

La rimozione di elementi muti può riguardare 3 casi diversi:

- 1 **rimozione di una singola coppia forma-lemma muta**
- 2 **rimozione di tutte le coppie forma-lemma mute relative ad uno o più lemmi**
- 3 **rimozione di tutte le coppie forma-lemma presenti nel corpus**

In tutti i casi se la rimozione effettuata comporta l'eliminazione di tutte le coppie forma-lemma associate ad un lemma, anche quest'ultimo viene eliminato automaticamente (si trattava quindi di un lemma cui erano associate solo coppie mute, motivo per il quale tale tipo di lemma è indicato col termine *lemma muto*).

CASO 1: RIMOZIONE DI UNA SINGOLA COPPIA FORMA-LEMMA MUTA

Questa rimozione si effettua semplicemente con una sorta di trucco, basato sul concetto che una coppia muta cessa di esserlo nel momento in cui viene assegnata ad almeno un'occorrenza nel corpus: a quel punto la successiva rimozione di tale unica occorrenza produce l'automatica cancellazione della coppia dagli archivi, secondo la logica standard seguita dal programma.

Si apra il corpus **Demo2** e si imposti la ricerca per lemmi di **caccia**. Avviare la ricerca con [AVVIO RICERCA]; si ottiene il lemma **caccia (s.f.)**, con **0** occorrenze (lemma muto); attivando **mostra forme** si ottiene la coppia muta lemma **caccia (s.f.)** + forma **caccia**. Questa è la coppia muta che vogliamo rimuovere (poiché ad essa fa riferimento un lemma muto, anch'esso al termine risulterà rimosso).

Passare all'ambiente **Lemmatizzazione sul corpus** mediante [ESCI] seguito da [ALTRO AMBIENTE], [LEMMATIZZAZIONE.. + SUL CORPUS] e [RICERCHE.. + PER FORME]; impostare la ricerca di **caccia** e procedere con [CONTESTI]. Si ottiene una sola occorrenza; tramite **N.lem/omo/ipe** lemmatizzarla con **caccia (s.f.)**. A questo punto, restando sull'occorrenza appena lemmatizzata, rimuovere la lemmatizzazione con **Canc. lemma**.

In questo modo l'operazione di rimozione è stata effettuata.

Per verifica, tornare all'ambiente **Ricerche** con **N. ricerca/Escl**, seguito da [ESCI], [ALTRO AMBIENTE], [RICERCHE] e [RICERCHE.. + PER LEMMI]; impostare la ricerca di **caccia** e procedere con [AVVIO RICERCA]. Si ottiene in risposta il messaggio **Nessun elemento localizzato**, a dimostrazione che il lemma **caccia** e la coppia forma-lemma ad esso associata sono stati rimossi dagli archivi.

CASO 2: RIMOZIONE DI TUTTE LE COPPIE FORMA-LEMMA MUTE RELATIVE AD UNO O PIÙ LEMMI

Restando nel corpus **Demo2** si imposti la ricerca per lemmi di **cagione**. Avviare la ricerca con [AVVIO RICERCA]; si ottengono 2 lemmi: **cagione (prep.)** (muto) e **cagione (s.f.)**; selezionarli entrambi e passare alle coppie forma-lemma associate con **mostra forme**. Si trova che a **cagione (prep.)** è associata l'unica coppia muta **cagione (prep.)** + **cagione**, mentre a **cagione (s.f.)** sono associate 3 coppie, una delle quali (**cagione (s.f.)** + **cagione**) è muta.

Si vogliono rimuovere le 2 coppie mute così evidenziate; al termine risulterà rimosso anche il lemma **cagione (prep.)**, mentre **cagione (s.f.)** rimarrà nel lemmario in quanto associato alle due coppie superstiti.

Portarsi alla finestra relativa alla rimozione tramite [ESCI] seguito da [ALTRO AMBIENTE], [GEST. BASE DATI] e [LEMMARIO.. + RIMOZIONE LEMMI E COPPIE FORMA-LEMMA MUTI]; scorrere la tabella dei lemmi fino a visualizzare **cagione (prep.)** e **cagione (s.f.)**. Selezionarli cliccando sul riquadro grigio a sinistra di ciascuna delle righe relative ai due lemmi (entrambe le righe appariranno in reverse). Usare **Annulla** in caso di errori nella selezione.

A questo punto effettuare l'operazione di rimozione delle coppie mute tramite **Rimuovi elementi muti (selez.)**, rispondendo **Sì** alla richiesta di conferma.

Come prima verifica, riposizionare la griglia in corrispondenza di **cagione**: si vedrà che il lemma (muto) **cagione (prep.)** è sparito. Passando poi all'ambiente **Ricerche**, impostando la ricerca del lemma **cagione** con **mostra forme** attivo ed avviandola con [AVVIO RICERCA]; si troveranno le **2** coppie forma-lemma superstiti mentre la terza (muta) non compare più in quanto cancellata dagli archivi, così come si voleva.

CASO 3: RIMOZIONE DI TUTTE LE COPPIE FORMA-LEMMA MUTE PRESENTI NEL CORPUS

Se nel corpus **Demo2** si effettua una ricerca per lemmi generica con **mostra forme** attivo, ad esempio di **as*** o **per***, tra i risultati ottenuti con [AVVIO RICERCA] ve ne sono alcuni evidenziati in rosso: si tratta di coppie forma-lemma mute.

Volendo rimuovere in un colpo solo tutte quelle presenti nel corpus, dalla finestra GESTIONE BASE DATI scegliere [LEMMARIO.. + RIMOZIONE LEMMI E COPPIE FORMA-LEMMA MUTI]. Poi cliccare su **Rimuovi elementi muti (tutti)** rispondendo **Sì** alla richiesta di conferma.

Per verifica, ripetere le ricerche per lemmi generiche accennate sopra: si troverà che tutte le coppie evidenziate in rosso sono sparite.

CAP. 27 - GESTIONE DEGLI IPERLEMMI

Nella Lezione 23 è stato visto come inserire nuovi iperlemmi nell'iperlemmario di un corpus.

Per completare il quadro della gestione degli iperlemmi vanno illustrate le operazioni di modifica e cancellazione degli iperlemmi già inseriti e di fusione di iperlemmi (quest'ultima operazione è analoga a quella descritta nella Lezione 56 in relazione ai lemmi). La finestra di gestione degli iperlemmi consente inoltre di navigare all'interno dell'iperlemmario, partendo da un iperlemma qualunque e passando a uno di quelli ad esso collegati, che diviene il nuovo punto di partenza dal quale si può continuare a spostarsi.

La modifica degli attributi è molto semplice. Va tenuto presente che sono modificabili l'assetto grafico e il disambiguatore, mentre non lo è il livello, per cambiare il quale bisogna sostituire l'iperlemma esistente con un altro del livello voluto.

Un iperlemma è modificabile anche nel senso di aggiungergli o togliergli connessioni agli iperlemmi di livello immediatamente superiore o inferiore (o ai lemmi se si tratta di un iperlemma di livello 1).

Se si tenta di cancellare anche l'ultima connessione al livello inferiore posseduta da un iperlemma l'operazione viene rifiutata, in quanto equivale alla cancellazione dell'iperlemma stesso. Volendo effettuare comunque questa operazione, lasciare la finestra di *modifica* e portarsi in quella di *cancellazione*.

Se si intende cancellare un iperlemma è importante avere ben chiaro che, se ad esso erano associati iperlemmi del livello superiore, ognuno di questi rimarrà nell'iperlemmario solo se disporrà di almeno un'altra associazione di livello inferiore, una volta persa quella fornita dall'iperlemma cancellato. Questo effetto si propagherà verso l'alto fino agli iperlemmi dei livelli via via più alti, per cui è importante valutare esattamente le conseguenze di una cancellazione, onde evitare la perdita non prevista di altri elementi dell'iperlemmario.

Anche la rimozione di lemmi, indipendentemente dalla causa, può produrre con lo stesso meccanismo la cancellazione di iperlemmi di livello 1 e poi dei livelli via via superiori.

La cancellazione di un iperlemma di livello 1 lo rimuoverà, ovviamente, anche da tutte le occorrenze alle quali sia stato specificamente associato in fase di lemmatizzazione.

Con esclusione di questi ultimi meccanismi di rimozione *indotta* di iperlemmi dall'iperlemmario, la loro cancellazione può essere operata solo manualmente; da questo punto di vista la gestione degli iperlemmi si differenzia nettamente da quella dei lemmi, basata essenzialmente su procedure automatiche.

Nella Lezione 25 è stato illustrato il procedimento per *esportare* un lemmario da un corpus e per importarlo successivamente in un altro corpus. Se il corpus di partenza includeva degli iperlemmi, questi vengono automaticamente trascritti nel database esportato; la loro successiva importazione in un altro corpus è invece opzionale, in funzione delle scelte dell'utente.

Nell'importazione degli iperlemmi in un nuovo corpus valgono le seguenti regole:

1. gli iperlemmi di livello 1 vengono importati solo se sono collegati ad almeno un lemma presente nel corpus di arrivo prima dell'importazione o inserito al suo interno nel corso dell'operazione;
2. gli iperlemmi di livello via via superiore vengono importati solo se sono collegati a iperlemmi di livello inferiore presenti nel corpus di arrivo prima dell'importazione o inseriti al suo interno nel corso dell'operazione;
3. le iperlemmatizzazioni applicate a specifiche occorrenze non sono mai importate: questo perché l'importazione degli iperlemmi non ha riferimenti con le occorrenze delle parole nei testi, ma attiva solo dei link più generali con i lemmi presenti nel corpus.

LEZIONE 59. COME MODIFICARE O CANCELLARE UN IPERLEMMA

MODIFICA DEGLI ATTRIBUTI DI UN IPERLEMMA

Si supponga di voler modificare in **Demo** l'iperlemma di livello **2 scienza e tecnica** per introdurre le modifiche seguenti:

- l'assetto grafico deve divenire **scienza/tecnica**,
- va inserito il disambiguatore **dizionario specifico**,
- va rimosso l'iperlemma associato di livello **1 matematica**,
- va associato l'iperlemma di livello **3 uomo**.

Per operare queste modifiche, dalla finestra GESTIONE BASE DATI scegliere [IPERLEMMI] e quindi [MODIFICA]. Selezionare nella griglia in alto l'iperlemma **scienza e tecnica** cliccando nel relativo riquadro grigio all'inizio della riga. Copiarlo nel riquadro a destra **iperlemma da modificare** cliccando sulla freccia blu. A questo punto si può procedere con le modifiche:

1. correggere l'assetto grafico modificandolo in **scienza/tecnica**,
2. nella casella **disambiguatore** digitare **dizionario specifico**,
3. nella griglia in basso a destra selezionare l'iperlemma associato **matematica** e rimuoverne l'associazione con <Canc>>,
4. nella griglia superiore selezionare **uomo** (di livello 3) e connetterlo a **scienza/tecnica** con (**Collega**),
5. rendere effettive tutte le modifiche impostate con (**Registra**) o annullarle con (**Annulla**).

CANCELLAZIONE DI UNO O PIÙ IPERLEMMI

Si voglia cancellare l'iperlemma **corpo**.

Dalla finestra GESTIONE BASE DATI scegliere [IPERLEMMI], e quindi [CANCELLA]. Selezionare nella griglia l'iperlemma **corpo** cliccando nel relativo riquadro grigio all'inizio della riga. Per cancellare l'iperlemma utilizzare il bottone (**Cancella**) e confermare l'operazione.

LEZIONE 60. COME FONDERE DUE IPERLEMMI INTRODOTTI ERRONEAMENTE COME DIVERSI

Scopo di questa operazione è di ovviare ai problemi generati dall'eventuale inserimento nell'iperlemmario di due iperlemmi identici come significato, ma registrati come diversi per errori di digitazione o per altri motivi.

Se in seguito a questi iperlemmi sono stati associati più lemmi e, magari, anche occorrenze direttamente iperlemmatizzate, può essere complicato ricondurre manualmente tutti i riferimenti presenti nel corpus, relativi all'iperlemma *sbagliato*, all'altro.

L'operazione di fusione di due lemmi in uno consente invece di effettuare in maniera automatica la riunificazione dei due iperlemmi, cancellandole uno e reindirizzando all'altro tutti i riferimenti presenti nel corpus. L'unico vincolo è che i due iperlemmi devono essere dello stesso livello.

Si supponga di voler fondere tra loro, all'interno del corpus **Demo**, gli iperlemmi **attività sensoriali e linguaggio**, nel senso di sopprimere quest'ultimo facendolo confluire nel precedente.

Dalla finestra GESTIONE BASE DATI scegliere [IPERLEMMI] e quindi [FONDI 2 IN 1]. Selezionare nella griglia l'iperlemma **attività sensoriali** e copiarlo nella casella **iperlemma da mantenere** cliccando sulla freccia blu superiore; notare che l'iperlemma è collegato superiormente con **uomo fisico** e inferiormente con i lemmi **occhio, orecchia, orecchio, vedere e vedere(-si)**; notare anche che nel corpus esistono 2 occorrenze lemmatizzate con **occhio** e iperlemmatizzate con **attività sensoriali**.

Ripetere l'operazione precedente con **linguaggio**, copiandolo nella casella **iperlemma da eliminare** tramite la freccia blu inferiore; questa volta vi è una connessione superiore all'iperlemma **uomo sociale** e connessioni inferiori ai lemmi **lingua, parlare, parola, segno, versigliare, verso e voce**; analogamente al caso precedente nel corpus sono presenti 2 occorrenze lemmatizzate con **lingua** e iperlemmatizzate con **linguaggio**.

Come si può notare i due iperlemmi trascritti nelle caselle sono seguiti dal carattere "/"; nel caso dell'operazione di fusione, infatti, in ciascuna delle due caselle appena riempite vengono copiati entrambi i dati iperlemma e disambiguatore, separati dal carattere "/". Poiché nel caso in questione non vi sono disambiguatori, la barra è l'ultimo carattere inserito nelle caselle.

Per procedere alla fusione dei due lemmi (in pratica all'assorbimento di **linguaggio** da parte di **attività sensoriali**) utilizzare il bottone (**Esegui**) e confermare l'operazione.

Si vede subito che l'iperlemma **linguaggio** non è più presente. Per vedere come si è trasformato **attività sensoriali**, scegliere [NAVIGA], selezionare l'iperlemma e copiarlo nella casella **iperlemma** tramite la freccia blu. È facile vedere che ora **attività sensoriali** raccoglie in sé tutte le connessioni, sia superiori che inferiori, che erano prima complessivamente associate ai due lemmi. Anche le 2 occorrenze prima direttamente iperlemmatizzate con **linguaggio** (e lemmatizzate con **occhio**) sono ora confluite nel nuovo iperlemma e possono essere agevolmente localizzate con una ricerca per iperlemmi (utilizzando [TROVA.. + OCCORRENZE ASSOCIATE A IPERLEMMI]).

Chiudere la finestra con (**Annulla**).

LEZIONE 61. COME NAVIGARE ALL'INTERNO DELLA STRUTTURA DELL'IPERLEMMARIO

È stato già utilizzato il menu [NAVIGA] per acquisire informazioni sugli iperlemmi senza effettuare alcuna operazione su di essi.

Vediamo con un esempio come usarlo per muoversi all'interno della struttura ad albero dell'iperlemmario.

Con riferimento al solito corpus **Demo** e allo schema riportato nel capitolo 12, immaginiamo di voler attraversare lo schema, partendo da **attività sensoriali** per passare poi nell'ordine a **uomo fisico**, **uomo**, **uomo sociale**, **scrittura**, **scienza e tecnica**¹² e infine **matematica**.

Un percorso di questo genere può servire, ad esempio, a verificare rapidamente la correttezza degli archivi realizzati e la loro rispondenza agli schemi che si intendevano realizzare.

La tecnica di navigazione è semplicissima.

Dalla finestra GESTIONE BASE DATI scegliere [IPERLEMMI] e poi [NAVIGA]. Selezionare nella griglia l'iperlemma **attività sensoriali** e copiarlo nella casella **iperlemma** cliccando sulla freccia blu; l'iperlemma è collegato superiormente (anche) a **uomo fisico**; selezionare quest'ultimo nella griglia in basso a sinistra e portarlo nella casella **iperlemma** (in luogo di **attività sensoriali**) tramite la freccia blu; i contenuti del riquadro **iperlemma corrente** e delle due griglie in basso riflettono ora la nuova situazione. Passare a **uomo** selezionandolo e trasferendolo con la freccia blu; procedere nell'ordine, allo stesso modo, con **uomo sociale**, **scrittura**, **scienza e tecnica** e **matematica**.

Chiudere la finestra con (**Annulla**).

CAP. 28 - GESTIONE DEGLI ARCHIVI BIBLIOGRAFICI

Nella Lezione 11 è stato mostrato come costruire un nuovo corpus, dotandolo contestualmente di un archivio bibliografico. Se l'archivio indicato non esiste viene creato automaticamente.

È possibile più in generale costruire un nuovo archivio bibliografico con un'operazione indipendente dalla creazione di un nuovo corpus.

È ovviamente prevista anche l'operazione contraria, ovvero la cancellazione di un archivio bibliografico; a questo riguardo va tenuto presente che è compito dell'utente assicurarsi che l'archivio bibliografico cancellato non sia più utilizzato da altri corpus esistenti, nel qual caso questi ultimi non potrebbero più essere aperti. Se un evento del genere si verificasse, l'apertura dei corpus potrebbe essere ripristinata ricreando, mediante il meccanismo illustrato nella Lezione che segue, un archivio bibliografico (anche vuoto) avente lo stesso nome e collocazione di quello cancellato; successivamente, per poter effettuare ricerche nei corpus interessati, andrebbe comunque ripristinato anche il contenuto dell'archivio bibliografico.

Quando si effettuano operazioni di gestione su un qualunque archivio bibliografico l'eventuale corpus correntemente aperto viene automaticamente chiuso, anche se l'archivio bibliografico cui è collegato non è quello del corpus..

Le altre operazioni di gestione di un archivio bibliografico sono inserimento, modifica e cancellazione di un record, ovvero del pacchetto di dati relativi ad un testo.

Nota

Non utilizzare programmi diversi da GATTO per aprire l'archivio bibliografico di un corpus e apportarvi modifiche. Infatti nell'eseguire le operazioni di gestione GATTO ne effettua anche altre, non rilevabili

¹² Se è stata svolta l'esercitazione della Lezione 57 ora dovrebbe chiamarsi **scienza/tecnica**.

dall'utente e non eseguite ovviamente da un programma diverso, il che potrebbe dare origine a situazioni di errore nell'uso successivo.

LEZIONE 62. COME EFFETTUARE LE OPERAZIONI DI GESTIONE DI UN ARCHIVIO BIBLIOGRAFICO

CREAZIONE DI UN NUOVO ARCHIVIO

Si supponga di voler creare l'archivio bibliografico **bibliobis.mdb** nella directory **C:\demobiblio**, già utilizzata per la bibliografia di **Demo**.

Dalla finestra GESTIONE BASE DATI scegliere [**BIBLIOGRAFIE.. + CREA**]; assegnare nome e collocazione del nuovo archivio e confermare con (**Salva**); notare che l'estensione **mdb** può anche essere omessa, nel qual caso verrà aggiunta automaticamente dal programma.

Un messaggio conferma l'avvenuta costruzione dell'archivio bibliografico.

CANCELLAZIONE DI UN ARCHIVIO

Per cancellare l'archivio **bibliobis.mdb**, testé creato nella directory **C:\demobiblio**, dalla finestra GESTIONE BASE DATI scegliere [**BIBLIOGRAFIE.. + ELIMINA**]; localizzare l'archivio che si vuole rimuovere dal disco e procedere con (**Apri**); un messaggio chiede conferma della cancellazione: effettuarla con (**Si**) oppure annullare l'operazione con (**No**).

Un messaggio conferma l'avvenuta cancellazione dell'archivio bibliografico.

INSERIMENTO DI UN NUOVO RECORD

Questa operazione è stata già descritta nella Lezione 12. Il testo relativo è qui riassunto per comodità di consultazione.

Per *record* di un archivio bibliografico si intende il pacchetto completo dei dati bibliografici di un testo.

Prima di procedere, costruire di nuovo l'archivio bibliografico **bibliobis.mdb**, secondo quanto già descritto in questa stessa Lezione.

Dalla finestra GESTIONE BASE DATI scegliere [**BIBLIOGRAFIE.. + APRI**]; localizzare l'archivio **bibliobis.mdb** e procedere con (**Apri**).

Tramite [**MODIF. DATI.. + INSERISCI**] attivare l'operazione di inserimento di un nuovo record nell'archivio (unica operazione consentita in un archivio vuoto): viene presentata la scheda di immissione dati.

Digitare **Anonimo** nel campo **autore**, **Lauda in lucchese** nel campo **titolo**, **Poes. an. lucch. XIV in. in abbr. tit., es1 in sigla, tosc. in area gen. e XIV in. in data desc..**

Infine digitare **13101300AL** nella **data codificata**.

Questi dati completano l'inserimento delle informazioni bibliografiche relative al testo **es1**. Per rendere effettivo l'inserimento (per ora solo impostato) non rimane che utilizzare il menu [**REGISTRA**], mentre con [**ANNULLA**] l'inserimento verrebbe fatto abortire. Il record di **es1** viene inserito nell'archivio e la scheda di inserimento viene disabilitata (sfondo grigio).

Per chiudere la finestra di gestione dell'archivio bibliografico usare [**CHIUDI FINESTRA**].

MODIFICA DEI DATI DI UN RECORD

Questa operazione è stata già descritta nella Lezione 12. Ne viene fornita qui una nuova descrizione per comodità di consultazione.

Sempre con riferimento alla bibliografia **bibliobis.mdb** si supponga di voler effettuare alcune modifiche ai dati del testo **es1**:

1. cambiare il titolo in **lauda in senese**,
2. rimuovere l'indicazione **tosc.** nel campo **area gen.**,
3. inserire l'indicazione **Siena** nel campo **area spec.**

Dalla finestra GESTIONE BASE DATI scegliere [**BIBLIOGRAFIE.. + APRI**]; localizzare l'archivio **bibliobis.mdb** e procedere con (**Apri**).

Cliccare in un punto qualunque della riga relativa al testo **es1** per selezionarlo (la riga apparirà evidenziata).

Tramite [**MODIF. DATI.. + MODIFICA**] attivare l'operazione di modifica dei dati: viene presentata la scheda di immissione dati.

Effettuare le modifiche indicate nei tre campi interessati, quindi salvarle con [**REGISTRA**] o annullarle con [**ANNULLA**].

Per chiudere la finestra di gestione dell'archivio bibliografico usare [**CHIUDI FINESTRA**].

CANCELLAZIONE DI UN RECORD

Si voglia infine rimuovere il record relativo al testo **es1** dall'archivio bibliografico **bibliobis.mdb**.

Dalla finestra GESTIONE BASE DATI scegliere [**BIBLIOGRAFIE.. + APRI**]; localizzare l'archivio **bibliobis.mdb** e procedere con (**Apri**).

Cliccare in un punto qualunque della riga relativa al testo **es1** per selezionarlo (la riga apparirà evidenziata).

Tramite [**MODIF. DATI.. + ELIMINA**] avviare la cancellazione del record. Viene chiesta conferma della cancellazione del record: procedere con (**Si**) o annullarla con (**No**).

Per chiudere la finestra di gestione dell'archivio bibliografico usare [**CHIUDI FINESTRA**].

PARTE VI - COMANDI E AZIONI COMUNI A PIU' AMBIENTI

CAP. 29 - COMANDI DI USO GENERALE

CHIUSURA O RIDUZIONE A ICONA DELLE FINESTRE

Finestre primarie

Col termine *primarie* si intendono le normali finestre di GATTO. In ogni momento è aperta una e una sola finestra primaria. Quando da una di queste finestre si passa ad un'altra viene normalmente eseguita una serie di operazioni legate all'ambiente attivo e più specificamente all'azione in corso di svolgimento, operazioni che non potrebbero aver luogo in caso di chiusura immediata. Per questo motivo in tutte le finestre primarie il simbolo **X** posto in alto a destra è disabilitato. Per contro è sempre possibile ridurle a icona col simbolo **_**.

Finestre secondarie

Col termine *secondarie* in GATTO ci si riferisce alle finestre aperte al di sopra di altre. In genere si tratta di finestre più piccole, che coprono quelle sottostanti solo parzialmente. Per chiuderle si può indifferentemente usare il tasto **<Esc>** o cliccare sul simbolo **X** posto in alto a destra (non avendo chiaro se la finestra considerata è o meno secondaria si può sempre provare: sulle finestre primarie questi due comandi non hanno effetto). Non è prevista la riduzione a icona delle finestre secondarie.

SELEZIONE DEGLI ELEMENTI DI UNA GRIGLIA

Molte delle finestre di GATTO contengono delle tabelle di dati visualizzate all'interno di apposite griglie.

Spesso è necessario selezionare una o più righe all'interno di queste tabelle per indicare al programma quali sono i dati, inclusi al loro interno, sui quali si intende operare. Può trattarsi ad esempio di una lista di forme ottenute nel corso della ricerca, di alcune delle quali (ma non di tutte) si vogliono vedere i contesti. Si pone quindi il problema di selezionare una o più righe non necessariamente consecutive di una tabella che può contenere anche un numero elevato di elementi.

Quando si vuole effettuare una selezione di questo genere in una qualunque tabella di GATTO valgono i comandi normalmente usati in Windows, con gli stessi significati: quindi si dovrà cliccare sulle righe desiderate, definendo il tipo di selezione mediante la contemporanea pressione sui tasti di controllo **<Ctrl>** e **<Maiusc>**, secondo lo schema che segue:

nessun tasto di controllo	la selezione sostituisce quelle preesistenti,
<Ctrl>	la selezione si aggiunge a quelle preesistenti (cliccando di nuovo un elemento selezionato lo si deselecta),
<Maiusc>	vengono selezionati (o alternativamente deselectati) tutti gli elementi compresi tra quello selezionato ora e l'ultimo selezionato in ordine di tempo; le selezioni preesistenti vengono rimosse,
<Maiusc> + <Ctrl>	vengono selezionati (o alternativamente deselectati) tutti gli elementi compresi tra quello selezionato ora e l'ultimo selezionato in ordine di tempo; le selezioni preesistenti vengono mantenute.

La seconda e la quarta modalità comportano il rischio di perdita non voluta delle selezioni effettuate in precedenza se ci si dimentica di mantenere premuto il tasto **<Ctrl>**. Per evitare questo rischio accanto a molte tabelle compare il *check box* **blocco selezionati** il quale, se attivato, mantiene la selezione preesistente, mentre si selezionano nuovi elementi, indipendentemente dal fatto che il tasto **<Ctrl>** sia premuto o meno.

In pratica, è come se il tasto **<Ctrl>** fosse sempre premuto.

Infine, un modo per selezionare o deselectare tutte le righe di una griglia è quello di utilizzare, dove presenti, i menu **[SEL/DESEL.. + SELEZIONA TUTTO]** e **[SEL/DESEL.. + DESELEZIONA TUTTO]** (o menu caratterizzati da diciture analoghe).

ESPORTAZIONE DEL CONTENUTO DI UNA GRIGLIA

I contenuti delle griglie presenti nelle diverse finestre di GATTO possono essere *esportati*, ovvero stampati su carta o trascritti su file RTF (Rich Text Format).

A questo scopo una prima cosa da fare è quella di decidere se esportare tutte le righe contenute nella griglia o solo una parte di esse; nel secondo caso bisogna provvedere a selezionare le righe desiderate utilizzando i comandi illustrati nel paragrafo precedente.

In partenza l'esportazione riguarda tutti gli elementi inclusi in una riga, ovvero tutte le colonne della stessa. È però possibile escludere dall'esportazione le colonne che non interessano: questa opzione può risultare particolarmente utile in presenza di un elevato numero di colonne, con conseguenti problemi di impaginazione dei dati esportati. Per escludere una colonna dall'esportazione è sufficiente spostare completamente a sinistra il separatore di campo successivo, ovvero la linea nera verticale che separa la sua intestazione da quella della colonna alla sua destra: l'operazione si effettua col mouse, cliccando sulla riga nera suddetta e trascinandola completamente a sinistra finché non si arresta, in prossimità di quella che la precede.

Una volta operata l'eventuale selezione di righe e colonne rimane comunque aperta la possibilità di esportare sia le sole righe selezionate che tutte quelle incluse nella griglia, scegliendo di avviare l'esportazione tramite i menu [STAMPA.. + TUTTE LE RIGHE] oppure [STAMPA.. + RIGHE SELEZIONATE] (o menu caratterizzati da diciture analoghe).

La successiva finestra IMPOSTAZIONI DI STAMPA offre la scelta tra stampa su carta, copia su file o uso di entrambe le possibilità.

Stampa su carta

La stampa su carta dà accesso a successive finestre di impostazione relative al formato della carta, all'orientamento della stampa nella pagina e, in generale, alle diverse opzioni che la stampante utilizzata offre.

Copia su file

La copia su file offre 3 opzioni alternative:

- uso di un nuovo file, del quale andranno indicati nome e collocazione mentre l'estensione, ovvero **rtf**, viene assegnata direttamente da GATTO,
- utilizzo di un file RTF esistente da specificare subito dopo, con accodamento dei dati della tabella al contenuto attuale del file di destinazione,
- utilizzo di un file RTF esistente da specificare subito dopo, con sostituzione dei dati della tabella al contenuto attuale del file di destinazione.

Facendo uso di Word è in seguito possibile convertire i dati trascritti nel file in una tabella vera e propria. Per fare questo, dopo averlo aperto con Word selezionare tutto il contenuto del file con esclusione delle prime righe includenti informazioni quali data, nome del corpus, natura e numero degli elementi riportati nel file o altro. In altre parole la parte selezionata deve partire dalle intestazioni delle colonne della griglia, laddove presenti, oppure direttamente dai dati inclusi. A quel punto fare uso del menu [TABELLA.. + CONVERTI IL TESTO IN TABELLA..], seguito da **(OK)**, avendo cura di definire come separatori le tabulazioni.

Si può infine ottenere una migliore formattazione della tabella così prodotta utilizzando, sempre in Word, il menu [TABELLA.. + ALTEZZA E LARGHEZZA CELLE] seguito da **(Adatta)**.

LEZIONE 63. COME STAMPARE IL CONTENUTO DELL'ACCUMULATORE

A titolo di esempio delle cose dette in questo capitolo si esporterà su carta e su file il contenuto dell'ACCUMULATORE. Questa particolare griglia presenta tra l'altro un tipo di esportazione aggiuntivo, non presente nelle altre griglie.

Esempio 1: esportazione dell'intero contenuto della griglia.

Nel corpus **Demo** dalla finestra MODALITÀ RICERCA avviare la ricerca di tutti i lemmi che cominciano per **car** tramite [RICERCHE.. + PER LEMMI] seguito dalla digitazione di **car*** in un selettore e da [AVVIO RICERCA]; si selezionino tutte le righe e si passi all'ACCUMULATORE con [COPIA IN ACC.]. Si avvii la stampa con [STAMPA.. + TUTTE LE RIGHE]; in IMPOSTAZIONI DI STAMPA si attivino entrambi i *check box* presenti, ovvero **stampante** e **file** e si proceda con **(OK)**, lasciando impostato **file nuovo**. Si scelga come nome del

file **prova14**, in una cartella a piacere, e si continui con (**Salva**). Per quanto riguarda la stampante lasciare le impostazioni proposte e procedere con (**OK**).

Volendolo, aprire con Word il file **prova14.rtf** così generato, selezionarne tutto il contenuto a partire da "**lemma cat. gr.**", usare il menu [TABELLA.. + CONVERTI IL TESTO IN TABELLA..], controllare che come separatori di testo siano previste le tabulazioni (in caso contrario selezionarle) e continuare con (**OK**); infine cliccare in un punto qualunque all'interno della tabella ottenuta, selezionare tutte le colonne della tabella e proseguire con [TABELLA.. + ALTEZZA E LARGHEZZA CELLE] seguito da (**Adatta**).

Il risultato, sostanzialmente equivalente a quello stampato su carta, dovrebbe essere quello qui riportato:

corpus DEMO Accumulatore <data>

lemma	cat. gr.	forma	disambiguatore	occ.	commento
cara	s.f.	*		1	
carbonchio	s.m.	*		1	
carbone	s.m.	*		1	
carcere	s.f.	*		1	
cardinale	s.m.	*		3	
carestia	s.f.	*		1	
caricare	v.	*		7	
caricare(-si)	v.	*		1	
carico	s.m.	*		1	
carità	s.f.	*		6	
carnale	agg.	*		9	
carne	s.f.	*		9	
carnefice	s.m.	*		1	
caro	agg.	*		10	
caro	s.m.	*	carestia	1	
carogna	s.f.	*		1	
carretta	s.f.	*		2	
carta	s.f.	*		7	

Esempio 2: esportazione delle righe selezionate, con esclusione di alcune colonne.

In questo caso la terza e sesta colonna erano prive di contenuti informativi, avendo scelto di trovare i lemmi a prescindere dalle forme associate ed essendo gli elementi individuati tutti privi di commenti. Si potrebbe quindi pensare di escluderle dalla esportazione.

Si supponga inoltre di essere interessati solo ai 7 lemmi compresi tra **carbone** e **carico**, oltre a **caro (s.m.)**.

Nella finestra ACCUMULATORE si rimuova l'eventuale selezione di righe con [SEL/DESEL.. + DESELEZIONA TUTTO] e si rifelezionino gli 8 lemmi indicati con le tecniche riportate in questo capitolo; poi si escludano dall'esportazione le colonne **forma** e **commento** minimizzandone la larghezza (vedere sopra in questo stesso capitolo); infine si proceda con [STAMPA.. + RIGHE SELEZIONATE] alla costruzione nel nuovo file **prova15**.

In Word, elaborare il file **prova15.rtf** allo stesso modo del file precedente.. Il risultato dovrebbe essere il seguente (come si può notare sono scomparse le colonne indesiderate):

corpus DEMO Accumulatore <data>

lemma	cat. gr.	disambiguatore	occ.
carbone	s.m.		1
carcere	s.f.		1
cardinale	s.m.		3
carestia	s.f.		1
caricare	v.		7
caricare(-si)	v.		1
carico	s.m.		1
caro	s.m.	carestia	1

Esempio 3: esportazione dall'ACCUMULATORE della **lista delle forme** associate al lemma **abbandonare**.

Si tratta di un particolare tipo di esportazione, specificamente finalizzata alla stesura di voci di vocabolario. Questa uscita è presente per il solo ACCUMULATORE e prevede unicamente la scrittura su file, secondo un formato particolare, della lista delle forme che compaiono nell'ACCUMULATORE, con rimozione automatica di eventuali doppioni.

Per sperimentare questa uscita, da MODALITÀ RICERCA impostare la ricerca per lemmi con [RICERCHE.. + PER LEMMI]; digitare **abbandonare** in un selettore, attivare **mostra forme** e procedere con [AVVIO RICERCA]; svuotare eventualmente l'ACCUMULATORE con [SVUOTA ACC.], selezionare tutte le righe con [SEL/DESEL.. + SELEZIONA TUTTO] e proseguire con [COPIA IN ACC.]. Nell'ACCUMULATORE ordinare la griglia *per forme* con **ordina per forme** e (A - Z). Infine trasferire la lista delle forme nel file **prova16** con [STAMPA.. + LISTA FORME (TUTTE)], scegliendo successivamente **file** e (OK). Un messaggio conferma in questo caso l'avvenuta trascrizione delle forme.

Il contenuto del file **prova16.rtf** dovrebbe essere il seguente, con le forme in corsivo, separate da virgole e spazi in tondo:

abandonano, abandonare, abandonari, abandonarj, abandonaru, abandonata, abandonato, abandonatu, abandonau, abandoni, abandugninu, abanduna, abandunare, abandunati, abandunau, habandonato

La stessa operazione può essere eseguita limitatamente alle righe selezionate della griglia mediante [STAMPA.. + LISTA FORME (RIGHE SELEZIONATE)].

APPENDICE A - LIMITI DI GATTO

Lunghezza massima di una forma, di un lemma, di una categoria grammaticale, di un iperlemma: 250 caratteri.

Numero massimo di lemmi associabili ad una singola forma: 99.

Lunghezza massima dei contesti a parole: 81 parole.

Lunghezza massima dei contesti a periodi: 15 periodi.

Numero massimo di parole cooccorrenti cercabili: 10.

Numero massimo di testi in un corpus: 16.350.

Numero massimo di sottocorpus definibili contemporaneamente: 6.

Numero massimo di testi nella finestra statistica: 100.

Numero massimo di contesti localizzabili in una singola ricerca: 20.000.000.

Numero massimo di corpus localizzabili contemporaneamente: 50.

APPENDICE B - IMPOSTAZIONE DELLE OPZIONI GENERALI

Le funzioni di GATTO sono regolate da una serie di parametri alfanumerici. La maggior parte di questi parametri non è accessibile all'utente per ovvi motivi di sicurezza di funzionamento del programma e quindi non è modificabile; alcuni possono invece essere cambiati e l'utente è libero, entro certi limiti, di attribuire loro valori diversi da quelli iniziali. Questo può essere fatto, a seconda dei casi, utilizzando dei menu, delle caselle di testo o altri elementi dell'interfaccia grafica: nel manuale sono stati indicati di volta in volta gli strumenti disponibili.

Questo meccanismo di modifica dei parametri di GATTO presenta però il limite che i nuovi valori hanno durata limitata: riavviando il programma vengono ripristinati i valori di partenza.

Ad esempio se, durante la localizzazione dei contesti nel corso di una ricerca, si raggiunge un certo numero limite, viene chiesto all'utente se continuare con la ricerca di ulteriori contesti o mostrare quelli trovati fino a quel momento. Il valore limite preimpostato per questo controllo è 300 contesti. L'utente può modificare questo dato e il nuovo valore rimarrà valido per le ricerche successive, anche cambiando corpus; se però si esce da GATTO e in seguito si riavvia il programma si trova che questo parametro è stato riportato a 300.

Volendo cambiare in modo permanente i valori iniziali dei parametri bisogna fare ricorso alla funzione di modifica delle opzioni generali, cui si accede dalla finestra di avvio del programma. I dati assegnati con questo meccanismo costituiscono i nuovi valori preimpostati dai quali parte GATTO ad ogni riavvio: naturalmente possono essere modificati temporaneamente durante la sessione così come lo erano quelli precedenti. I nuovi valori preimpostati si mantengono inalterati anche in caso di *aggiornamento* di GATTO con nuove revisioni.

Reinstallando il programma le nuove impostazioni vengono invece cancellate.

Per accedere alla funzione di modifica delle opzioni generali avviare GATTO e cliccare su [OPZIONI]. vengono presentate le opzioni generali di GATTO ripartite in 8 pagine a seconda dell'argomento cui si riferiscono:

- ricerca per forme,
- ricerca per lemmi,
- ricerca di cooccorrenze,
- ricerche (in generale),
- costruzione di un nuovo corpus,
- inserimento di un nuovo testo,
- lemmatizzazione,
- gestione dei dati.

Per impostare i nuovi valori delle opzioni utilizzare i *check box* o le caselle di testo inserite nelle varie pagine; al termine, per rendere effettive le modifiche, usare [REGISTRA]; poi, per tornare alla finestra di avvio di GATTO, usare [CHIUDI]. Chiudendo la finestra OPZIONI DI PERSONALIZZAZIONE con [CHIUDI] senza aver prima usato [REGISTRA] verranno mantenuti i valori precedenti le eventuali modifiche apportate.

Il bottone [IMPOSTA I VALORI DI DEFAULT] predispone il ripristino dell'intero set di valori di tutte le 8 pagine a quelli preimpostati all'atto dell'installazione di GATTO: anche in questo caso, per rendere effettiva la riassegnazione, va poi usato [REGISTRA].

Verranno ora passate in rassegna le opzioni disponibili, indicando dove presenti i limiti dei valori assegnabili. Qualora il dato introdotto superi il limite impostato esso verrà automaticamente riportato a quel valore.

OPZIONI IN RICERCA PER FORME

mostra lemmi

Pilota l'attivazione iniziale o meno del *check box* corrispondente.

ricerca espansa

Pilota la simultanea attivazione iniziale o meno di tutti i 10 *check box* di questa opzione.

varianti

Pilota la simultanea attivazione iniziale o meno di tutti i 10 *check box* di questa opzione.

anche forma con iniziale raddoppiata

Pilota la simultanea attivazione iniziale o meno di tutti i 10 *check box* di questa opzione.

input brano: spezza dopo diacritico

Pilota l'attivazione iniziale o meno del *check box* corrispondente nelle ricerche avviate a partire da un brano di testo.

OPZIONI IN RICERCA PER LEMMI**mostra forme**

Pilota l'attivazione iniziale o meno del *check box* corrispondente.

ricerca espansa

Pilota la simultanea attivazione iniziale o meno di tutti i 10 *check box* di questa opzione.

forma pronominale

Pilota la simultanea attivazione iniziale o meno di tutti i 10 *check box* di questa opzione.

OPZIONI IN RICERCA DI COCCORRENZE**entro 1 periodo**

Pilota l'attivazione iniziale o meno del *check box* che limita l'intervallo di validità ad un periodo.

entro N parole

Pilota l'attivazione iniziale o meno del *check box* che limita l'intervallo di validità ad un certo numero di parole e ne fissa il valore iniziale. Il valore preimpostato è **10**; i valori accettati sono compresi tra **1** e **10**.

OPZIONI NELLE RICERCHE IN GENERALE**blocco selezionati**

Pilota l'attivazione iniziale o meno dei *check box* corrispondenti.

contesti di N parole

Imposta la dimensione iniziale dei contesti visualizzati e dimensionati a numero di parole. Il valore preimpostato è **31**; i valori accettati sono compresi tra **1** e **81**.

dopo N contesti chiede se continuare

Imposta il numero di contesti da localizzare comunque, prima di chiedere se proseguire o meno. Il valore preimpostato è **300**; il valore minimo è **1**.

numero massimo di elementi ordinati automaticamente: N

Imposta il numero massimo di forme o lemmi trovati oltre il quale non si procede automaticamente al loro ordinamento alfabetico prima della presentazione a video; l'utente può sempre effettuarla in un secondo tempo. Il valore preimpostato è **300**; i valori accettati sono compresi tra **2** e **30.000**.

distanza massima per 'raffina': N

Imposta la distanza massima, in parole rispetto alla parola centrale, entro la quale vengono cercate le parole cooccorrenti. Il valore preimpostato è **10**; i valori accettati sono compresi tra **1** e **20**.

stampa lemmi

Pilota l'attivazione iniziale o meno del *check box* corrispondente in fase di esportazione dei contesti.

stampa note/trad

Pilota l'attivazione iniziale o meno del *check box* corrispondente in fase di esportazione dei contesti.

cartella di uscita

Imposta la cartella nella quale, per default, vengono collocati i file contenenti le uscite richieste dall'utente. La cartella preimpostata è la radice del disco C, ovvero "C:\".

OPZIONI NELLA CREAZIONE DI UN NUOVO CORPUS

Le opzioni incluse in questa pagina assegnano i valori iniziali a tutti i parametri di un nuovo corpus, al momento della sua creazione.

Nella parte sinistra della finestra sono attivabili le opzioni seguenti:

- **possibilità di definire ' caratteri "?" e "!" come punteggiatura forte**, ovvero come separatori tra periodi;
- **possibilità di attribuire la stessa valenza ai puntini di sospensione "..."**;
- **possibilità di utilizzare anche le 10 cifre arabe come componenti di forme e lemmi**, per esempio volendo indicizzare le date presenti nel testo.
- **possibilità di utilizzare, come coordinate delle parole nel testo, numeri di pagina e di riga** (questa scelta implica il riferirsi ad una specifica edizione del testo); sono possibili 3 alternative:
 - **num. di pag. non visualizzati**: i riferimenti organici non vengono mostrati, indipendentemente dal fatto che i testi (o anche parti di testi) li contengano o meno;
 - **num. di pag. facoltativi**: i riferimenti organici vengono mostrati solo se relativi ad occorrenze incluse in testi (o parti di testi) che li riportano; perché una parte di testo non dia luogo ai numeri di pagina e riga, marcarla col codice **\$0\$** (cioè pagina 0)
 - **num. di pag. obbligatori**: i riferimenti organici vengono mostrati e devono essere sempre presenti nei testi, pena la mancata accettazione dei testi stessi all'interno del corpus.

I valori numerici preimpostati (numeri di caratteri) per le **lunghezze massime** sono:

- **26** per forme, lemmi, disambiguatori dei lemmi, iperlemmi, disambiguatori degli iperlemmi,
- **10** per le categorie grammaticali,
- **50** per i commenti dei lemmi.

Per tutte le opzioni numeriche i valori accettati sono compresi tra **10** e **180**.

OPZIONI NELL'INSERIMENTO DI UN NUOVO TESTO

limite numero segnalazioni stesso errore + N segnalazioni massime stesso errore

Pilota l'attivazione iniziale o meno del *check box* che limita il numero di segnalazioni di errori di uno stesso tipo da parte dell'interprete dei testi e fissa il valore di questo limite. Il valore preimpostato è **10**; i valori accettati sono compresi tra **1** e **32.000**.

uso memoria

Pilota l'attivazione iniziale o meno del *check box* corrispondente.

OPZIONI IN LEMMATIZZAZIONE

N: massima distanza di ricerca in lemmatizzazione sequenziale

Limita il numero di parole entro il quale viene cercata la successiva occorrenza valida (nel caso di limitazione alle sole occorrenze lemmatizzate o alle sole non lemmatizzate) in lemmatizzazione sequenziale, nell'ambiente di lemmatizzazione per singolo testo. Il valore preimpostato è **300**; i valori accettati sono compresi tra **2** e **30.000**.

OPZIONI IN GESTIONE DEI DATI

N: livello massimo iperlemmi

Imposta il livello massimo attribuibile ad un iperlemma. Il valore preimpostato è **20**; i valori accettati sono compresi tra **0** e **99**.

APPENDICE C - CODICI UTILIZZABILI NEI FILE TESTO, NOTE, TRAD

I codici utilizzabili all'interno dei file associati di tipo **Note** e **Trad** si trovano all'interno del paragrafo **CODICI DI APPARATO**.

Notazioni usate:

- <spazio> indica uno spazio.
- nn, mm indicano numeri interi, formati da una o più cifre.
- x indica una lettera, maiuscola o minuscola.

VERSI O PROSA

Parte di testo in prosa &P

La parte di testo che segue è in prosa. Codice superfluo ad inizio testo in quanto ogni testo è assunto essere in prosa, salvo la presenza di codici di versi.

Parte di testo in versi &V

La parte di testo che segue è in versi.

Più in particolare si può precisare la modalità di conteggio dei versi. La lista completa dei codici accettati è riportata in tabella.

Codice	Modalità di conteggio dei versi
&V	numerazione per versi singoli
&V@2@	numerazione per distici
&V@3@	numerazione per terzine
&V@4@	numerazione per quartine
&V@8@	numerazione per ottave
&V@nnS@	numerazione per stanze di nn versi (nn compreso tra 1 e 99)

Numero di verso &/Vnn&

Assegna il numero del prossimo verso, sostituendosi al meccanismo di conteggio automatico (es.: per indicare che il prossimo verso è il 25, la sequenza è "&/V25&"). Si usa quando nel file vengono omessi dei versi.

Numero di verso - 1 &/Rnn&

Analogo al precedente e mantenuto per compatibilità con altre versioni di GATTO (es.: per indicare che il prossimo verso è il 25, la sequenza è "&/R24&").

Rima al mezzo &T

Questo codice viene associato alla forma immediatamente precedente e la dichiara costituire rima al mezzo.

RIFERIMENTI TOPOGRAFICI

Numero di volume o tomo, pagina e colonna \$VmmVnnx\$

- **VmmV** indica il **volume**, ad esempio **V3V** = vol. 3, **V091V** = vol. 91, ecc. Il numero di volume deve essere compreso tra 1 e 999.
- Il numero **nn**, obbligatoriamente presente, indica il **numero di pagina**.
- **x** (maiuscola o minuscola) se presente indica la **colonna nella pagina**, secondo la codifica **A**= col. 1, **B** = col. 2, e così via, con le due eccezioni **r** = "recto" **v** = "verso". Se manca, viene assunta la colonna 1.

Il massimo numero di pagina accettato è un miliardo.

Il massimo numero di riga generabile è 32.000.

Eventuali spazi inclusi tra i simboli \$ vengono ignorati.

Il numero di volume rimane invariato fino a che non ne viene incontrato uno diverso, quindi non occorre ripeterlo per ogni pagina. Per indicare che da una certa pagina in poi (ad esempio dalla pagina 35) il numero di volume non va più riportato, usare il codice **\$VmmVnny\$** con mm=0 (ad esempio, arrivati alla pagina 35, per indicare che da lì in poi il volume non va più indicato inserire il codice **\$V0V35\$**).

Numero di volume o tomo, pagina e colonna (altro formato) \$xnnx\$

(Questo formato è supportato per compatibilità con le precedenti versioni di GATTO.)

La lettera iniziale (maiuscola o minuscola), se presente, indica il volume, secondo la codifica **a** = vol.1, **B** = vol.2, ecc. Eccezione: **r** indica numero di pagina scritto in numeri romani; da visualizzare con una **r** immediatamente precedente il numero.

Il numero **nn**, obbligatoriamente presente, indica il numero di pagina.

La lettera finale (maiuscola o minuscola), se presente, indica la colonna nella pagina, secondo la codifica **A** = col.1, **b** = col.2, e così via, con le due eccezioni **r** ("recto") e **v** ("verso").

Gli spazi, se presenti all'interno di questa sequenza, vengono ignorati.

Se questa sequenza è l'unica presente in una riga, quest'ultima non conta come riga di testo.

Il massimo numero di pagina è un miliardo.

Numero di riga nella pagina &/Lnn&

Assegna il numero della prossima riga, sostituendosi al meccanismo di conteggio automatico (es.: per indicare che la prossima riga è la 25, la sequenza è "&/L25&"). Si usa quando nel file vengono omesse delle righe di testo oppure quando il riferimento organico cade all'interno di una riga di testo.

Altro esempio: il riferimento organico cade all'interno della riga 12. In questo caso le parole successive verrebbero attribuite automaticamente (ed erroneamente) al rigo seguente (13); la sequenza "&/L12&" posta all'inizio della riga successiva indica che tale riga è ancora la 12.

Numero di riga nella pagina - 1 &/Pnn&

Analogo al precedente e mantenuto per compatibilità con altre versioni di GATTO. (es.: per indicare che la prossima riga è la 25, la sequenza è "&/P24&").

RIFERIMENTI ORGANICI

Riferimento organico:

%

%%

%%%

%%%%

%%%%%

Possono comparire, con valore di codifica, solo all'inizio di una riga; il resto della riga costituisce un riferimento organico che si applica al testo che segue, fino al raggiungimento del riferimento organico successivo.

Il riferimento organico non può essere inserito all'interno di un campo formula.

Ogni occorrenza all'interno del testo deve essere collegata ad un riferimento organico; pertanto in ogni testo deve essere presente almeno un riferimento organico, collocato all'inizio e riferito in tal caso a tutto il testo.

I riferimenti organici possono essere organizzati secondo una struttura a più livelli (massimo 5). Un riferimento organico al livello più basso (livello 1) è individuato dal codice "%"; un riferimento organico a livello 2 è individuato dal codice "%%" e così via a salire fino al 5° livello (codice "%%%%%"). L'inserimento di un riferimento organico di un certo livello automaticamente sostituisce il preesistente riferimento organico di quel livello e azzerà quelli di livello inferiore (quindi dovrà essere immediatamente seguito, in genere, dai nuovi riferimenti organici di livello inferiore).

Il riferimento organico completo associato ad un contesto è formato dalla concatenazione dei riferimenti organici non nulli definiti a quel momento, a partire da quello di livello 5 fino a quello di livello 1, separati dal carattere "-". La lunghezza del riferimento organico complessivo non può superare 150 caratteri.

DIACRITICI

Cediglia ¨ (ANSI 184) oppure 6

Indicano entrambi la presenza di una cediglia sulla lettera precedente. Si applicano solo alle lettere **t** e **T**. Il codice "6" è utilizzabile solo all'interno dei corpus non impostati per accettare le cifre arabe all'interno delle forme.

Accento grave ` (ANSI 96)

Indica la presenza di un accento grave sulla lettera precedente. Si applica solo alle lettere **j**, **J**, **y** e **Y**.

Apostrofo ' (ANSI 39)

Appartiene alla parola con cui è a contatto; se è interno a dei caratteri, forma con essi un'unica parola in *notazione anglosassone* (vedere sotto), altrimenti appartiene alla parola precedente. Nelle forme polirematiche è accettato solo se queste sono in notazione anglosassone.

Notazione anglosassone &A &I

Inizio e fine di una zona di testo in notazione anglosassone. Modifica il trattamento degli apostrofi.

Trattino di unione 7 oppure - (ANSI 45) oppure _ (ANSI 95)

Fa parte della parola con la quale è a contatto; nei contesti visualizzati viene sempre rappresentato col trattino "-" (ANSI 45). Questa codifica è valida solo all'esterno dei campi formula.

La tabella seguente riassume i codici previsti per indicare il trattino di unione:

impostazione del corpus	testo normale	forme polirematiche	campi formula
le cifre non sono accettate nelle forme	7	7 oppure - (ansi 45)	(non ammesso)
le cifre sono accettate nelle forme	_ (ansi 95)	- (ansi 45)	(non ammesso)

Punto in alto · (ANSI 183) oppure &·

In entrambe le codifiche (senza e con &) indica la caduta di una consonante; nei contesti visualizzati viene sempre rappresentato col solo punto in alto "·". Questo carattere appartiene alla parola con cui è a contatto; se è interno a dei caratteri vanno distinti 2 casi: se è preceduto da & forma un'unica parola con i caratteri circostanti; altrimenti si unisce solo alla parola precedente. All'interno delle forme polirematiche è accettato solo se preceduto da &.

Punto di abbreviazione 8 oppure . (ANSI 46) oppure © (ANSI 169)

Indica la presenza di un punto di abbreviazione; appartiene alla parola con cui è a contatto e nei contesti visualizzati viene sempre rappresentato come punto ".".

La tabella seguente riassume i codici previsti per indicare il punto di abbreviazione:

impostazione del corpus	testo normale	forme polirematiche	campi formula
le cifre non sono accettate nelle forme	8	8 oppure . (ansi 46)	&. (ansi 38 e 46)
le cifre sono accettate nelle forme	© (ansi 169)	. (ansi 46)	&. (ansi 38 e 46)

(Punto sottoscritto)

È utilizzato nelle edizioni di testi in versi per la segnalazione di lettere o sillabe soprannumerarie, le quali vengono invece codificate in GATTO racchiudendole con i codici di **espunzione** "<" e ">" (vedere **CARATTERI SPECIALI**).

FORME POLIREMATICHE

Inizio e fine di forma polirematica @/ /@

Per le forme polirematiche valgono le regole seguenti:

- è ammesso il carattere spazio, alfabeticamente precedente ogni altro carattere. Lo spazio non è però ammesso a inizio e fine parola, cioè subito dopo @/ o subito prima a /@;
- il trattino di unione può essere indicato, indifferentemente, con i codici "-" o "7" (quest'ultimo solo nei corpus non impostati per accettare cifre all'interno delle forme);
- il punto di abbreviazione può essere indicato, indifferentemente, con i codici "." o "8" (quest'ultimo solo nei corpus non impostati per accettare cifre all'interno delle forme): ne consegue che il carattere *punto* è sempre interpretato come abbreviazione, mai come interpunzione; il punto debole "&." non è ammesso;
- il punto in alto è accettato solo nella codifica "&." (altrimenti spezzerebbe la forma, il che sarebbe in conflitto col concetto di polirematica);
- se la polirematica deve essere trattata in notazione anglosassone, i codici relativi a questa notazione, ovvero &A e &I, vanno collocati esternamente ai codici di inizio e fine polirematica: un esempio di codifica corretta è &A@/Bartolomeo di Bonaccorso/@&I;
- l'apice (o apostrofo) è ammesso solo se la polirematica è inserita in una parte di testo in notazione anglosassone (altrimenti spezzerebbe la forma, il che sarebbe in conflitto col concetto di polirematica);
- una forma polirematica non può essere inserita all'interno di un campo formula;
- una forma polirematica non può essere ripartita su due (o più) righe di testo.

INTERPUNZIONI

I segni di interpunzione possono essere deboli, forti o definibili dall'utente. Una sequenza di più segni di interpunzione è forte se contiene almeno un segno di interpunzione forte. Un segno di interpunzione debole è solo visualizzato, senza altri effetti; uno forte costituisce separatore tra periodi ed entra in gioco nel delimitare i contesti quando questi sono dimensionati in base al numero dei periodi. La maggior parte dei segni di interpunzione è ricercabile. La ricerca di uno o più specifici segni di interpunzione localizza tutte le sequenze che includono i caratteri cercati. con l'eccezione dei puntini di sospensione le interpunzioni operano allo stesso modo nel testo indicizzabile e nei campi formula.

Caratteri di interpunzione

Punteggiatura forte:	. ; :
Punteggiatura debole o forte:	! ? ...
Punteggiatura debole (cercabili):	" , - « »
Punteggiatura debole (non cercabili):	()

Apice &"

Nei contesti visualizzati viene sempre rappresentato con l'apice " ' " (ANSI 39). Questa codifica indica un apice inteso come segno di interpunzione.

Punto debole &.

Punto (interpunzione) non avente funzione di punteggiatura forte; può trattarsi di un punto di abbreviazione in un campo formula, di un elemento di numerazione dei riferimenti organici, ecc.. Questa codifica vale comunque in tutte le parti del testo, interne o esterne ai campi formula. Nei contesti visualizzati viene sempre rappresentato come punto ".". A differenza del punto forte, il punto debole è accettato anche prima della prima parola di un riferimento organico.

Punti di lacuna (vedere Puntini di sospensione)

Puntini di sospensione

Sono rappresentati da una successione di 2 o più punti consecutivi (non preceduti da &). Possono costituire punteggiatura forte o debole, a seconda dell'impostazione scelta al momento della creazione del corpus e a seconda del loro inserimento o meno all'interno di parentesi quadre. Sono possibili 3 casi:

- **puntini di sospensione nel testo.** Sono considerati punteggiatura forte o debole a seconda della scelta operata all'atto della creazione del nuovo corpus (il fatto che i puntini siano o meno racchiusi tra parentesi quadre è ininfluyente);
- **puntini di sospensione in campo formula #...@ o in campo parlante \...\.** Sono sempre e comunque punteggiatura debole (il fatto che i puntini siano o meno racchiusi tra parentesi quadre è ininfluyente);
- **puntini di sospensione in campo formula &L-...&L+.** Se sono privi di parentesi quadre sono considerati punteggiatura forte o debole a seconda della scelta operata all'atto della creazione del nuovo corpus; se invece sono racchiusi tra parentesi quadre sono comunque considerati punteggiatura debole.

CARATTERI SPECIALI

&(&)

Nella notazione adottata all'OVI sono utilizzati come apertura e chiusura di **scioglimento interno**. La stringa racchiusa tra queste sequenze appartiene alla forma. La presenza di & davanti alle parentesi tonde evita che la parola venga spezzata; la & non è obbligatoria all'interno dei campi formula, dove le parentesi anche se interne a una parola non la spezzano comunque.

< >

Nella notazione adottata all'OVI sono utilizzati come apertura e chiusura di **espunzione esterna**. Le parole racchiuse tra queste sequenze vengono regolarmente indicizzate; all'OVI la funzione espuntiva è stata soppressa poiché con lo stesso codice sono state indicate le sillabe soprannumerarie nei testi in versi.

&< &>

Nella notazione adottata all'OVI sono utilizzati come apertura e chiusura di **espunzione interna**. La presenza di & davanti alle parentesi angolari evita che la parola venga spezzata. La stringa racchiusa tra queste sequenze appartiene alla forma; all'OVI la funzione espuntiva è stata soppressa poiché con lo stesso codice sono state indicate le sillabe soprannumerarie nei testi in prosa.

[]

Nella notazione adottata all'OVI sono utilizzati come apertura e chiusura di **integrazione esterna**. Le parole racchiuse tra queste sequenze vengono regolarmente indicizzate.

&[&]

Nella notazione adottata all'OVI sono utilizzati come apertura e chiusura di **integrazione interna**. La presenza di & davanti alle parentesi quadre evita che la parola venga spezzata. La stringa racchiusa tra queste sequenze appartiene alla forma.

&C &c

Apertura e chiusura di **corsivo esterno**.

&K &k

Apertura e chiusura di **corsivo interno**.

CAMPI FORMULA

Racchiudono sequenze di caratteri particolari (numerici ecc..) e/o forme non indicizzate. Non è ammessa la presenza di campi formula innestati o interlacciati. L'unica differenza di comportamento tra i due tipi di campo formula previsti riguarda il trattamento dei puntini di sospensione (vedere **INTERPUNZIONI**).

Campo formula # @

Apertura e chiusura di un campo formula.

Campo formula &L- &L+

Apertura e chiusura di un campo formula.

CAMPI SPECIALI

Citazione }| |{

Nella notazione adottata all'OVI sono utilizzati come **inizio e fine citazione**. Il primo codice è formato dai caratteri ANSI 125 e 124; il secondo dagli ANSI 124 e 123. Le parole racchiuse tra queste sequenze non vengono indicizzate.

Parlante \ \

Nella notazione adottata all'OVI sono utilizzati come apertura o chiusura di un **parlante**. Questo campo può trovarsi all'interno del testo e in questo caso il suo contenuto viene indicizzato, oppure in un campo formula con le sintassi #\...\@ oppure &L-\...\&L+, nei quali casi il suo contenuto non viene indicizzato; questo campo non può invece trovarsi a cavallo di un estremo di un campo formula. Occorrenze del carattere "\" all'interno di campi formula ma in posizioni diverse dagli estremi vengono interpretate come presenze di un semplice carattere di interpunzione. All'interno del campo *parlante* non può trovarsi un campo formula (quindi non sono ammesse notazioni del tipo \...\#\...\@\...\ oppure \...\&L-...\&L+...\).

CODICI DI APPARATO

Consentono di associare delle note a singole parole o parti di testo, oppure di visualizzare in parallelo i contesti tratti da due testi diversi¹³, in presenza di traduzioni o diverse edizioni. La nota o il contesto parallelo saranno mostrati ogni qualvolta verrà localizzata, a seguito di una ricerca, una delle parole del *testo base* incluse nella zona interessata alla nota. Nel caso di traduzioni o diverse edizioni le parti di testo corrispondenti nelle due opere sono indicate col termine *brani*.

@@NT

Costituisce la prima riga di un testo base che prevede al suo interno la presenza di note.

@@TR

Costituisce la prima riga di un testo base che prevede la presenza di un testo associato contenente una traduzione o una diversa edizione.

@@A2

Costituisce la prima riga di un testo base che prevede la presenza sia di note al suo interno che di un testo associato contenente una traduzione o una diversa edizione.

Inizio di una nota *nn

Nel *testo base*, inizio della zona di testo cui è associata la nota numero **nn**.

Nel *testo associato di tipo Note*, inizio della nota numero **nn**.

Fine di una nota */

Nel *testo base*, fine della zona di testo cui è associata una nota.

Nel *testo associato di tipo Note*, fine di una nota.

Separatore fra brani *

Nel *testo base* e nel *testo associato di tipo Trad*, separatore tra *brani* consecutivi.

LEMMI E IPERLEMMI

I codici di lemma e iperlemma sono inseriti nei file e gestiti completamente da GATTO. Si suggerisce di non modificarli. Per maggiori informazioni vedere ad esempio la Lezione 50.

¹³ Detti *testo base* e *testo associato di tipo Trad*; nel caso di note, il *testo associato di tipo Note* conterrà invece queste ultime.

Codice di lemma =nn=

La parola immediatamente successiva a questa sequenza è lemmatizzata col lemma di codice (relativo al testo) **nn**.

Codice di lemma più iperlemma =nn,mm=

La parola immediatamente successiva a questa sequenza è lemmatizzata col lemma di codice **nn** (relativo al testo) e ad essa è anche associato l'iperlemma di codice **mm** (sempre riferito al testo).

SEQUENZE DA EVITARE

Le seguenti 4 sequenze, formate ciascuna da 3 caratteri, non sono accettate nei testi in quanto genererebbero nei contesti visualizzati la stringa <spazio> + "/" + <spazio>, la quale potrebbe erroneamente essere intesa come indicatore di fine verso.

<spazio>	+	"/"	+	<spazio>
<spazio>	+	"/"	+	"@"
"#"	+	"/"	+	<spazio>
"#"	+	"/"	+	"@"

APPENDICE D - AVVIO DI GATTO CON PARAMETRI

CASI IN CUI UTILIZZARE QUESTO TIPO DI AVVIO

Questa modalità di avvio del programma è utile in due casi: schedulazione di GATTO da parte di un'altra procedura oppure avvio da linea comando, ovvero dalla finestra di comando DOS presente negli ambienti Windows.

In entrambi i casi GATTO opererà esclusivamente in modalità interrogazione e solamente sul corpus specificato nei parametri; fermi restando questi due vincoli una volta 'entrati' nel programma questo funzionerà normalmente, consentendo di accedere a tutte le modalità di ricerca normalmente disponibili, con tutte le opzioni previste per tali ricerche. Si potranno effettuare più ricerche e anche quella iniziale, ovvero quella impostata con i parametri di avvio, potrà essere modificata prima di essere eseguita.

PARAMETRI DA FORNIRE

- 1 lettera identificativa del disco nel quale si trova il corpus da interrogare (es. "C"),
- 2 nome del corpus da interrogare (es. "demo"),
- 3 codice che denota il tipo di ricerca iniziale da effettuare (es. "f" per ricerca per forme),
- 4 elementi da cercare (es. "casa" per cercare tale forma).

CODICI IDENTIFICATIVI DEI TIPI DI RICERCHE

- c ricerca per categorie grammaticali
- cl generazione di concordanze lemmatizzate
- coo ricerca di cooccorrenze
- d ricerca per disambiguatori
- f ricerca per forme
- i ricerca per iperlemmi
- ilf generazione di indices locorum per forme
- ill generazione di indices locorum lemmatizzati
- l ricerca per lemmi

SINTASSI DEL COMANDO

Il comando è formato dal nome del programma ("**gatto**" o "**gatto.exe**"), seguito da uno o più spazi e poi dalla lista dei parametri, separati tra loro da virgole. Spazi immediatamente precedenti o successivi alle virgole sono ininfluenti.

Nel caso di ricerche di cooccorrenze gli elementi da cercare vanno scritti separati tra loro da spazi, non da virgole, esattamente come quando vengono digitati nella finestra di avvio ricerche di cooccorrenze.

I parametri non possono includere il carattere virgola al loro interno. Sono ammessi i parametri polirematici, da digitare senza accorgimenti di sorta.

Tutti i parametri possono essere scritti indifferentemente in lettere minuscole o maiuscole, ad eccezione degli elementi da cercare nelle cooccorrenze, per i quali vanno rispettate le regole specifiche relative all'uso di maiuscole e minuscole.

Il numero di elementi da cercare deve essere compreso tra 1 e 10; nel caso di ricerche per forme il numero massimo è invece sostanzialmente illimitato. Nel caso che gli elementi siano più di 10 la ricerca per forme verrà iniziata a partire dalla finestra di *ricerca da un brano*.

ESEMPI DI CHIAMATE A GATTO

gatto c,demo,f,casa

Schedula GATTO per interrogare il corpus c:\demo. Viene cercata la **forma casa**.

gatto.exe d,demo,l,andare, ab eternu, venire

Il corpus interrogato è **d:\demo**. Vengono cercati i **lemmi andare e venire** e il lemma polirematico **ab eternu**.

gatto c,demo,i, a*

In **c:\demo** vengono cercati tutti gli **iperlemmi** che iniziano per **a**.

gatto c, demo, coo, Ltornare "per altra via"

In **c:\demo** vengono cercate le **cooccorrenze** di occorrenze lemmatizzate col lemma **tornare** in prossimità della stringa **"per altra via"**.

COME EFFETTUARE UNA CHIAMATA COMPLETA DALLA FINESTRA DOS

Supponiamo che GATTO sia stato installato nella cartella **C:\Programmi\Gatto**; se GATTO è installato in una cartella e/o un disco diverso, le istruzioni che seguono vanno modificate di conseguenza.

- 1** aprire la finestra DOS (ad esempio dal menu **Start => Programmi => Accessori => Prompt dei comandi**).
- 2** se il prompt è **C:\>** (cioè siamo già nel disco C), portarsi nella cartella che contiene **Gatto.exe** con il comando **cd Programmi\Gatto**
- 3** se invece il prompt presenta una lettera diversa da C, dare il comando **cd /D c:\Programmi\Gatto**
- 4** infine avviare GATTO, ad esempio con **gatto c,demo,f,casa**

APPENDICE E - RISOLUZIONE DI PROBLEMI PARTICOLARI

CIFRATURA DEI DATI SCAMBIATA PER UN VIRUS

Sia pure in casi estremamente infrequenti (dell'ordine di una volta su alcune migliaia di testi caricati) può capitare che una delle sequenze di caratteri generate dalla cifratura dei testi sia scambiata per un virus da un programma antivirus.

Per rimuovere questa situazione per prima cosa osservare il nome e l'*estensione* (ovvero i caratteri dopo il punto) presente nel nome del file sospettato di includere un virus (ricordarsi di configurare *Gestione risorse* di Windows in modo da visualizzare le estensioni dei file: vedere capitolo 2): uno di questi due dati includerà la sigla del testo cui il file si riferisce.

A quel punto cancellare il testo dal corpus (Lezione 48) e inserirlo nuovamente (Lezione 51): GATTO provvederà a cambiare la cifratura rimuovendo l'occasionale coincidenza con la firma virale individuata dall'antivirus.

Se il problema permane ricostruire il corpus cambiando l'ordine di inserimento dei testi oppure cancellare il testo che genera il problema, alterarlo lievemente, magari inserendo un secondo spazio tra una parola e la successiva (un intervento ogni 2 o 3 righe, o anche ogni riga) e quindi reinserirlo nel corpus così modificato.

ARRESTO ANOMALO DI GATTO PER DIFETTO DEL PROGRAMMA DURANTE LE OPERAZIONI DI LEMMATIZZAZIONE O IPERLEMMATIZZAZIONE

Per quanto il programma sia stato e continui ad essere testato, è possibile che nel corso delle operazioni di lemmatizzazione o iperlemmatizzazione si manifesti un qualche errore non ancora rilevato e rimosso.

Questo evento può danneggiare gli archivi del corpus in uso in quel momento.

Per ripristinarne l'integrità, effettuare le operazioni seguenti:

1. se il problema è avvenuto in lemmatizzazione *per singolo testo*, estrarre il testo dal corpus (Lezione 49) prendendo nota degli eventuali messaggi riportati, del tipo "**attenzione: lemma (o iperlemma) perso. Codice del lemma (o iperlemma) = N (rimuoverne i riferimenti dal file filgat)**". In questo messaggio N indica il numero del lemma (o iperlemma) perso. Supponiamo che si sia perso il lemma numero 5: in tal caso rimuovere completamente dal file estratto **filgat.<sigla>.txt** (dove <sigla> indica la sigla del testo in oggetto), con un editore qualunque, tutti i riferimenti del tipo "...=5=..." o "...=5,M=..." (nel secondo caso M indica un qualunque numero di iperlemma). Supponiamo invece che si sia perso l'iperlemma numero 7: analogamente, rimuovere dal file estratto **filgat.<sigla>.txt** tutti i riferimenti del tipo "...=M,7=..." (dove M indica un qualunque numero di lemma) o trasformarli in "...=M=...", lasciandovi cioè solo il lemma senza l'iperlemma. A questo punto trasferire i file **filgat.<sigla>.txt** e **codlemmi.<sigla>**, ottenuti con l'*estrazione* dei testi, dalla sottodirectory **out** del corpus alla *directory dei testi* sostituendoli a quelli già esistenti e procedere alla sostituzione dei testi nel corpus (Lezione 51).
Attenzione: ulteriori segnalazioni di errore possono comparire anche in questa fase. Correggere i file **filgat.<sigla>.txt** sulla base delle indicazioni fornite e poi, per i testi coinvolti, ripetere la procedura di sostituzione.
2. se invece il problema si è manifestato in lemmatizzazione sul corpus, eseguire le operazioni sopra descritte su tutti i testi del corpus, estraendoli in blocco per fare prima (Lezione 54).

ARRESTO ANOMALO DI GATTO PER CAUSA ESTERNA DURANTE LA LEMMATIZZAZIONE PER SINGOLO TESTO

Durante la lemmatizzazione gli archivi sono aperti non solo in lettura ma anche in scrittura. Se durante questa fase di lavoro si verificasse un'interruzione nel funzionamento del PC, per mancanza di corrente o altra causa esterna, gli archivi di lemmatizzazione potrebbero risultarne danneggiati; in particolare si potrebbe ottenere un errore non eliminabile ogni volta che si tornasse sulla specifica occorrenza in corso di lemmatizzazione al momento dell'interruzione.

La procedura di recupero degli archivi è la seguente:

1. prendere nota del testo cui apparteneva l'occorrenza che si stava lemmatizzando al momento dell'interruzione,
2. *estrarre* (non cancellare) dal corpus il testo in oggetto, seguendo la procedura descritta nella Lezione 49,
3. a questo punto cancellare il testo dal corpus (Lezione 48) ignorando l'eventuale messaggio di avvertimento dato alla fine,
4. trasferire i file **filgat.<sigla>.txt** e **codlemmi.<sigla>** (dove <sigla> indica la sigla del testo in oggetto) ottenuti con l'*estrazione* del testo, dalla sottodirectory **out** del corpus alla *directory dei testi* sostituendoli a quelli già esistenti,
5. reinsertire il testo nel corpus (Lezione 51): se viene segnalato un errore portarsi con un editore nel file **filgat.<sigla>.txt** nel punto in cui è localizzato l'errore (che molto probabilmente sarà **codice di lemma associato a un lemma inesistente** o qualcosa del genere) e rimuovere il codice =nnn= o =nnn,mmm= cui l'errore si riferisce (nnn e mmm indicano numeri interi), poi ripetere l'inserimento del testo.

In questo modo il testo viene recuperato integralmente, completo della sua lemmatizzazione, con rimozione della sola lemmatizzazione (danneggiata) attribuita alla occorrenza cui si stava lavorando al momento dell'interruzione anomala.

ARRESTO ANOMALO DI GATTO PER CAUSA ESTERNA DURANTE LA LEMMATIZZAZIONE SUL CORPUS OD OPERANDO NELL'AMBIENTE GESTIONE BASE DATI

Questo evento può danneggiare in modo irreparabile gli archivi del corpus in uso in quel momento. Se si dispone di una copia di riserva, conviene sostituire il corpus con quest'ultima.

In caso contrario:

1. estrarre dal corpus (Lezione 49) i testi alla cui lemmatizzazione siano state apportate modifiche dopo il loro inserimento nel corpus stesso (estrarli uno per volta, in modo che eventuali *crash* del programma, dovuti agli archivi danneggiati, inibiscano l'operazione su un solo testo alla volta),
2. trasferire i file **filgat.<sigla>.txt** e **codlemmi.<sigla>** (dove <sigla> indica la sigla del testo in oggetto) ottenuti con l'*estrazione* dei testi, dalla sottodirectory **out** del corpus alla *directory dei testi* sostituendoli a quelli già esistenti,
3. cancellare il corpus e ricostruirlo (capitoli 24 e 7),
4. inserire nel nuovo corpus tutti i testi, lemmatizzati e non, utilizzando il contenuto della *directory dei testi* (Lezione 51).

ARRESTO ANOMALO DI GATTO PER CAUSA ESTERNA DURANTE UN'OPERAZIONE SUI TESTI (INSERIMENTO, SOSTITUZIONE, CANCELLAZIONE O ALTRA OPERAZIONE), CON DANNEGGIAMENTO DEGLI INDICI

Questo evento impedisce la ricostruzione degli indici del corpus, che vengono temporaneamente rimossi durante queste operazioni. Il risultato è che il corpus potrebbe diventare inutilizzabile e mandare in crash il programma con messaggi del tipo "...is'nt an index...".

Si può ripristinare l'integrità del corpus con la procedura seguente:

1. entrare nell'ambiente **Gestione base dati**,
2. cliccare su [CORPUS],
3. selezionare il corpus da ripristinare cliccando sulla riga corrispondente,
4. cliccare su [GESTIONE.. + RIPRISTINA INDICI].

Se anche dopo questa operazione il corpus continuasse a presentare dei problemi, ricostruirlo seguendo la procedura descritta nel paragrafo precedente "**Arresto anomalo di GATTO per causa esterna durante la lemmatizzazione sul corpus od operando nell'ambiente Gestione base dati**".

APPENDICE F - GLOSSARIO

accumulatore

Tavola che raccoglie in modo cumulativo tutte le forme, i lemmi o le coppie forma-lemma localizzati nel corpus (o in sottocorpus) in quanto soddisfacenti le ricerche effettuate su di esso.

ambiente

Ognuna delle parti in cui è logicamente strutturato il programma GATTO. Ogni *ambiente* contiene l'insieme di finestre e di istruzioni necessarie per svolgere una certa categoria di operazioni. Gli *ambienti* sono quattro: **ricerche**, **lemmatizzazione sul corpus**, **lemmatizzazione per singolo testo**, **gestione base dati**.

ansi

American National Standard Institute.

Set di 256 caratteri riproducenti lettere maiuscole e minuscole con e senza accenti, numeri, segni di interpunzione e alcuni simboli speciali. I primi 128 caratteri sono uguali ai corrispondenti caratteri *ASCII*.

area generica

Sigla che denota un ambito linguistico esteso; nel sistema adottato all'OVI si tratta di regola di un ambito regionale.

area specifica

Sigla che denota un ambito linguistico ristretto ad un preciso luogo di origine del testo o della sua lingua (città o paese).

bottone

Elemento grafico presente in una finestra e caratterizzato da una scritta descrittiva dell'azione che verrà svolta dal programma cliccando sul *bottone* stesso col tasto sinistro del mouse (sinonimo di *pulsante*).

box a tendina

Elemento grafico normalmente composto da un riquadro e una freccia in basso, posta sulla parte destra: cliccando la freccia apparirà una lista di dati alternativi tra i quali scegliere, cliccandolo, quello da inserire nel riquadro, come dato *selezionato*. Una volta effettuata la selezione, la lista di dati del *box a tendina* si richiude e resta visibile il solo dato selezionato.

campo formula

Parte di testo non indicizzata e perciò non raggiungibile dalle ricerche effettuate col programma.

Sono in *campo formula*, ad es.:

- i numeri arabi e romani, forme frammentarie (nel caso in cui le edizioni portino a testo parole parzialmente illeggibili del manoscritto, per es. "cav..."), annotazioni che si vogliono rendere visibili a chi esamina un contesto,
- brani in lingua non italiana.

Un campo formula è racchiuso tra i codici #...@ oppure &L-...&L+.

carattere jolly

Carattere che, inserito in una parola, indica la presenza non di se stesso ma di altri caratteri. In particolare ? sta per 'un carattere qualunque'; * sta per 'un qualunque gruppo di caratteri o nessun carattere'.

cartella

Parte di un disco fisso, di un floppy disc, di un CD-Rom o di un DVD, logicamente separata dalle altre e contenente file, documenti o programmi tra loro connessi da vincoli di attinenza. A sua volta il contenuto di una *cartella* può essere suddiviso, in base a criteri di convenienza, in *sottocartelle*. Sinonimo di *directory*.

cartella dei testi

Cartella destinata a contenere file testo, file associati di tipo Note o di tipo Trad e file di lemmatizzazione relativi ai testi da includere in un corpus.

Ogni corpus ha una *cartella dei testi* assegnatagli al momento della sua creazione. Una stessa *cartella dei testi* può essere utilizzata per più corpus e può anche contenere file non connessi con GATTO. Sinonimo di *directory dei testi*.

check box

Riquadro che permette di attivare o disattivare una certa opzione del programma effettuando un singolo clic al suo interno. Un *check box* è attivato quando visualizza al suo interno una \surd . Lo stato di attivazione di ogni *check box* è indipendente da quello di altri eventualmente presenti nella stessa finestra.

citazione

Espunge dall'indicizzazione parole non appartenenti all'autore del testo, ma da lui soltanto riferite.

contesto

Porzione di testo variabile contenente una occorrenza della forma cercata o analizzata.

contesto corrente

Negli ambiente di lemmatizzazione, contesto cui si applicheranno le successive operazioni di lemmatizzazione.

cooccorrenza

Presenza contemporanea di più elementi di testo, occorrenze e/o segni di interpunzione, entro una zona delimitata con precisione (*intervallo di ricerca*).

corpus

L'insieme dei testi che costituiscono la tavola dei testi citati.

corpus in uso

Corpus al quale si applicheranno tutte le successive azioni eseguite con GATTO.

data codificata

La data codificata è una stringa di caratteri associata al testo, generata sulla base della datazione attribuitagli e di altri elementi ad essa connessi. I contesti risultanti dalle ricerche vengono ordinati inizialmente in base dell'ordinamento alfabetico delle date codificate attribuite ai testi di appartenenza.

default

Nel caso di possibilità di scelta tra più opzioni alternative, questo termine identifica la scelta impostata automaticamente dal programma in assenza di una indicazione esplicita da parte dell'utente.

diacritico

Carattere convenzionale, presente nelle liste testo, avente uno speciale significato.

directory

(vedi *cartella*).

directory dei testi

(vedi *cartella dei testi*).

disambiguatore

All'interno di GATTO indica un attributo, di uso facoltativo, che permette di distinguere tra loro lemmi denotati dalla stessa entrata lessicale e dalla stessa categoria grammaticale, oppure iperlemmi aventi lo stesso assetto grafico e livello.

dizionario di macchina

Lista dei lemmi associati ad una forma nell'ambito dell'intero corpus.

dominio di ricerca

Insieme di testi entro il quale viene effettuata una ricerca. Può coincidere con l'intero corpus, con un sottocorpus o con una combinazione di più sottocorpus.

espunzione

Porzione di testo all'interno di una parola (*espunzione interna*) o comprendente una o più parole (*espunzione esterna*) che l'editore ha stampato, segnalando che però essa non appartiene al testo, o a suo giudizio o perché lo stesso manoscritto la presenta espunta (i due casi non vengono distinti).

file di lemmatizzazione

File che contiene la lemmatizzazione associata ad un testo. Ha nome codlemmi.xxx.txt, dove xxx è la sigla del testo cui il file si riferisce. Va inserito nella *directory dei testi* del corpus in cui si intende inserirlo.

file testo

Il testo a stampa, inserito in un file, codificato opportunamente per servire a tutte le operazioni di ricerca e lemmatizzazione previste in GATTO (sinonimo di *testo base*). Ha nome filgat.xxx.txt, dove xxx è la sigla del testo cui il file si riferisce. Va inserito nella *directory dei testi* del corpus in cui si intende inserirlo.

forma (parola)

Ogni singola parola, distinta dalle altre esclusivamente in base all'assetto grafico, che può comparire un qualsiasi numero di volte (*occorrenze*) nel corpus o in un testo.

forma (riferita ad un testo)

Dato bibliografico associabile ad un testo. I codici utilizzati all'OVI sono:

M => testo misto versi/prosa.

P => testo in prosa.

V => testo in versi.

forma esclusiva

Dicitura riferita al *dominio di ricerca* corrente. Indica una forma utilizzata solo al suo interno, nel senso che non compare nei testi del corpus non appartenenti al *dominio di ricerca*.

forma polirematica

Forma composta da più parole separate da spazi e considerata come un tutto unico, sia in ricerca che in lemmatizzazione.

formario

Elenco di tutte le forme diverse presenti nei testi del corpus.

formario esclusivo

Insieme di tutte le *forme esclusive* presenti nel *dominio di ricerca*.

genere

Sigla che connota un'opera come appartenente ad un determinato genere letterario o di scrittura.

integrazione

Parte di testo comprendente uno o più caratteri all'interno di una parola (*integrazione interna*) o una o più parole intere (*integrazione esterna*) che l'editore segnala come sua aggiunta al manoscritto.

interrogazione

(vedi *ricerca*).

intervallo di ricerca

Zona del testo entro la quale devono essere racchiuse delle occorrenze per essere prese in considerazione nella formazione di una cooccorrenza; l'*intervallo di ricerca* può essere definito fissandone le dimensioni in numero di parole, facendone coincidere gli estremi con i limiti di un periodo o combinando entrambi i vincoli.

iperlemma

Elemento che può essere messo in connessione con più lemmi, consentendone la localizzazione simultanea in fase di ricerca. Data la struttura gerarchica nella quale possono essere collocati gli iperlemmi, ognuno di loro, anziché con dei lemmi, può essere collegato con altri iperlemmi *di livello inferiore*; un iperlemma può inoltre essere direttamente associato a occorrenze specifiche di forme specifiche.

iperlemmario

Elenco di tutti gli iperlemmi utilizzati nel corpus.

kwic

(KeyWord In Context): i contesti sono presentati con la forma cercata disposta al centro della riga di lettura.

lemma

La forma, presente nei testi o ricostruita, cui viene riportato un insieme di forme che si distinguono fra loro soltanto per l'assetto grafico (varianti grafiche, con o senza valore di varianti fonetiche) e/o perché sono forme della flessione dello stesso verbo o sostantivo o aggettivo. Corrisponde di regola, ma non necessariamente, ad un'entrata del vocabolario.

lemma esclusivo

Dicitura riferita al *dominio di ricerca* corrente. Indica un lemma utilizzato solo al suo interno, nel senso che non compare nei testi del corpus non appartenenti al *dominio di ricerca*.

lemma impostato

È il lemma che compare nel riquadro celeste nella testata delle finestre di lemmatizzazione. Utilizzabile per accelerare le operazioni di lemmatizzazione, può avere origini diverse:

- al termine di ogni operazione di lemmatizzazione il lemma utilizzato viene assegnato anche al *lemma impostato*,
- può essere ridefinito manualmente, sulla base del lemma associato ad una occorrenza di una forma,
- viene riassegnato ogni volta che il *dizionario di macchina* viene aperto o viene comunque riferito ad una forma diversa, purché siano presenti dei lemmi nel *dizionario di macchina* di quest'ultima.

lemmario

Elenco di tutti i lemmi emersi dallo spoglio dei testi del corpus.

lemmario esclusivo

Insieme di tutti i *lemmi esclusivi* presenti nel *dominio di ricerca*.

lemmatizzazione

Operazione che associa a una occorrenza di una forma un lemma ed eventualmente un iperlemma.

notazione anglosassone

Codifica che mantiene unite in un'unica forma due stringhe separate da un apostrofo. In sua assenza le due stringhe danno origine a due parole separate, con l'apostrofo appartenente alla prima.

occorrenza

Ogni singola parola di un testo, intendendo per *parola* una stringa di caratteri delimitata a sinistra e a destra da uno spazio¹⁴ o da un carattere di interpunzione.

¹⁴ Fanno eccezione a questa definizione le forme polirematiche, le cui parole componenti sono separate da spazi.

Intesa anche come *evento*, nel senso di singola comparsa in un testo dell'elemento linguistico *forma*.

option box

Finestrina circolare che permette di selezionare un'opzione del programma, scegliendo tra più opzioni alternative, ognuna controllata da uno specifico *option box*; gli *option box* tra i quali va operata una scelta mutuamente esclusiva sono racchiusi, ed eventualmente separati da altri *option box* presenti nella stessa finestra, da appositi pannelli o riquadri. Per attivare l'opzione associata ad un *option box*, disattivando automaticamente le opzioni alternative, fare clic su di esso. Un *option box* è attivato quando evidenzia al suo interno un dischetto nero.

parola

Stringa di caratteri inclusa in un testo e delimitata a sinistra e a destra da uno spazio o da un carattere di interpunzione.

pseudocarattere

Carattere diacritico appartenente ad una forma. Se presente, determina un ordinamento alfabetico tra forme che, in sua assenza, risulterebbero uguali. Non agisce invece sull'ordinamento alfabetico tra forme che in sua assenza risulterebbero diverse.

pulsante

(vedi *bottone*).

punteggiatura debole

Segno di interpunzione che non costituisce separatore tra periodi.

Sono segni di punteggiatura debole " « » , - () . (punto debole)

Possono essere definiti punteggiatura debole ! ? ...

punteggiatura forte

Segno di interpunzione che costituisce separatore tra periodi.

Sono segni di punteggiatura forte . : ;

Possono essere definiti punteggiatura forte ! ? ...

punto di abbreviazione

Un punto che segnala un'abbreviazione o che sta per un carattere mancante dal testo, distinto da quello che ha valore di segno di interpunzione.

reverse

Evidenziazione grafica di un elemento selezionato all'interno di un insieme, tipicamente una riga di una griglia. L'elemento selezionato appare in genere visualizzato in bianco su sfondo blu.

ricerca

Operazione che richiede a GATTO una informazione specifica: lemmi o insiemi di lemmi, corredate o meno dalle rispettive forme; forme o insiemi di forme, corredate o meno dalle rispettive occorrenze; iperlemmi o insiemi di iperlemmi corredate o meno degli elementi associati; cooccorrenze; altre informazioni su corpus, sottocorpus, singoli testi (sinonimo di *interrogazione*).

riferimento organico

Stringa, generalmente abbreviata, che corrisponde alla ripartizione del testo (capitolo, paragrafo, ecc.). È costituita dalla concatenazione dei riferimenti organici non nulli, dei vari livelli, definiti a quel momento, a partire dal riferimento organico di livello 5 fino a quello di livello 1, separati da singoli caratteri "-".

riferimento topografico

Dato che permette di individuare precisamente la collocazione di una parola all'interno del testo di riferimento a stampa. È costituito da: volume (se multiplo), numero di pagina, numero di riga nella pagina, numero di colonna (se la pagina contiene più colonne).

rima al mezzo

Codifica attribuita a parole interne al verso che rimano con parole alla fine o all'interno di altri versi.

rtf

Rich Text Format.

Codifica standard che integra il testo vero e proprio con codici atti a memorizzare le istruzioni di formattazione, consentendo la lettura ed il recupero integrale di un documento completo di tutte le caratteristiche tipografiche da parte di qualunque programma in grado di interpretare questa codifica.

scioglimento

Traduzione in chiaro di segni tachigrafici o di abbreviazioni presenti nel manoscritto segnalata dall'editore.

selettore

Riquadro destinato a contenere una forma, un lemma, una categoria grammaticale, un iperlemma, un disambiguatore o un gruppo di tali elementi per indicarlo come obiettivo da localizzare in una ricerca.

selettore multiplo

Insieme di due o più *selettori* consecutivi, collegati tra di loro per imporre simultaneamente più condizioni allo stesso elemento oggetto di una ricerca.

selettore semplice

Usato come sinonimo di *selettore*, in contrapposizione a *selettore multiplo*.

selettore virtuale

Le finestre di ricerca mostrano sempre 10 selettori. Questo numero individua quindi il numero massimo di elementi o gruppi di elementi ricercabili esplicitamente nel corso di una stessa interrogazione. Alcune opzioni, come **Varianti**, o l'uso delle sequenze [...] o <...>, generano ulteriori selettori, non visualizzati, che ampliano di fatto notevolmente questo limite: il termine *selettore virtuale* indica ognuno dei selettori generati in questo modo.

sequenza jolly

Gruppo di caratteri che, inserito in una parola, indica la presenza non di se stesso ma di altri caratteri. In particolare:

[x-z]	sta per <i>una lettera qualunque tra x e z (comprese)</i> ;
[ack]	sta per <i>a, oppure c oppure k</i> ;
<str1,str2,str3,...>	sta per <i>la stringa str1 o la stringa str2 o la stringa str3 o ...</i> (una stringa può anche essere la stringa nulla). Per stringa si intende una qualunque successione di caratteri alfanumerici.

sigla

Stringa formata da uno a tre caratteri alfanumerici che identifica univocamente un testo.

sottocorpus

Sottoinsieme dei testi inclusi nel corpus, utilizzabile per effettuare ricerche su porzioni limitate di esso. È possibile definire contemporaneamente fino a 6 sottocorpus ed utilizzarne una qualunque combinazione.

sottodirectory

Una *directory* logicamente inserita a sua volta in una *directory* gerarchicamente superiore, come suo sottoinsieme.

tabella degli omografi

Lista dei lemmi presenti nel *lemmario* e omografi del lemma proposto. L'uso della tabella è riservato alla lemmatizzazione.

tabella delle varianti

Tabella associativa, valida a livello di installazione di GATTO (quindi utilizzabile per tutti i corpus), nella quale vengono inserite delle coppie di forme, dette rispettivamente *forma base* e *variante*. Una stessa *forma base* può far parte di più coppie, in associazione con più *varianti*, così come una stessa *variante* può comparire in associazione con più *forme base*.

Quando, durante una ricerca per forme, in un *selettore* è inserita una delle forme che compare come *forma base* in questa tabella ed è attiva per quel selettore l'opzione **vAr**, vengono cercate non solo la *forma base* ma anche tutte le eventuali *varianti* che risultino ad essa associate nella *tabella delle varianti*.

testo associato di tipo Note

File contenente il corpo delle note inserite in un testo. Ha nome *note.xxx.txt*, dove *xxx* è la sigla del testo cui il file si riferisce. Va inserito nella *directory dei testi* del corpus.

testo associato di tipo Trad

File contenente una traduzione o una diversa edizione di un testo base. Ha nome *trad.xxx.txt*, dove *xxx* è la sigla del testo cui il file si riferisce. Va inserito nella *directory dei testi* del corpus.

testo base

(sinonimo di *file testo*).

tipo

Dato bibliografico associabile ad un testo. I codici utilizzati all'OVI sono:

M => misto originale/volgarizzamento.

O => originale.

P => parafrasi.

T => traduzione.

V => volgarizzamento.

APPENDICE G - CARATTERI ANSI

Generazione dei caratteri accentati

I caratteri accentati inclusi nella tavola ANSI sono ottenibili, a prescindere dalla loro presenza nella tastiera, usando il tastierino numerico a destra (controllare che sia acceso il led **BlocNum** o **NumLock**).

Per digitare una lettera accentata mediante il corrispondente codice ANSI procedere come segue.

Supponiamo di voler digitare la lettera È (codice **200**):

1. premere il tasto <ALT>,
2. tenendo premuto <ALT>, digitare <0> seguito dalle cifre costituenti il codice, nel caso in questione <2>, poi due volte <0>,
3. rilasciare <ALT>.

Tabella delle lettere accentate ANSI

à	0224	À	0192
á	0225	Á	0193
ã	0227	Ã	0195
â	0226	Â	0194
ä	0228	Ä	0196
å	0229	Å	0197
ç	0231	Ç	0199
è	0232	È	0200
é	0233	É	0201
ê	0234	Ê	0202
ë	0235	Ë	0203
ì	0236	Ì	0204
í	0237	Í	0205
î	0238	Î	0206
ï	0239	Ï	0207
ñ	0241	Ñ	0209
ò	0242	Ò	0210
ó	0243	Ó	0211
õ	0245	Õ	0213
ö	0246	Ö	0214
ù	0249	Ù	0217
ú	0250	Ú	0218
û	0251	Û	0219
ü	0252	Ü	0220
ý	0253	Ý	0221
ÿ	0255		