Martin-Dietrich Gleßgen (Zürich)

Esigenze della tecnologia informatica nella filologia e lessicografia storica

Tematica

Tratterò qui di seguito gli aspetti tecnologici e organizzativi della creazione di banche dati testuali come base per la lessicografia storica. È evidente che in questi campi un informatico ha delle competenze che mancano al letterato. Ma, allo stesso tempo, l'ottica con la quale lo storico della lingua affronta i problemi della tecnologia informatica corrisponde alle esigenze dettate dai contenuti. Peraltro, il problema della codificazione dei dati testuali è una tematica filologica e semiotica come lo sono l'elaborazione della scrittura o quella della stampa[1].

1. Potenzialità e pericoli della tecnologia informatica

1.1. Innovazioni informatiche recenti

L'informatica ha conosciuto nell'ultimo decennio del Novecento sviluppi importanti anche per quel che riguarda la storiografia linguistica. I parametri decisivi sono:

= la diffusione dei pc e del web e dunque la disponibilità dei mezzi informatici;

= la codificazione *xml* sviluppata sulla base di *sgml* (*Standard General Markup Language*) che struttura il testo puro con tag univoci, neutri e scelti liberamente (cfr. xml [2004])[2]. I tag possono servire tanto per rappresentare il testo a stampa o sotto forma elettronica quanto alla sua interrogazione informatica. Questo principio è un passo decisivo per la longevità dei dati testuali[3];

---

1 Ringrazio Sergio Lubello per la sua amichevole correzione del testo italiano.

2 La proposta della TEI (*Text Encoding Initiative*) per la scelta dei tag nelle banche dati testuali costituisce un primo consenso all'applicazione del *xml*; cfr. TEI [2001/2005], particolarmente importante è la versione TEI Lite. [La citazione dei documenti web segue le proposte della MLA (cfr. l'acuta sintesi di Gerstenberg 2001)].

3 L'applicazione di una struttura *xml* a un dato testo richiede un certo investimento di tempo: i tag e la loro interpretazione devono essere documentati tramite una *dtd* (*document type declaration*) o – oramai piuttosto – uno schema interpretabile da un parser (per es. XMLSpy, gratuito per la *home edition* 2005).

= la costituzione di *unicode* che dà spazio – con due milioni di posizioni – ai caratteri di tutte le lingue standardizzate del mondo e anche ad alcuni caratteri fuori uso[4]. Anche se *unicode* non permette di rendere conto integralmente della varianza grafematica antica, questo nuovo standard migliora l'internazionalità nella codificazione di dati;

= infine, lo sviluppo degli strumenti *open-source*, come *linux* che introducono nei mezzi informatici la trasparenza costituendone così la loro scientificità: uno strumento che trasforma dei dati testuali o linguistici, senza che il linguista possa seguire questo processo non risponde all'esigenza fondamentale della scienza, vale a dire la riproducibilità. Allo stesso tempo, l'elaborazione aperta e intersoggettiva di un programma può garantire una maggiore longevità, dato che l'evoluzione tecnologica è documentata e radicata nella comunità dei programmatori. Il programma non dipende più da un solo individuo o da una ditta commerciale, ma è patrimonio comune.

Disponiamo solo da un decennio di una base tecnologica che permetta anche lo sviluppo di corpus linguistici sotto forma informatica: i dati testuali sono oramai facilmente disponibili, possono avvalersi, con la dovuta attenzione, di una longevità reale e di una interpretabilità anche a fronte di esigenze future non prevedibili; gli strumenti informatici rendono possibile per la prima volta una trasparenza e, dunque, anche la riproduzione di dati e programmi con strumenti alternativi.

1.2. Utilità linguistica dei corpus testuali

I corpus testuali aprono nella linguistica nuove prospettive empiriche e teoriche grazie alle loro possibilità di quantificazione. Le trasformazioni della lingua possono essere seguite non solo nel tempo e nello spazio, ma anche nei diversi generi testuali e, quindi, nelle tradizioni discorsive e nelle varietà linguistiche[5]. L'interrogazione informatica si applica bene agli elementi formali come i grafemi, i lessemi e i morfemi, il che spiega la sua particolare utilità nella lessicografia storica: è facile sapere se la frequenza di una parola cambia in funzione dei parametri diasistematici; si può anche quantificare il contesto sintagmatico e identificare così eventuali fraseologismi[6].

Più problematici risultano la sintassi per la sua difficile segmentazione e, generalmente, gli aspetti funzionali e semantici (cfr. Stein 1995). Ma, per riprendere l'esempio del lessico, le evoluzioni formali danno anche indizi per il cambio semantico: è difficilmente immaginabile che esso non sia accompagnato da un cambio di frequenza o da un cambio sintagmatico, almeno all'interno di una varietà linguistica. Come in questo esempio le marche del linguaggio sono normalmente ridondanti. Se possiamo provare delle definite caratteristiche linguistiche di un testo o di una epoca in un campo della lingua, questo rende probabile l'esistenza di caratteristiche anche in altri campi. La linguistica di corpus apre la prospettiva di una possibile sintesi fra le interrogazioni della linguistica

---

4 Cfr. *unicode* [1991/2005].

5 Cfr. per l'argomentazione di questo paragrafo Gleßgen 2005.

6 Cfr. infra n. 12 per la linguistica di corpus.

testuale, variazionale e pragmatica e quelle della storia linguistica interna. Tale sintesi metterebbe in opera una concezione della linguistica basata sull'atto della parola, sull'*energeia*.

1.3. Utilizzazione attuale di banche dati nella lessicografia storico-etimologica

La linguistica di corpus ha un'utilità particolare nella lessicografia storico-etimologica data la facile identificabilità delle parole nel testo e la necessità, nei dizionari, di grandi quantità di dati. Oggigiorno è perfettamente realizzabile, anche se con un importante investimento di lavoro, un'integrazione complessiva di banche dati in un'opera lessicografica. Un'operazione simile ha diverse implicazioni:

= una gran parte delle fonti del dizionario dovrebbe essere disponibile sotto una forma informatica;

= la scelta e la strutturazione delle parole dovrebbero essere preparate in parte tramite programmi informatici;

= questi ultimi dovrebbero intervenire anche nel processo di redazione;

= il prodotto finale dovrebbe anch'esso essere interrogabile in quanto banca dati;

= nel caso ideale, gli articoli del dizionario dovrebbero essere legati a una banca dati testuale evolutiva.

Ma una tale realizzazione, come avviene per la romanistica nella coppia TLF-*Frantext* e giustamente nell'OVI[7], è piuttosto l'eccezione: il FEW, il LEI, il DI o anche il DEAF ed il DEM, cioè i dizionari che costituiscono i riferimenti metodologici della scienza attuale, non includono l'informatica nella loro redazione. Sfruttano le banche dati esistenti per migliorare la documentazione antica, ma la loro concezione resta indipendente dai mezzi informatici.

Un'impermeabilità, questa, che acquista importanza, dato che è indubbia l'utilità di disporre di dati linguistici codificati e di correlare dati testuali a dati interpretativi in una rete evolutiva. I problemi tecnologici che solleva l'utilizzazione dell'informatica nella lessicografia storica sono seri e di natura strutturale.

1.4. Imponderabilità e pericoli nella previsione

Abbiamo visto che la tecnologia informatica è operativa solo da poco tempo nella sua applicazione a grandi progetti filologici. Avventurarsi in quella via prima degli anni Novanta implicava rischi notevoli. La genesi del TLF e di *Frantext* ne sono una prova: l'evoluzione delle tecniche di codificazione dagli anni

---

7 Cfr. Beltrami in questo volume.

Sessanta in poi ha obbligato a copiare manualmente tre volte di seguito gli stessi testi, prima su carte perforate, poi su bande perforate, infine con dati magnetici: così sono stati sprecati forse 200 anni-lavoro[8].

Un altro progetto visionario era il *Dictionary of the Old Spanish Language* (DOSL), concepito per la prima volta negli anni 1930 e ripreso su una base informatica dal Hispanic Seminary of Medieval Studies a Madison dagli anni 1970 in poi. La trascrizione diplomatica dei manoscritti medievali – oggi disponibili per la maggior parte sotto forma di microschede – doveva creare la base filologica per un dizionario che non è mai stato iniziato. L'unica produzione dizionaristica, il *Diccionario de la antigua lengua médica* (DETEMA), fu preparata a Salamanca da un gruppo di lavoro indipendente, sotto la direzione di Maria Teresa Herrera e di Maria Nieves Sánchez.

Non è del tutto inusuale che talora i progetti lessicografici sortiscano l'effetto contrario a quello previsto: almeno nel passato, una base informatica sembra piuttosto aver aumentato i rischi senza allo stesso tempo facilitare la redazione dell'opera.

1.5. Tempo investito

Anche con una tecnologia operativa, la preparazione di una banca dati richiede un tempo considerevole, tanto per la concezione intellettuale quanto per l'elaborazione di dati e programmi. Il dizionario dei dialetti germanici dell'Austria (*Bayerisch-österreichisches Wörterbuch*) codifica attualmente i 4 milioni di schede dialettali in una banca dati programmata con il linguaggio-script *tustep*. Questa operazione occuperà quattro e più segretarie per 12 anni[9]. Quel tempo verrà probabilmente recuperato prima del 2010 grazie alla rapidità di redazione, nel frattempo aumentata (cfr. DBÖ). Ma l'investimento in termini di tempo è comunque considerevole.

Una soluzione più flessibile consisterebbe nel codificare i dati linguistici al momento della redazione degli articoli: l'autore integrerebbe le forme dialettali e antiche in una maschera piuttosto che elaborarli a mano e batterli poi al computer. Rimane tuttavia la necessità di una riflessione preliminare sulla struttura dell'insieme e di una programmazione che integri anche le informazioni già disponibili. Non è dunque affatto una soluzione di facilità, anche perchè nelle scienze umane le conoscenze informatiche non fanno parte del canone di formazione.

Lo sviluppo di strumenti informatici, la ricerca linguistica e l'elaborazione filologica sono tre entità che richiedono degli orizzonti di esperienza diversi. Di conseguenza, sono rari i progetti come gli studi dialettometrici di Hans Goebl

---

[8] Il nostro calcolo è una prudente proiezione sulla base dello studio in corso di Radermacher sulla storia del TLF (cfr. Radermacher 2004).

[9] Cfr. DBÖ: «acht DatentypistInnen».

che integrano in modo equilibrato dati consistenti con interrogazioni linguistiche rilevanti servendosi di strumenti informatici adeguati (cfr. DM).

2. Esigenze filologiche e tecnologiche per le banche dati testuali

2.1. Qualità dei dati testuali

Tralascio in questa sede la complessità delle interrogazioni linguistiche e dell'elaborazione lessicografica concreta per concentrarmi sulle esigenze filologiche e informatiche nella creazione di banche dati testuali. Le questioni inseparabili della qualità e della longevità dei dati formano la pietra miliare di tutta la linguistica di corpus: senza la garanzia di una scadenza a lungo termine un progetto di ricerca informatizzata nelle scienze umane non ha basi sicure.

Il primo problema nella linguistica di corpus storica è la base filologica. Nonostante la grande tradizione di critica editoriale della romanistica, le preoccupazioni filologiche sono facilmente messe fra parentesi davanti alle difficoltà informatiche. Per il francese medievale, per prendere l'esempio che attualmente mi interessa maggiormente, esistono oramai numerosi testi informatizzati e varie banche dati testuali: la base del DMF, concentrata sugli anni 1350-1500, che dovrebbe però essere allargata verso un corpus *Frantext* medievale (cfr. ATILF 2005); la *Base du Français Médiéval* elaborata sotto la direzione di Christiane Marchello-Nizia (cfr. BFM 2005); un insieme di testi riuniti nel *Laboratoire de Français ancien* di Ottawa (cfr. LFA 2004, ARTLF 2005) e che dovrebbero, come i testi della Marchello-Nizia, essere integrati dalla base *Frantext* antica; il *Corpus d'Amsterdam* di testi letterari del Duecento riuniti per la redazione dell'*Atlas des formes littéraires* di Antonij Dees e Piet Van Reenen[10].

Manca attualmente una descrizione accessibile e filologicamente convincente per questo insieme di testi (datazione e localizzazione dei testi e dei manoscritti, identificazione della tradizione testuale e del contesto sociologico di genesi, qualità dell'edizione)[11]. Nella maggior parte dei casi si tratta ancora di trascrizioni basate su edizioni di testo e non su singoli manoscritti, cosa che fa perdere certi vantaggi dello studio tramite mezzi informatici.

Ma persino una trascrizione basata su manoscritti non è a priori una garanzia di qualità. Un solo esempio: l'edizione informatica di Karl D. Uitti del *Chevalier de la Charrette* si basa sugli otto manoscritti del testo. I dati sono correttamente codificati con *sgml* e disponibili sul web. Le condizioni sarebbero dunque ideali per le interrogazioni linguistiche che Cinzia Pignatelli ha intrapreso a Poi-

---

[10] Cfr. Gleßgen 2003b, 1.6 e, più generalmente, Kunstmann 2000, 2003, Stein 2004, 2.2 e Stein / Gleßgen 2005; cfr. anche infra 1.4. per la creazione di una banca dati per i testi documentali.

[11] Cfr. la descrizione dei testi informatizzati in francese medievale e rinascimentale sulla pagina *Ménéstrel* (2004): Textes / France (René Pellen, [01.02.2005]).

tiers. Ma nella verifica la trascrizione si è avverata incoerente, erronea e inutilizzabile senza una correzione integrale (cfr. Pignatelli 2003). Uno scienziato come Uitti non avrebbe mai osato pubblicare delle trascrizioni del genere in un libro a stampa, altrimenti si sarebbe esposto a una critica feroce. Questo caso preso dalla storiografia del francese e forse anche estremo, illustra i pericoli che l'assenza di una critica sistematica della base filologica procura alle banche dati. È vero che il TLIO è esemplare da questo punto di vista, ma non si tratta affatto di uno standard generalizzato.

2.2. Programmi di elaborazione

La codificazione e la qualità dei programmi hanno la stessa importanza dei dati testuali o di quelli linguistici. Ma qui si toccano questioni epistemologiche non ancora risolte. Lo studioso o il gruppo di lavoro che vuol creare una banca dati ha attualmente quattro possibilità:

= una programmazione completa con un linguaggio basico come c^{++} o *java*, pesante e quasi esclusa per un linguista data la pesantezza della procedura e le competenze necessarie;

= una programmazione, più leggera, con strumenti preparati per quei fini come i linguaggi-script (*perl*, *python* o *tustep*) e gli strumenti di interpretazione sviluppati per l'*xml* (come X-Query e XSLT), un programma di stampa (come *tex*/*latex*), un sistema di banca dati *open source* (come MySYL);

= l'adattamento di programmi commerciali come *file maker* o *Access*, di facile accesso e diffusi; qui però l'utente è limitato alle configurazioni previste, non controlla bene la relazione fra i dati testuali e il programma per il quale non ha nessuna garanzia di longevità;

= l'utilizzazione di programmi non commerciali elaborati per grandi progetti di banche dati e corpus linguistici come *STELLA* di *Frantext*, *GATTO* o *PhiloLogic* dell'OVI e dell'ARTFL; la trasparenza di questi programmi è spesso ridotta: per *STELLA* tanto il codice quanto il programma sono inaccessibili alla comunità scientifica.

Gli strumenti disponibili richiedono uno sforzo importante o contengono rischi per la longevità. Anche le possibilità di interrogazioni linguistiche sono spesso ridotte: l'utente è dipendente da istanze non trasparenti. I meccanismi di controllo e di intersoggettività, normalmente costitutivi per la scienza, sono ristretti o assenti. Lo sviluppo di una critica sistematica degli strumenti informatici utili per la filologia e la linguistica è dunque uno dei desiderata più urgenti in materia[12]. Aspettando le inevitabili evoluzioni, è necessario garantirsi al massimo contro la perdita dei dati tanto testuali quanto di programmazione; l'unica via è

---

[12] Le numerose introduzioni alla linguistica di corpus danno naturalmente indicazioni molteplici sugli strumenti, ma senza gerarchizzare le loro qualità e senza indicare precisamente la loro utilità per i diversi fini (per es. Biber / Conrad / Reppen 1998: 281-287; Kennedy 1998: 259-267; Oakes 1998: 156-158; 182s.; 193-195; il problema è posto più chiaramente da Habert / Fabre / Issac 1998: 294-298; 305-308); sono rari i commenti più dettagliati e le interrogazioni più specifiche sulla finalità e l'applicabilità degli strumenti (cfr. per es. Stein 1995).

quella di scegliere dei programmi trasparenti, documentati e possibilmente *open source* (cfr. infra 2.6). Aggiungiamo che nessuno strumento garantisce né garantirà mai tutte le funzionalità al linguista storico, il che aumenta la necessità di dati e programmazioni aperti all'esportazione e all'importazione.

2.3. Esigenze fondamentali per una linguistica di corpus

Una linguistica di corpus deve dunque tener conto di diverse esigenze fondamentali:

= nella filologia: i dati testuali devono essere preparati con lo stesso rigore delle pubblicazioni a stampa, e devono essere sottoposti alla stessa critica;

= nella tecnologia di codificazione: la codificazione dei dati testuali deve seguire i principi di neutralità del *xml* e prendere in considerazione le premesse di *unicode* e della TEI[13]. Peraltro è necessario prevedere l'aggiornamento dei dati al di là dell'individuo, eventualmente in collaborazione con le grandi biblioteche nazionali o universitarie;

= nella tecnologia di programmazione: è necessario lo sviluppo di strumenti non commerciali, facilmente accessibili, trasparenti, documentati e sottoposti alla critica. I programmi dovrebbero permettere l'importazione o l'esportazione di dati testuali e linguistici. Come per i dati testuali, è necessario prevedere un aggiornamento sistematico dei programmi di elaborazione.

Aggiungiamo che i programmi utilizzati non devono impedire una qualsiasi interrogazione linguistica per la quale sia possibile definire un algoritmo. È certo che i mezzi hanno un influsso sulla maniera di muoversi; ma non devono determinare la via da prendere.

2.4. Un esempio di applicazione: il programma per l'edizione e l'analisi linguistica di testi romanzi antichi (PHOENIX)

Concludo con una parola su un mio progetto: lo sviluppo di uno strumento che permetta l'edizione di testi romanzi antichi, sotto forma elettronica o a stampa, e la loro interrogazione a fini grafematici, morfo-sintattici e lessicologici. La base empirica è costituita dai documenti non-letterari in antico francese, nel quadro dei *Plus anciens documents linguistiques de la France*, diretti da Françoise Vielliard, Olivier Guyotjeannin e l'autore delle presenti righe. Si tratta per me di preparare l'edizione informatica e l'elaborazione dei glossari che nutriranno una banca dati lessicale, complementare al dizionario di Godefroy[14].

Per il programma, elaborato con Matthias Kopp e Matthias Osthof (Tübingen), abbiamo scelto il linguaggio-script *tustep*, strumento efficace e sicuro per

---

[13] Già il programma di scrittura *Open office* (gratuito e basato su *Linux*, compatibile con i programmi commerciali attuali) garantisce una codificazione basica di tipo *xml*.

[14] Cfr. Gleßgen 2003a/b con ulteriori rinvii.

scopi di tipo filologico e linguistico. Un'altra opzione sarebbe stata una combinazione fra *perl* (per la parte di interrogazione) e *tex/latex* (per la parte editoriale). Ma *tustep*, anche se meno conosciuto e dunque meno accessibile, ha il vantaggio di riunire le qualità di entrambi e di essere tecnicamente superiore per le problematiche filologiche. Permette anche, tramite l'interfaccia *cgi*, una riproduzione facile dei dati sotto forma di *browser*. La messa a disposizione pubblica delle fonti del linguaggio-script è prevista per prima del 2007.

L'architettura del programma *PHOENIX* è ben strutturata, trasparente e commentata per rispondere all'eventualità di una riprogrammazione con altri strumenti nel futuro (cfr. Gleßgen / Kopp 2005). Se il programma ha un futuro, una tale riprogrammazione sarà prima o poi inevitabile perché gli strumenti informatici non sono eterni. Nel frattempo la programmazione dovrebbe servire come *tertium comparationis* per valutare le qualità di altri strumenti informatici utilizzati a fini filologico-linguistici. La programmazione si trova attualmente, dopo quattro anni, a metà strada e, naturalmente, prevede alla fine una messa in rete delle fonti commentate.

Le strutture di codificazione sono aperte e permettono l'integrazione di altri tipi di testi in antico o medio francese, ma anche in altre lingue romanze come l'italiano, l'occitanico o il latino medievale. Finora sono state integrate solo trascrizioni di manoscritti, filologicamente sicure e codificate secondo un modello *xml* che prevede variazioni secondo i generi testuali. Un altro principio di base è la collocazione dei singoli manoscritti e dei testi nel diasistema storico-linguistico tramite una griglia analitica precisa che permetta delle interrogazioni sotto forma di banca dati. Una collaborazione con Pierre Kunstmann e Achim Stein prepara la interazione con dati linguistici e programmi esterni (cfr. Stein / Gleßgen 2005); per le interrogazioni valutative di linguistica storica possono intervenire procedure programmate con X-Query.

Il programma *PHOENIX* è nato con obiettivi diversi da quelli di *GATTO* o di *STELLA*. Anche se può servire per una banca dati importante, i suoi obiettivi di partenza sono l'edizione e l'elaborazione filologica di un testo individuale per un utente individuale. Ma certe funzionalità sono parallele e bisognerà dunque interrogarsi sulle interfacce potenziali. Indipendentemente da tali prospettive speriamo che lo sforzo solido di riflessione informatica e di programmazione si giustifichi per lo studio linguistico dei testi documentari in antico francese.

Bibliografia

ATILF = *ATILF. Analyse et traitement informatique de la langue française* [2005, s.d.].
<http://www.atilf.fr> (15.01.2005)

ARTLF = *American and French Research on the Treasury of the French Language* [2005, s.d.].
<http://humanities.uchicago.edu/orgs/ARTFL> (15.01.2005)

BFM = *Base du Français Médiéval*, in: *Corpus, Ressources et Apprentissages Linguistiques: Diachronie du français* [luglio 2003].
<http://www.ens-lsh.fr/labo/corpus> (15.01.2005)

Biber, D. / Conrad, S. / Reppen, R., *Corpus Linguistics. Investigating Language Structure and Use*, Cambridge University Press, 1998 (repr. 2000).

DBÖ = *Datenbank der bairischen Mundarten in Österreich*, in: *Institut für Österreichische Dialekt- und Namenwörterbücher* [22.09.2004].
<http://oeaw.ac.at/dinamlex> (15.01.2005)

DETEMA = Herrera, M. T. / Sánchez, M. N., *Diccionario español de textos médicos antiguos*, Madrid, Arco/Libros, 1996 [e Herrera, M. T. / González de Fauve, M. E., *Textos y concordancias electrónicos del Corpus Médico Español*, ed. CD-ROM, Madison, 1997].

DM = *Das Dialektometrieprojekt der Universität Salzburg* [febbraio 2004].
<http://ald.sbg.ac.at/dm> (15.01.2005)

Gerstenberg, A., *Zitieren von Internetdokumenten*, in: Italienisch 23 (2002), 172-175.

Gleßgen, M.-D., *L'élaboration philologique et l'étude lexicologique des 'Plus anciens documents linguistiques de la France' à l'aide de l'informatique*, in: F. Duval (ed.), *Frédéric Godefroy. Actes du X^e colloque international sur le moyen français (Metz, 12-14 juin 2002)*, Paris, 2003, 371-386 (2003a).

–, *La lemmatisation de textes d'ancien français: méthodes et recherches*, in: Kunstmann 2003, 55-75 (2003b).

–, *Diskurstraditionen zwischen pragmatischen Regeln und sprachlichen Varietäten*, in: A. Schrott / H. Völker (edd.), *Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen*, Göttingen, Universitätsverlag, 2005, 207-228.

Gleßgen, M.-D. / Kopp, M., *Linguistic annotation of texts in non-standardized languages : the program procedures of the tool PHOENIX*, in: Kabatek / Pusch / Raible, 2005, 147-154.

Habert, B. / Fabre, C. / Issac, F., *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*, Paris, Masson, 1998.

Kabatek, J. / Pusch, C. / Raible, W. (edd.), *Romance Corpus Linguistics II : Corpora and Diachronic Linguistics* (ScriptOralia, 130), Tübingen, Narr, 2005.

Kennedy, G., *An introduction to Corpus Linguistics*, London / New York, Longman, 1998.

Kunstmann, P., *Ancien et moyen français sur le web: textes et bases de données*, RLiR 64 (2000), 17-42.

Kunstmann, P. (ed.), *Ancien et moyen français sur le Web: enjeux méthodologiques, Colloque d'Ottawa (4-6 octobre 2002)*, Ottawa, 2003.

LFA = *Laboratoire de Français ancien* [18.11.2004].
<http://www.uottawa.ca/academic/arts/lfa> (15.01.2005)

Ménéstrel = *Ménéstrel. Un portail pour les médiévistes* [30.09.2004]. <http://ccr.jussieu.fr/urfist/omedirht.htm> (15.02.2005)

Oakes, M.P., *Statistics for Corpus Linguistics*, Edinburgh University Press, 1998.

Pignatelli, C., *L'archive du Projet 'Charrette': huit témoins prêts à se livrer*, in: Kunstmann 2003, 203-220.

Radermacher, R., *'Le Trésor de la langue française'. Une étude historique et lexicographique*, Thèse de doctorat dactylographiée, Université Marc Bloch de Strasbourg, 2004 [in corso di stampa].

Stein, A., *Maschinenlesbare Textkorpora für das Französische*, in: Zeitschrift für französische Sprache und Literatur, 105 (1995), 1-25.

–, *Wörterbücher und Textkorpora für Französisch und Italienisch*, RK XVI (2004), 107-124.

Stein, A. / Gleßgen, M.-D., *Resources and Tools for Analyzing Old French Texts*, in: Kabatek / Pusch / Raible, 2005, 135-145.

TEI = *Text Encoding Initiative. TEI Website* [15.08.2001 / 06.09.2003]. <http://www.tei-c.org> (15.01.2005)

unicode = *Unicode Home Page* [1991-2005]. <http://www.unicode.org> (15.01.2005)

xml = *W3C – Extensible Markup Language* (XML) [04.02.2004]. <http://www.w3.org/TR/REC-xml> (15.01.2005)