Burr, Elisabeth (2003): "Elektronische italienische Zeitungskorpora", in: Rainer, Franz / Stein, Achim (eds.): *I nuovi media come strumenti per la ricerca linguistica* (= Sprache im Kontext 18). Franfurt am Main etc.: Peter Lang 11-26.

ELISABETH BURR

ELEKTRONISCHE ITALIENISCHE ZEITUNGSKORPORA

1. Korpora und Korpuslinguistik

Unter Korpora verstehen wir heute zunächst einmal grundsätzlich elektronische Korpora: „la nozione di corpus è da subito associata a quella di raccolta di porzioni di lingua in formato elettronico (*machine readable form*)." (Spina 2001: 53). Korpus und Digitalisierung sind also synonym: „Le terme de corpus est devenue indissociable de celui d'informatisation." (Blanche-Benveniste 2000: 14). Solche Korpora können in vielen Zweigen der Linguistik als Hilfsmittel eingesetzt werden, sie sind aber auch die Grundlage einer neuen Disziplin, nämlich der Korpuslinguistik. Bei der Korpuslinguistik handelt es sich zumindest potentiell um eine Linguistik der *parole*, der Performanz oder des Sprechens. Ihre Fragestellungen und Methoden werden maßgeblich von den Möglichkeiten bestimmt, die die computationelle Verarbeitung von Sprachdaten bietet. Korpora und Korpuslinguistik sind eben auch ein Kind der technologischen Entwicklung im 20. Jahrhundert.[1]

1.1. Korpuslinguistik

Der Beginn der Korpuslinguistik wird traditionell auf die 60er Jahre festgelegt, weil damals die ersten elektronischen Korpora entstanden. Das aller erste elektronische Korpus, das selbst wieder zum Modell für viele der in der Folge erstellten Korpora wurde, ist das an der Brown Universität unter der Leitung von W. Nelson Francis und Henry Kučera ausdrücklich *for use with Digital Computers* erstellte *Brown Corpus*, das einen Umfang von 1 Millionen Wortformen hat.[2]

Aus europäischer und besonders italienischer Sicht ließe sich jedoch der Anfang der Korpuslinguistik auch gut und gern auf das Ende der 40er Jahre festlegen. Damals fing nämlich Pater Busa an, das Gesamtwerk von Thomas von Aquin mit seinen 10,6 Millionen Wörtern auf Lochkarten zu übertragen. Sein Ziel war, mit einer Sortiermaschine den *Index Thomisticus* zu erstellen. Beendet

[1] Das soll natürlich nicht heißen, dass es vor dem Erscheinen des Computers keine Korpora und Korpus-basierte Sprachuntersuchungen gab. Interessant sind in diesem Zusammenhang u.a. Svartvik (1992) und Francis (1992).

[2] Für eine genaue Beschreibung cf. Francis / Kučera (1964 / 1979).

wurde die Arbeit dann 33 Jahre später auf einem großen IBM Mainframe-Computer.[3]

Was dagegen die Etablierung des Terminus *Korpuslinguistik* angeht, so scheint es diesbezüglich keine Zweifel an der Terminierung zu geben. Denn 1984 veröffentlichten Jan Aarts und Willem Meijs einen Sammelband mit dem Titel *Corpus linguistics* und führten damit den Terminus ein.

Seit den 90er Jahren ist die Korpuslinguistik nun dabei, sich von einer Methodologie zu einem eigenständigen theoretischen Ansatz und damit zu einer neuen Disziplin zu entwickeln (cf. Leech 1992). Diese wird zwar nach wie vor v.a. im angelsächsischen Raum und mit Blick auf das Englische betrieben, in der letzten Zeit wächst aber in den romanisch-sprachigen Ländern und in der Romanistik im deutschsprachigen Raum das Interesse an der Korpuslinguistik beträchtlich. Als Beweis für diese Entwicklung können zum einen die Publikationen von Habert / Nazarenko / Salem (1997), Bilger (2000) und Spina (2001) sowie die *Journées de la Linguistique de Corpus* an der *Université de Bretagne-Sud* (cf. http://www.univ-ubs.fr/crellic/) oder *Corpus Use and Learning to Translate* (CULT) an der *Università di Bologna* (cf. http://www.sslmit.unibo.it/cult2k/) gelten, zum anderen die *1. Freiburger Arbeitstagung zur Romanistischen Korpus-Linguistik* (cf. Pusch / Raible im Druck).[4]

Dass sich auch die internationale Italianistik verstärkt der Erstellung von Korpora und der Korpuslinguistik zuwendet, hat der im Jahre 2000 von der Duisburger Italianistik ausgerichtete VI Internationale Kongress der *Società Internazionale di Linguistica e Filologia Italiana* (SILFI) gezeigt, wo nicht nur alle wichtigen elektronischen Korpora zum *italiano parlato* (*spontaneo* oder *trasmesso*) und die auf ihrer Basis laufende Forschung diskutiert wurden (cf. Burr im Druck), sondern auch Korpora und Korpus-gestützte Untersuchungen zu (historischen) Varietäten des *italiano scritto* (cf. Burr in Vorbereitung). Als Gradmesser für die Aufmerksamkeit, die die deutschsprachige Italianistik der Korpuslinguistik widmet, lässt sich der Italianistentag 2001 in Dresden werten.

Die Korpuslinguistik als Disziplin ist natürlich der seit Chomsky hauptsächlich betriebenen (theoretischen) Linguistik diametral entgegengesetzt. Ihr geht es nicht primär um die Kompetenz bzw. das Universelle, sondern um die Erforschung natürlicher Sprachen bzw. Varietäten über das Auffinden und die Analyse von Regelmäßigkeiten: „l'obiettivo è ricercare e identificare fenomeni che si ripetono con regolarità" (Spina 2001: 53). Das heißt allerdings nicht, dass sie nicht aus den dabei gewonnenen Erkenntnissen allgemeine Prinzipien ableitet. Das *idiom-principle* von John Sinclair (1991) ist z.B. ein solches Prinzip.

---

[3] Näheres hierzu in Busa (1980).

[4] Für das wachsende Interesse der deutschsprachigen Linguistik an der Korpuslinguistik cf. die Arbeit von Lenz (2000).

1.2. Korpora

Die Grundlage der Korpuslinguistik sind, wie gesagt, elektronische Korpora. Definiert werden sie heute wie folgt: „Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage" (Habert / Nazarenko / Salem 1997: 11). Korpora stellen also Sammlungen von natürlichem Sprachmaterial dar, d.h. von Texten, die tatsächlich in einem wirklichen sozialen Kontext und ohne Eingriff von Linguistinnen oder Linguisten schriftlich oder mündlich geäußert wurden. Sie konservieren damit das Sprechen im Sinne von realisierter Sprache. Als weitere Bedingung kommt hinzu, dass ihrer Erstellung linguistische Kriterien zugrunde liegen und diese explizit gemacht werden.

Exhaustive Korpora

Der heutige Begriff *Korpus* ist allerdings selbst das Ergebnis einer Metamorphose. Korpus bedeutet nämlich, wie Blanche-Benveniste ausführt, zunächst Jahrhunderte lang eine Sammlung aller Dokumente eines bestimmten Studienbereichs, d.h. „L'exhaustivité des données rassemblées faisait partie de la définition." (Blanche-Benveniste 2000: 11). Als Beispiele führt sie an:

- *Corpus Juris*,
- Korpus der griechischen und lateinischen Inschriften,
- Korpus der Kataloge mittelalterlicher Bibliotheken.

Stehen im Zentrum des Interesses nun nicht die Texte als solche, sondern die darin enthaltenen sprachlichen Daten, dann lassen sich unter einen so definierten Korpusbegriff natürlich streng genommen nur Sammlungen von Dokumenten toter Sprachen (*Thesaurus Linguae Graecae*); früherer Sprachstufen (*Tesoro della lingua italiana delle origini*), verstorbener Schriftstellerinnen und Schriftsteller (*Index Thomisticus*) oder ähnlicher Entitäten fassen, deren Grenzen sich höchsten noch durch die Entdeckung bisher unbekannter Schriftstücke verschieben lassen. Exhaustive Sammlungen von Daten lebender Sprachen sind dagegen nicht möglich.

Repräsentative Korpora der Sprache

Deshalb geht es 1961 bei der ersten Anwendung des Terminus *Korpus* auf eine Sammlung von Daten lebender Sprachen auch nicht um Vollständigkeit im eigentlichen Sinne, sondern um etwas, was dieser ähnlich ist, d.h. um „une bonne représentativité de la langue" (Blanche-Benveniste 2000: 11). Um nämlich zu gewährleisten, dass das *Brown Corpus* das *Present-Day Edited American English*

repräsentiert, werden zufällig 500 Textfragmente von je 2000 Wortformen aus 15 verschiedenen Genres erhoben.

Die Frage, was unter Vollständigkeit im Sinne statistischer Repräsentativität zu verstehen ist und aufgrund welcher Kriterien sich eine Sprache einschließlich ihrer Varietäten als solche abbilden lässt, so dass die anhand eines begrenzten Materials erzielten Untersuchungsergebnisse für die Sprache als solche gelten können, war damit aber nicht gelöst. Stattdessen hat sie noch fast 30 Jahre lang die Gemüter bewegt. Ausdruck dieses Ringens sind eigentlich alle frühen Korpora, bei deren Erstellung eine Auswahl aus einer unendlichen oder doch zumindest sehr großen Menge von sprachlichen Realisierungen zu treffen war.

Obwohl diese frühen Korpora für die Herausbildung der Korpuslinguistik von großer Bedeutung waren und noch heute wertvolle Grundlagen für die Forschung darstellen, so darf doch auch die hemmende Wirkung, die die Diskussion um statistische Repräsentativität auf die Entwicklung von Korpusprojekten hatte, nicht übersehen werden. Genaue Kriterien waren schließlich nicht zu bestimmen und kaum jemand sah sich in der Lage, die jeweils diskutierten Kriterien auch nur annähernd zu erfüllen.

Repräsentative Korpora des Sprechens

Selbstverständlich sind auch die neueren Referenz- oder *general* Korpora dem Kriterium der Repräsentativität verpflichtet. In der Zwischenzeit wurde jedoch die ursprünglich aus den Sozialwissenschaften mit ihrer genau definier- und abgrenzbaren Population übernommene proportionale Erhebung von Stichproben durch eine Erhebung möglichst ganzer Texte aufgrund von Strata und *sampling frames* ersetzt, die die zumeist unendliche Natur sprachlicher Populationen, das komplexe Zusammenspiel von sprachlicher Produktion, Rezeption und sprachlicher Produkte sowie die Kontextbedingtheit sprachlicher Phänomene berücksichtigen[5] und so der kulturellen Bedeutung der Arten des mündlichen und schriftlichen Sprechens sowie ihrer Verbreitung in der Sprachgemeinschaft und bei den Sprechenden eher Rechnung tragen. Damit werden repräsentative Korpora des Sprechens und des sprachlichen Wissens denkbar, wie ich sie in Burr (1997) gefordert habe. Da sich zudem der Fokus von der Größe auf das Forschungsziel verschiebt[6] und Machbarkeit zu einem zulässigen Kriterium wird, gelten jetzt auch kleinere Korpora als wertvoller Beitrag zur Schaffung der Voraussetzungen für eine korpuslinguistische Forschung. Denn je mehr Korpora untersucht werden können, desto besser lassen sich die Erstellungskriterien über-

[5] Für die Diskussion um das *corpus design* cf. u.a. Sinclair (1991), Clear (1992), Biber (1994) und Engwall (1994). Ein Korpus, bei dessen Erstellung diese Diskussion beachtet wurde, ist das *British National Corpus* (cf. Burnard im Druck).

[6] So ist Engwall (1994: 51) der Meinung, dass "no scientific criteria exist for determining the size of any corpus" und Clear (1992: 27) spricht von einem „corpus for a purpose".

prüfen und weiterentwickeln (cf. Clear 1992: 30). Zudem kann nur durch die Untersuchung des Sprechens dem allseits beklagten Mangel an Kenntnissen über das Sprechen begegnet werden.

Mit Markup angereicherte Korpora

Aber selbst wenn ein Korpus alle angesprochenen Bedingungen erfüllt, heißt das noch nicht, dass es also solches schon für linguistische Untersuchungen geeignet ist. Reine Textkorpora oder *suites de mots «nus»*, wie sie Habert, Nazarenko und Salem (1997: 7) nennen, können zwar die Grundlage interessanter Untersuchungen v.a. zur Lexik sein, vielen anderen Fragestellungen werden sie aber nicht gerecht. So sind zum einen sprachliche Phänomene nicht gleichmäßig über Texte verteilt, zum anderen können mit Hilfe des Computers nur solche Daten erhoben werden, die ausdrücklich definiert worden sind. Deshalb werden Korpora annotiert, d.h. die Texte werden mit Informationen angereichert. Solche Annotationen können unterschiedliche Grade der Verfeinerung erreichen. So gibt es Korpora, die nur mit lexikographischen oder textstrukturellen Informationen angereichert wurden, andere sind dagegen auf allen Ebenen annotiert. Wieder andere enthalten darüber hinaus syntaktische Analysen.

Aufweichung des Korpusbegriffs

Während aber einerseits die direkt an Fragen des *corpus design* Interessierten die für die Korpuserstellung relevanten Text-externen Kategorien diskutieren und die Repräsentativität von Korpora bezüglich der Verteilung von sprachlichen Phänomenen in Texten und über Texte und Textkategorien hinweg analysieren (cf. z.B. Biber / Conrad / Reppen 1998), andererseits internationale Textkodierungsstandards entwickelt werden,[7] wird der Begriff *Korpus* in der einzelsprachlichen Linguistik immer mehr zu einem Synonym für eine wie auch immer geartete Sammlung elektronischer Texte: „Aujourd'hui que le texte électronique foisonne, des documents se trouvent parfois agrégés avant tout parce qu'ils sont faciles d'accès, sans que leur mise en relation ait été réellement pensée. La définition raisonnée d'un regroupement adéquat à l'objectif poursuivi cède le pas à la seule disponibilité des ressources. " (Habert / Nazarenko / Salem 1997: 143).

Nun lassen sich zwar aus großen Sammlungen strukturierte Sub-Korpora extrahieren, in der Korpuslinguistik herrscht aber trotzdem Konsens, dass Korpora von digitalen Textarchiven oder Sammlungen zu unterscheiden sind. Ein Archiv wie die LIZ, die CD-ROM von *Le Monde* oder die im Folgenden vorgestellten italienischen Zeitungsarchive wurden schließlich nicht anhand linguistischer Kriterien erstellt. Auch sind sie nicht primär auf eine Analyse der Daten ausge-

[7] Cf. die *Guidelines* der *Text Encoding Initiatve* (TEI) (cf. <http://www.tei-c.org/>).

richtet, sondern es geht letztendlich um die Konservierung und Archivierung von Textmaterial und das Wiederauffinden der verschiedenen Komponenten.

2. Digitale Zeitungsarchive

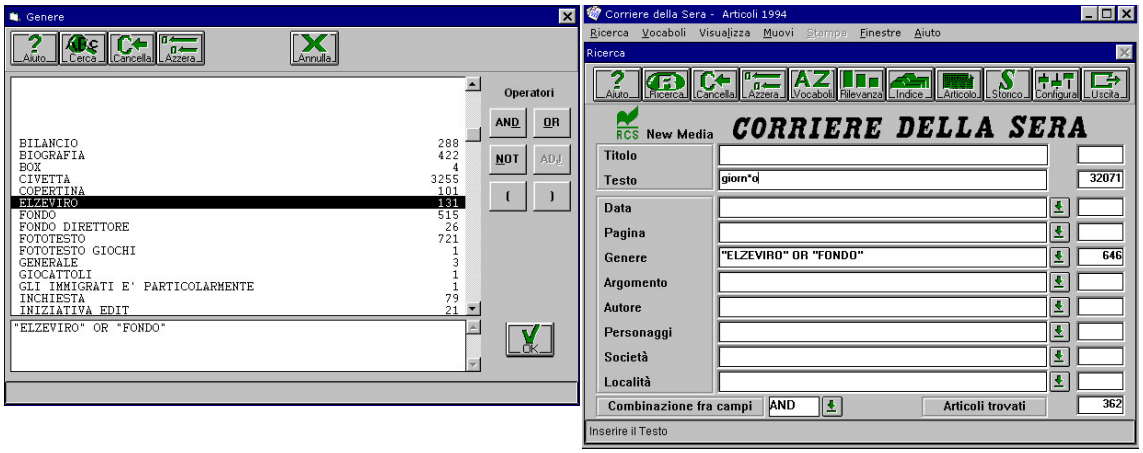
2.1. Corriere della Sera

Die erste CD-ROM des *Corriere della Sera* erschien 1992. Hierbei handelte es sich aber zunächst nur um einen Index der zwischen 1984 und 1991 in der Zeitung publizierten Artikel.[8] Erst mit der Jahresausgabe des *Corriere* von 1992 auf CD-ROM standen dann auch die Artikel selbst zur Verfügung.

Dem Klappentext der Ausgabe 1994,[9] die mir hier als Beispiel dient, lässt sich nicht nur entnehmen, als was diese CD-ROMs verstanden wurden:

> Questo archivio elettronico su CD-ROM (Compact Disc – Read Only Memory) contiene 76.921 articoli pubblicati sul quotidiano e sui supplementi (ad eccezione del supplemento illustrato “Sette”).[10]

sondern auch an welches Publikum sich die einzelnen Archive richten, nämlich an „broker dell’informazione, biblioteche, centri studio e ricerca, centri documentazione, società di consulenza, studi legali, scuole superiori, etc.“



Die Oberfläche des beigegebenen Textretrieval-Programms macht sofort klar, weshalb digitale Texte mit *Markup* anzureichern sind. Die *Campi* können nämlich nur dann als Suchkriterien fungieren, wenn die entsprechenden Textstellen selbst als *Titolo*, *Generi* etc. indiziert wurden.

Aufgrund dieses Markup kann nach einzelnen Wortformen entweder nur in den Titeln (*Titolo*) oder sowohl in den Titel als auch in den Artikeln (*Testo*) eines

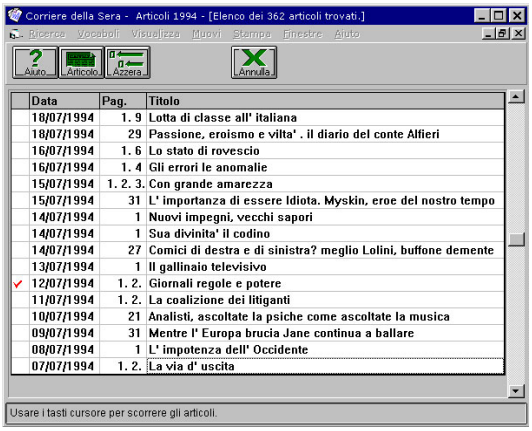
---

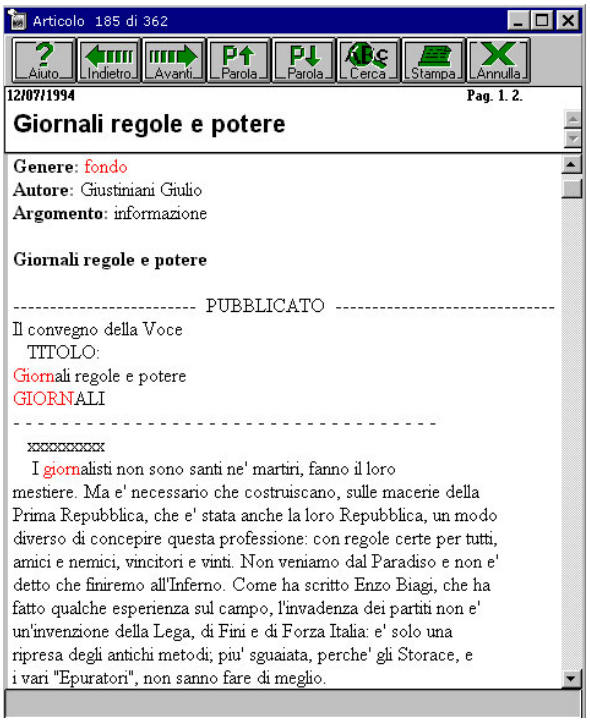
8 Lit. 750.000 ohne MwSt; die Preise werden nur aufgeführt, weil Korpuserstellung und -erforschung immer auch von den finanziellen Gegebenheiten abhängen.

9 Lit. 600.000 incl. MwSt.

10 Auch die Werbung wurde nicht archiviert; die *Pubblicità Redazionale* ist vorhanden.

bestimmten *Genere* (*elzeviro*, *fondo* etc.) zu einem bestimmten *Argomento* in einer oder mehreren Ausgaben des *Corriere* 1994 – die Auswahl wird jeweils in einem *drop-down* Fenster vorgenommen - gesucht werden. Welche Wortformen (*tokens*) im Archiv mit welcher Frequenz enthalten sind, erfahre ich übrigens unter dem Menüpunkt *A-Z*. Das Setzen einer *Wildcard* ist zugelassen (*giorn*o*), berücksichtigt wird bei der Suche allerdings nur die Trunkierung (*giorn**).





Als Ergebnis wird eine Liste der Artikel ausgeworfen, die den Suchkriterien entsprechen. Jeder dieser Artikel kann als Ganzes eingesehen, ausgedruckt oder als Datei abgespeichert werden. Nicht möglich ist allerdings, mehrere Texte gleichzeitig zu exportieren. Soll aus dem Archiv also ein Korpus entstehen, muss jeder der 76.401 Artikel einzeln auf einem Datenträger archiviert werden.

An diesen Funktionen hat sich auch bei der letzten CD-ROM des *Corriere della Sera* von 1998, mit der das Erscheinen von digitalen Jahresausgaben eingestellt wurde, nichts geändert. Die Oberfläche und auch die Restriktionen sind die gleichen geblieben. Auch der Klappentext sagt noch immer das, was ich weiter oben schon zitiert habe, nur sind zu dem ausgeschlossenen *supplemento Sette* noch *TV Sette* und *Donna* hinzugekommen, d.h. *supplementi*, die 1994 noch nicht existierten. Die CD-ROMs des *Corriere della Sera* sind also bis zuletzt das geblieben, was sie von Anfang an sein wollten: ein *Archivio Elettronico Articoli*.

2.2. Il Sole 24 ore - Banche Dati online

Das Finanzblatt *Il Sole 24 ore* stellt in seinem *Archivio* eine ganze Reihe von u.a. auf Wirtschaft und Finanzen spezialisierten *online*-Zeitungen zur Verfügung

(cf. <http://www.banchedati.ilsole24ore.com/>), darunter z.B. auch *Tutto La Stampa*.[11] Die Anzahl der Datenbanken wird fortwährend vergrößert. Daten können in einer oder in allen Datenbanken zugleich erhoben werden. Die Suche nach einer bestimmten Wortform z.B. erbringt wie beim *Corriere* eine Liste aller Dokumente, in denen sich der Suchbegriff findet. Von hier aus kann dann entweder eine Kurzfassung oder der ganze Text eingesehen werden. Beide lassen sich auch abspeichern.

Auch die *online*-Datenbanken von *Il Sole 24 ore* sind aber, trotz ihres beachtlichen Umfangs und den komfortableren Such- und Exportmöglichkeiten, vom Prinzip her nichts anderes als die CD-ROMs des *Corriere*, d.h. elektronische Archive von Dokumenten, aus denen nach bestimmten Kriterien Texte abgerufen werden können. Korpora für sprachwissenschaftliche Untersuchungen sind sie nicht und wollen sie auch gar nicht sein.

2.3. Internazionale

Zuletzt soll noch die *CD-ROM Internazionale 1993-1998* zur Sprache kommen, die als Supplement zur Nr. 319 der Zeitschrift vom 28. Januar 2000 erschienen ist.[12] *Internazionale* ist allerdings keine Tageszeitung, sondern eine Art von Wochenzeitung, die Artikel aus anderen Zeitungen sammelt:

> Ogni settimana Internazionale pubblica in italiano i migliori articoli comparsi sulla stampa straniera. Una finestra sul mondo della cultura, della politica, dell'economia, della tecnologia e della scienza.
> Tutti gli articoli sono tradotti integralmente e vengono scelti da un comitato di esperti e consulenti a partire da circa trecento giornali quotidiani, settimanali e mensili.

Auf der CD-ROM sind “tutti i testi pubblicati sul settimanale internazionale fino al 1998” archiviert. Im Unterschied zu den beiden anderen Archiven können hier entweder alle Texte eines ganzen Jahres auf einmal, nur ein Artikel oder eine Sammlung von Texten zu einem bestimmten Thema / zu einer bestimmten Suchanfrage exportiert werden. Zudem bietet die mitgelieferte Software sehr ausgefeilte Suchfunktionen.

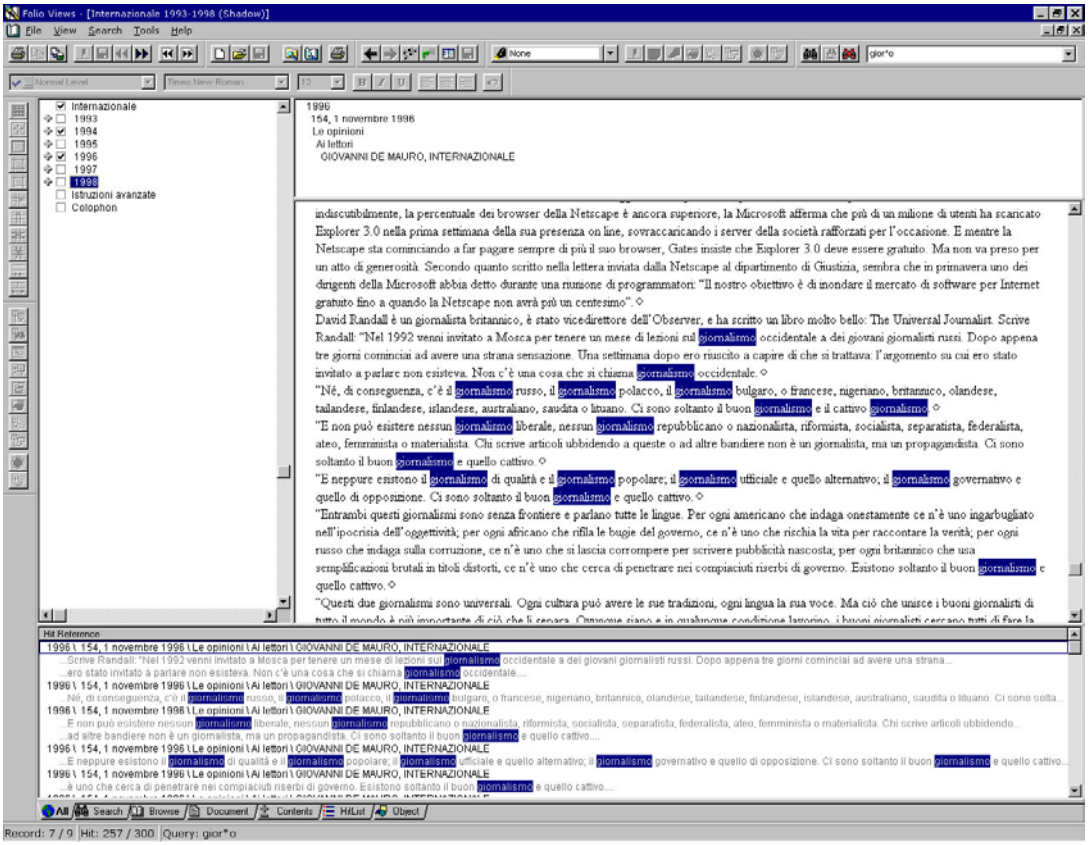
Auch die Ausgabe der Ergebnisse ist viel besser gelöst. So haben wir in einem Fenster den gesamten Text und in einem anderen die Belegstellen zusammen mit der Angabe zu ihrer Herkunft. Der Kontext, der hier zusammen mit dem Schlüsselwort auf einer Zeile ausgegeben wird, lässt sich bis auf 50 Wortformen erweitern. Ein Nachteil ist allerdings, dass das Schlüsselwort nicht, wie bei Konkordanzen üblich, grundsätzlich im Zentrum erscheint und die Belegstellen nicht

---

[11] Während es Ende 2001 noch möglich war, 30 Tage lang die Ausgaben der ersten drei Monate des Jahres 2000 aller Zeitungen der Datenbank gratis zu nutzen, muss jetzt eine Gebühr von zumindest Euro 60 entrichtet werden (cf. <http://www.banchedati.ilsole24ore.com/>).

[12] z. Zt. Euro 7,69 (cf. <http://www.internazionale.it>). *Internazionale* ist 1993 zum ersten Mal erschienen.

nach dem linken oder rechten Kontext sortiert werden können. Zudem lassen sich die Belegstellen nicht zur weiteren Verarbeitung exportieren, sondern können nur gedruckt werden.



Auch bei der *CD-ROM Internazionale 1993-1998* handelt es sich also letztendlich um ein Archiv von Texten, dem zwar eine recht ausgefeilte *Retrieval*-Software beigegeben ist, das aber für linguistische Analysen insgesamt nicht geeignet ist. Korpuserstellung und lexikalische Analysen können davon jedoch mehr als von den ersten beiden Archiven profitieren.

3. Das Korpus der italienischen Zeitungssprache

Was nun die eigentlichen Zeitungskorpora betrifft, so gibt es, wenn ich das richtig sehe, eigentlich nur zwei. Das erste Korpus ist Teil des von der *Evaluations and Language resources Distribution Agency* (ELDA) als *W0023* vertriebenen Komplexes multilingualer und paralleler Korpora (MLLC)[13] und besteht aus Artikeln des Finanzblattes *Il Sole 24 ore* von 1992. Seine Größe wird mit 1,88 Millionen Wortformen angegeben. Das zweite ist das von mir selbst erstellte *Korpus der italienischen Zeitungssprache*. Eine Version desselben ist auch auf

[13] Mitglieder der ELDA zahlen für dieses Korpus EURO 450, Nichtmitglieder EURO 1.200 (cf. <http://www.elda.fr/cata/tabtxt1.htmltext/.html>).

der von der *European Corpus Initiative* (ECI) erstellten CD-ROM enthalten (cf. <http://www.elsnet.org/eci.html>), eine weitere steht über das *Oxford Text Archive* (OTA) zur Verfügung (cf. <http://ota.ahds.ac.uk/textinfo/1723.html>).

Natürlich enthalten auch Referenzkorpora wie etwa das *CORpus di Italiano Scritto* (cf. <http://www.cilta.unibo.it/Portale/RicercaLinguistica/coris_ita.html>) eine den Zeitungen gewidmete Komponente, die selbst wieder als Zeitungskorpus betrachtet werden kann, da es hier aber um eigenständige Zeitungskorpora geht, werde ich mich damit nicht befassen. Stattdessen werde ich im Folgenden nicht zuletzt auch mit dem Ziel, zum praktischen und historischen Verständnis etwas beizutragen, die Erstellung des *Korpus der italienischen Zeitungssprache* etwas näher erklären.

3.1. Hintergrund

Ende der 80er Jahre des 20. Jahrhunderts existierten nur wenige frei verfügbare Korpora und diese waren zumeist dem Englischen gewidmet. Am *Istituto di Linguistica Computazionale* in Pisa war zwar ein Korpus zum Italienischen im Entstehen begriffen, dieses konnte aber, da es der Produktion von Wörterbüchern dienen sollte und Wörterbuchverlage an seiner Erstellung beteiligt waren, nicht für anderweitige Forschungen zur Verfügung gestellt werden. Da eine systematische Untersuchung zur Abgrenzung von Varietäten anhand von Verbalkategorien, wie ich sie plante (cf. Burr 1993) aber nur mit Hilfe des Computers realisiert werden konnte, sah ich mich schließlich 1989 gezwungen, aus den vier italienischen Tageszeitungen *Corriere della Sera*, *La Repubblica*, *La Stampa* und *Il Mattino* selbst ein Korpus zu erstellen. Tageszeitungen wurden v.a. deshalb gewählt, weil ihre Sprache selbst als eine Einheit aus diaphasischen Varietäten gilt. Die Wahl der einzelnen Blätter sollte sowohl der diatopischen als auch der diastratischen Variation Rechnung tragen.[14]

3.2. Die Erstellung des Korpus

1989 gab es noch keine CD-ROM. Zudem war es den meisten Zeitungen, so auch dem *Corriere* und dem *Mattino*, damals nicht möglich, Textmaterial in elektronischer Form zur Verfügung zu stellen. Die Zeitungstexte wurden zwar am Bildschirm erstellt, war die Ausgabe aber einmal gedruckt, wurden sie gelöscht. Die Wochenausgabe der beiden Blätter, die in das Korpus eingehen sollte, musste deshalb eigens digitalisiert werden.

Bei der Turiner *La Stampa* ist es dagegen gelungen, zwei Mitarbeiter von *Sistemi di Processo La Stampa* dazu zu bewegen, die eingetippten Texte einer vorab festgelegten Wochenausgabe eigens zu archivieren. Da es aber nicht mög-

[14] Für eine genauere Beschreibung der Kriterien, die der Korpuserstellung zugrunde gelegt wurden, cf. Burr (1993: bes. 125-157).

lich war, auch die getrennt von den Texten produzierte Titulatur aufzubewahren, musste diese später manuell eingefügt werden.

Allein *La Repubblica* besaß damals schon ein digitales Archiv, in dem die Texte, versehen mit bibliographischen Informationen wie *Tematica*, *Luoghi*, *Persone* etc. und Angaben zu den Textkategorien (*Occhiello*, *Titolo* etc.) abgelegt wurden. Nach langem Hin und Her waren die *Servizi Informativi Redazionali* und das *Centro Documentazione Arnaldo Mondadori* schließlich bereit, eine Wochenausgabe aus diesem Archiv zur Verfügung zu stellen.

Digitalisierung der Zeitungstexte

Liegen Texte nur in gedruckter Form vor, so gibt es zwei Möglichkeiten ihrer Digitalisierung: Eintippen oder Einscannen. Beide haben Vor- und Nachteile. Bei der Erstellung des Korpus wurde das Einscannen mit Hilfe einer OCR-Software gewählt. Wie sich aber bald herausstellen sollte, war die gedruckte Form der Texte des *Corriere* und des *Mattino* dazu, zumindest damals, nicht geeignet. Stattdessen mussten die Texte in ihre einzelnen Kolumnen zerschnitten werden. Zu welchen Problemen es sonst noch kam, zeigt die dritte Spalte: nicht nur werden akzentuierte Lettern und Sonderzeichen nicht richtig interpretiert, sondern die Mängel der Vorlage verhindern das Erkennen ganzer Textstellen.


CORRIERE DI

Commovente abbraccio in Vaticano tra il premier e il Pontefice dopo la grande svolta polacc

La Polonia ringrazia il Papa

E Wojtyla risponde a Mazowiecki: era la mia missione universal

«Speriamo che non si torni indietro» - «Profonda comprensione nei momenti difficili» - Appell del leader: aiuti entro la metà di novembre - Senza condizioni politiche l'intervento dell'Ital

QUEL RICORDO DI DE GASPERI

Papandreu «Vincerò ancora»

Nel cuore del terren

Un enorme squarcio ha minacc

Il presidente del Senato ha affrontato fra l'altro la questione dell'ondata migratoria dai Paesi più poveri rilevando che «i problemi non sono solo economici ma civili, morali e culturali e solo un'azione politica lungimirante può essere in grado di controllarli prima che diventino esplosivi oppure alimentino, come avviene in qualche parte d'Italia, forme preoccupanti e vergognose di razzismo».

Il presidente del Senato si è quindi soffermato sui problemi relativi alla mobilità all'interno dell'area metropolitana, sull'urgenza di assicurare un sistema di efficienti collegamenti con le grandi vie di comunicazione (aeroporti e ferrovie) e con l'hinterland, e ha sottolineato la scadenza europea del 1992 e i problemi che a quella data dovranno essere risolti.

```
  Il pr>-53de~~te del Senato
ha aflrcntat') r           la l altro
<F10,R>utt>+¯L1e <F7,R>u'ø~:>-1~~a ini-
i;r~I<>ria Ll>/'~ 1,-:> pi— po-
veri riie'~andc H' p~1-
blemi non so¯, :`.~l1
nomici ma ci'>1"~ ø--. ----
culturali e solo li' --`-----
politica lungimirani pi®->'-
essere in grado di control-
larli prima che diventino
esplosivi oppure alimenti-
no, come avviene in qual-
che parte d'Italia, forme
preoccupanti e vergognose
di razzismo¯.
  <F8,R>Il <F7,R>presidente del Senato
si Š quindi soffermato sui
problemi relativi alla mo-
bilit… all'interno dell'area
metropolitana, sull'urgen-
za di assicurare un sistema
di efficienti collegamenti
con le grandi vie di comu-
nicazione (aeroporti e fer-
rovie) e con lhilteriand, e
ha sottolineato la scaden-
za europea dal 1992 e i Ero-
blemi che a ouella data do-
vranno essere risolti.
```

Markup-System

Die Entscheidung, mit welchem der damals bei der Erstellung von Korpora verwendeten Markup-Formate die Texte kodiert werden sollten, war ein ziemlich

schwieriges Unterfangen. Erstens entwickelte fast jede/r ein eigenes Markup-System und schrieb dafür eigene Programme, zweitens hing diese Entscheidung mit der einzusetzenden Software zusammen, die allerdings erst dann getestet werden konnte, wenn schon ein entsprechend kodiertes Korpus vorhanden war. Da sich schließlich Micro-OCP als das für die Untersuchung am besten geeignete Programm erwies und dieses nach einem im COCOA-Format kodierten Korpus verlangte, wurde das COCOA-Format gewählt und folgendes Markup-System entwickelt:

Variable	Kodierung	Beispiel
Zeitung als Fragment des Korpus	<Z>	<Z Stampa>
Ausgabe	<E>	<E 211089>
Sparte	<S>	<S Politica>
Positionierung des Textes	<C>	<C MEA01>
AutorIn	<N>	<N Ferrara Anna>
Sprache	<L>	<L Inglese>
Autorenschaft	<A>	
signiert		<A firmato>
anonym		<A Non firmato>
Redaktion		<A Redazione>
Texttype	<T>	
Vorzeile		<T Occhiello>
Schlagzeile		<T Titolo>
Untertitel		<T Sottotitolo>
Zusammenfassung		<T Sommario>
Zwischenüberschrift		<T Catenaccio>
Ankündigung		<T Civetta>
Artikel		<T Articolo>
'Aufmacher'		<T Spalla>
Leitartikel		<T Fondo>
Nachricht		<T Notizia>
Glosse		<T Corsivo>
Interview		<T Intervista>
Kolumne		<T Rubrica>
Kritik		<T Critica>
Kurzmeldung		<T Flash>
Kurznachricht		<T Breve>
Fernseh-, Kinoprogramm		<T Programma>
Filminhalt		<T Film>
Leserbrief		<T Lettera>
Liste		<T Elenco>
Wetterbericht		<T Tempo>
Buch-, Film-, Liedtitel, etc.		<T Nome>
Bildunterschrift		<T Foto>
Art des Sprechens	<P>	
fortlaufender Text		<P Prosa>
Zitat von schriftlicher Quelle		<P Citazione>
Zitat von mündlicher Quelle		<P Discorso>
Frage im Interview		<P Domanda>
Antwort im Interview		<P Risposta>

Zusammensetzung des Korpus

Das Korpus der italienischen Zeitungssprache besteht in der Zwischenzeit aus zwei unterschiedlichen Komponenten und zwar aus dem Korpus *‚Deutsche Einigung 1989'* und aus dem Korpus *‚Europawahlen 1994'*, das zugleich eine Komponente des im Rahmen meines Habilprojektes erstellten *Korpus romanischer Zeitungssprachen* ist (cf. Burr 1997). Die beiden Komponenten setzen sich wie folgt zusammen:

‚Deutsche Einigung 1989'

Zeitung	Quelle	Ausgabe	Wortformen/tokens
Corriere della Sera	Zeitung	19.10.89	88.158
Corriere della Sera	Zeitung	20.10.89	79.030
Corriere della Sera[15]	Zeitung	21.10.89	91.099
Gesamt *Corriere*			258.287
Il Mattino	Zeitung	20.10.89	82.102
Il Mattino	Zeitung	21.10.89	89.399
Gesamt *Mattino*			171.501
La Repubblica	Zeitung	20.10.89	88.961
La Repubblica	Zeitung	21.10.89	85.997
Gesamt *Repubblica*			174.958
La Stampa	Zeitung	20.10.89	69.964
La Stampa	Zeitung	21.10.89	49.807
Gesamt *Stampa*			119.771
Gesamt Korpus			724.517

‚Europawahlen 1994'

Zeitung	Quelle	Ausgabe	Wortformen/tokens
Corriere della Sera	CD-ROM	13.06.94	102.976
Corriere della Sera	CD-ROM	14.06.94	102.441
Corriere della Sera	CD-ROM	15.06.94	98.224
Gesamt *Corriere*			303.641

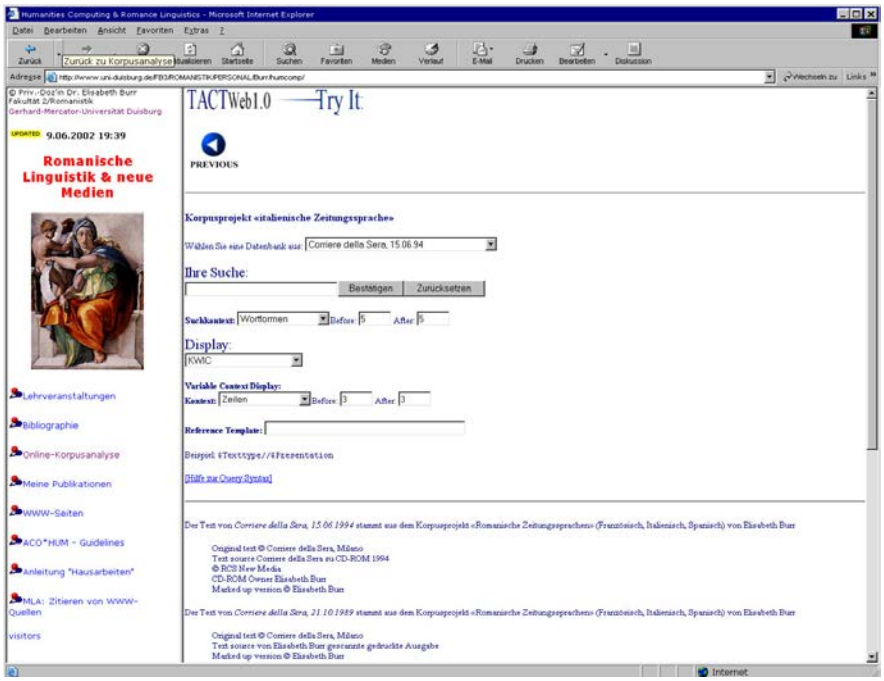
TACTWeb

Als 1998 mit dem Erscheinen von *TACTWeb*, der WWW-Version des von mit Korpora arbeitenden Literatur- und Textwissenschaftlern entwickelten und auch heute noch sehr weit verbreiteten interaktiven Textanalyseprogramms *TACT 2*.1, die Erhebung von Daten aus online-verfügbaren Korpora möglich wurde, stellte ich drei der oben aufgeführten Komponenten des Korpus zur italienischen Zeitungssprache v.a. für meine eigenen Lehrzwecke ins Web: <http://www.uni-duisburg.de/FB3/Romanistik/Personal/Burr/humcomp/>.[16]

In seinem derzeitig Zustand lassen sich anhand dieser Korpuskomponenten alle Arten von Analysen durchführen, die von bestimmten Wortformen ausgehen oder bestimmte formale Regelmäßigkeiten nutzen können. Da die Verbalkategorien in den beiden Teilkorpora von 1989 mit Hilfe eines Codes explizit gemacht wurden (cf. Burr: 1993), sind auch dazu Untersuchungen möglich. Zudem kann aufgrund des Markup bei der Erhebung der Belegstellen nach den von den Zeitungen vorgegebenen nicht-sprachlichen Einheiten differenziert und somit auch die sprachliche Variation berücksichtigt werden.

[15] Diese Ausgabe kam erst nach Abschluss der Untersuchung hinzu, für die das Korpus ursprünglich erstellt wurde.

[16] Zu den verschiedenen Phasen meiner Korpus-basierten Lehre, den damit verfolgten Zielen und den dabei gemachten Erfahrungen cf. z.B. Burr (2001).



4. Schluß

Digitale Textarchive eignen sich, wie wir gesehen haben, vor allem als materielle Grundlagen für die Erstellung von Korpora, d.h. zur Entnahme von Texten nach bestimmten Kriterien. Zudem bieten sie die Möglichkeit, Texte auf ihren Nutzen zu prüfen. So kann es, will ich nur bestimmte Lexeme untersuchen oder Lehrmaterial zu ihrem Gebrauch zusammenstellen, unter Umständen sinnvoll sein, mein Korpus bzw. meine Sammlung auf Texte zu beschränken, in denen diese Lexeme auch tatsächlich vorkommen. Systematische sprachwissenschaftliche Analysen lassen sich mit der Software, mit denen die Archive ausgestattet sind, jedoch nicht durchführen. Es geht hier nun mal nicht um die Erstellung von Konkordanzen, um Kollokationen etc., sondern um ein Auffinden von Texten oder Texstellen zu z.B. bestimmten Themen, d.h. um *Information-* oder *Document-Retrieval* im ureigendsten Sinne.

Damit ist natürlich auch schon gesagt, dass korpuslinguistische Untersuchungen vom Einsatz eines dafür geeigneten Analyseprogramms abhängen. Was aber letztendlich damit überhaupt analysiert werden kann, wird maßgeblich davon bedingt, welche sprachlichen und nicht-sprachlichen Einheiten eines Korpus mithilfe von Markup explizit gemacht wurden. Ohne ein solches Markup kann nämlich auch das beste Analyseprogramm nur Einheiten oder Kombinationen erheben, die sich formal genau abgrenzen lassen. An der Annotierung von Korpora führt also kein Weg vorbei.

Daraus ergibt sich insgesamt auch die weitere Perspektive mit Blick auf das Korpus der italienischen Zeitungssprache. Damit unter Einsatz von entsprechenden Analyseprogramme wie *TACTWeb*, *TACT 2.1*, *Monoconc* etc. nämlich die

Untersuchungen möglich werden, die bisher z.B. wegen der fehlenden Kodierung der *parts of speech*, von Eigennamen etc. nicht oder nur sehr mühsam durchgeführt werden können, ist eine viel weitreichendere Anreicherung mit Markup unter Berücksichtigung der von der *Expert Advisory Group on Language Engineering Standards* (EAGLES) für die Kodierung von Korpora erstellten *Guidelines* (cf. <http://www.ilc.pi.cnr.it/> geplant. Dabei soll dann auch das Markup insgesamt an den von der TEI entwickelten Textkodierungsstandard angepasst werden. Des weiteren sollen Mittel und Wege gefunden werden, das Korpus anhand der schon digitalisierten bzw. in digitaler Form gesammelten Zeitungsausgaben so auszubauen, dass es schließlich aus je einer Wochenausgabe der einzelnen Zeitungen besteht.

Mit nach linguistischen Kriterien aufgebauten und annotierten Korpora sowie für ihre Untersuchung geeigneten Analyseprogrammen ist es aber nicht getan. Stattdessen muss sich in der Linguistik auch eine neue Form der Darstellung von Untersuchungen durchsetzen. Da es einerseits für die Zusammensetzung von Korpora keine Ideallösung gibt und wir zum anderen bei korpuslinguistischen Untersuchungen das konservierte Sprechen als Daten betrachten, müssen nämlich nicht nur die Kriterien der Zusammensetzung des Korpus, sein Umfang und das Markup, mit dem es angereichert wurde, genau beschrieben werden, sondern auch die erhobenen Daten und die bei ihrer Analyse zur Anwendung kommenden Methoden. Korpuslinguistische Untersuchungen verlangen also nach einer wissenschaftlichen Kultur, wie sie den empirischen Wissenschaften eigen ist.

Bibliographie

Aarts, Jan / Meijs, Willem (eds.) 1984: Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research (= Costerus, N.S. 45). Amsterdam: Rodopi

Atkins, Sue B.T. / Zampolli, Antonio (Hrsg.) 1994: Computational Approaches to the Lexicon. Oxford: Oxford University Press

Biber, Douglas 1994: Representativeness in Corpus Design, in: Zampolli / Calzolari / Palmer (Hrsg.) 1994, 377-407

Bilger, Mireille (Hrsg.) 2000: Corpus. Méthodologie et Applications Linguistiques (= Bibliothèque de l'INaLF, Les français parlés - textes et études 3). Paris: Honoré Champion

Blanche-Benveniste, Claire 2000: Introduction, in: Bilger (Hrsg.) 2000, 11-15

Burnard, Lou im Druck: Where did we go wrong? A retrospective look at the British National Corpus, in: Burr im Druck

Burr, Elisabeth 1993: Verb und Varietät. Ein Beitrag zur Bestimmung der sprachlichen Variation am Beispiel der italienischen Zeitungssprache (= Romanische Texte und Studien 5). Hildesheim: Olms

Burr, Elisabeth 1997: Wiederholte Rede und idiomatische Kompetenz. Französisch, Italienisch, Spanisch. Habilitationsschrift, Gerhard-Mercator-Universität GH Duisburg, Fachbereich 3: Sprach- und Literaturwissenschaften (Manuskript)

Burr, Elisabeth 2001: Romance Linguistics and Corpora of French, Italian and Spanish Newspaper language, in: Fiormonte / Usher (Hrsg.) 2001, 85-104

Burr, Elisabeth (Hrsg) in Vorbereitung: Tradizione & Innovazione II u. III. Atti del VI Convegno Internazionale della SILFI, Duisburg 28.06.-02.07.2000. Firenze: Cesati

Burr, Elisabeth (Hrsg.) im Druck: Tradizione & Innovazione I: Il parlato: teoria - corpora - linguistica dei corpora. Atti del VI Convegno Internazionale della SILFI, Duisburg 28.06.-02.07.2000. Firenze: Cesati

Busa, Roberto 1980: The Annals of Humanities Computing. The Index Thomisticus, in: Computers and the Humanities 14, 2: 83-90

Clear, Jeremy 1992: Corpus sampling, in: Leitner (Hrsg.) 1992, 21-31

Engwall, Gunnel 1994: Not Chance but Choice: Criteria in Corpus Creation, in: Atkins / Zampolli (Hrsg.) 1993, 9-82

Fiormonte, Domenico / Usher, Jonathan (eds.) 2001: New Media and the Humanities: Research and Applications. Proceedings of the first seminar Computers, Literature and Philology, Edinburgh 7-9 September 1998. Oxford: HCU - University of Oxford

Francis, W. Nelson (1992): Language corpora B.C., in: Svartvik (Hrsg.) 1992, 17-32

Francis, W. Nelson / Kučera, Henry 1964 / 1979: Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Brown University: Department of Linguistics

Habert, Benoît / Nazarenko, Adeline / Salem, André 1997: Les linguistiques de corpus. Paris. Armand Colin

Leech, Geoffrey 1992: Corpora and theories of linguistic performance, in: Svartvik (Hrsg.) 1992, 105-122

Leitner, Gerhard (Hrsg.) 1992: New Directions in English Language Corpora. Methodology, Results, Software Developments (= Topics in English Linguistics 9). Berlin: de Gruyter

Lenz, Susanne 2000: Korpuslinguistik (= Studienbibliographien Sprachwissenschaft 32). Tübingen: Groos Brigitte Narr

Pusch, Claus D. / Raible, Wolfgang (Hrsg.) im Druck: Romanistische Korpuslinguistik. Romance Corpus Linguistics. Korpora und gesprochene Sprache. Corpora and Spoken Language (= ScriptOralia 126). Tübingen: Narr

Sinclair, John 1991: Corpus, Concordance, Collocation. Oxford: Oxford University Press

Spina, Stefania 2001: Fare i conti con le parole. Introduzione alla linguistica dei corpora. Perugia: Guerra

Svartvik, Jan 1992: Corpus linguistics comes of age, in: Svartvik (Hrsg.) 1992, 7-13

Svartvik, Jan (Hrsg.) 1992: Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991 (= Trends in Linguistics. Studies and Monographs 65): Berlin: de Gruyter

Zampolli, Antonio / Calzolari, Nicoletta / Palmer, Martha (Hrsg.) 1994: Current Issues in Computational Linguistics: In Honour of Don Walker (= Linguistica Computazionale IX-X). Pisa: Giardini

Hochschuldozentin Dr. Elisabeth Burr
Institut für fremdsprachliche Philologien - Romanistik
Gerhard-Mercator-Universität
Geibelstraße 41
47058 Duisburg
e-mail: Elisabeth.Burr@uni-duisburg.de