Paroubek, Patrick/Schabes, Yves/Joshi, Aravind K. (1992), XTAG − a Graphical Workbench for Developing Tree-adjoining Grammars. In: *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP)*. Trento, Italy, 223−230.

Pollard, Carl/Sag, Ivan A. (1994), *Head-driven Phrase Structure Grammar*. Stanford, CA: CSLI, and Chicago: University of Chicago Press.

Rabin, Michael O./Scott Dana (1959), Finite Automata and their Decision Problems. In: *IBM Journal of Research and Development* 3(2), 114−125.

Savitch, Walter J./Bach, Emmon/Marsh, William/Safran-Naveh, Gila (eds.) (1987), *The Formal Complexity of Natural Language*. (Studies in Linguistics and Philosophy 33.) Dordrecht: Reidel.

Shannon, Claude E. (1948), A Mathematical Theory of Communication. In: *Bell System Technical Journal* 27, 379−423 and 623−656.

Stabler, Edward (1997), Derivational Minimalism. In: Retore, Christian (ed.), L*ogical Aspects of Computational Linguistics, LACL'96*. Berlin: Springer, 68−95.

Stede, Manfred (1992), The Search for Robustness in Natural Language Understanding. In: *Artificial Intelligence Review* 6, 383−414.

Steedman, Mark (1985), Dependency and Coordination in the Grammar of Dutch and English. In: *Language* 61, 523−568.

Steedman, Mark (1996), *Surface Structure and Interpretation*. (Linguistic Inquiry Monograph No. 30.) Cambridge, MA: MIT Press.

Steedman, Mark (2000), *The Syntactic Process*. Cambridge, MA: MIT Press.

Tesnière, Lucien (1959), *Eléments de syntaxe structurale*. Paris: Éditions Klincksieck.

Viterbi, Andrew J. (1967), Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. In: *IEEE Transactions on Information Theory* 13(2), 260−269.

Weaver, Warren (1949 [1955]), *Translation*. Reprinted in: Locke, William Nash/Booth, Andrew Donald (eds.), *Machine Translation of Languages: Fourteen Essays*. New York: Wiley, 15−23.

Younger, Daniel H. (1967), Recognition and Parsing of Context-free Languages in Time n^3. In: *Information and Control* 10(2), 189−208.

van Zaanen, Menno (2000), ABL: Alignment-based Learning. In: *Proceedings of the 18th Conference on Computational Linguistics*, Volume 2. Saarbrücken, Germany, 961−967.

Stefanie Dipper, Bochum (Germany)

6. Corpus linguistics and sociolinguistics

1. Introduction

Sociolinguistics and corpus linguistics share a natural affinity. There is a sense in which one could say that sociolinguistics is corpus linguistics, at least with respect to one prominent branch of sociolinguistics devoted to the study of spoken and written language in

context. One goal of this kind of sociolinguistics is to compile a corpus of data suitable for quantitative analysis of linguistic and social variables, such as social class, gender, region, ethnicity, style, and age. Although sociolinguistics began before the use of electronic corpora and computers became widespread, today new technologies assist and enhance methods linguists and philologists have used for a very long time. Like many of the great grammarians, lexicographers, and dialectologists, the earliest sociolinguists worked from manually compiled and analyzed corpora (cf. article 1). Most of these consist of tape recordings and transcriptions (often not in electronic form) that are not in the public domain.

Despite the fact that most contemporary sociolinguists use computers to analyze the data they collect, and store it in electronic databases, most still design and compile their own corpora based on the particular variables under investigation and annotated for their own specific purposes (cf. articles 9 and 53) rather than rely on commercially available electronic corpora. There are a variety of reasons for this. Perhaps the main one is the emphasis within corpus linguistics on standard written forms of language. Texts found within most corpora do not contain the kind of material of greatest interest to most sociolinguists, namely, casual everyday speech, often from non-standard language varieties. Large corpora of spontaneously occurring spoken data are still expensive and time-consuming to compile due to problems of transcription and input (cf. articles 11 and 47).

This article provides examples of how one can use existing corpora to investigate some common social variables based primarily on English because it is the language for which the largest collections of data exist, much of it acquired for academic, industrial or commercial research (cf. article 20). However, resources for corpus-based sociolinguistic research on other large European languages such as German, French, Spanish, and smaller ones such as Dutch are steadily increasing. The Institute for German language in Mannheim houses 38 spoken corpora in its Archive for spoken German (Archiv für Gesprochenes Deutsch or AGD) (http://www.ids-mannheim.de/ksgd/agd/). The Meertens Institute in Amsterdam has a unit devoted to variationist studies (http://www.meertens.knaw.nl/meertensnet/wdb.php?url=/variatielinguistiek/), and the Institute for Dutch lexicology has a number of electronic corpora (http://www.inl.nl). The Spanish Royal Academy of Language makes available on-line its *Diccionario de la Lengua Española* (Dictionary of the Spanish Language) and the *Banco de datos del español* (Spanish language database) (http://www.rae.es). Pusch (2002) provides an overview of Romance language corpora (cf. article 21 for other languages). An increasing number of parallel corpora also present opportunities for sociolinguistic research (cf. article 16). The Europarl Corpus of proceedings from the European Parliament features 11 languages (French, Italian, Spanish, Portuguese, English, Dutch, German, Danish, Swedish, Greek and Finnish). The Nordic Teenage Language Project (UNO), a network of researchers investigating the language of teenagers, have compiled or made use of corpora in various Nordic languages (http://www.uib.no/uno/). Hasund's (2002) comparison of the discourse markers *like* and *liksom* among English and Norwegian teenagers relied on a Norwegian and English corpus of teenage language (see also Hasund/Stenström 2005).

The principal social dimensions sociolinguists have been concerned with are social class, age, ethnicity, sex, and style. Of these, social class has been one of the most researched. Most sociolinguists take as their starting point the notion that social stratification will be an important dimension in accounting for linguistic variation in all speech

communities. Most studies have employed what can be referred to as quantitative variationist methodology (sometimes also called the quantitative paradigm or variation theory) to reveal and analyze sociolinguistic patterns, i. e. correlations between variable features of the kind usually examined in sociolinguistic studies of urban speech communities, such as post-vocalic /r/ in New York City (Labov 1966), initial /h/ in Norwich (Trudgill 1974), etc., and external social factors (e. g. social class, age, ethnicity, sex, network, and style). A major finding of urban sociolinguistic work is that differences among social dialects are quantitative and not qualitative. The usual sorts of queries/searches routinely performed on corpora produce various kinds of data that can be analyzed using sociolinguistic methods (Milroy/Gordon 2003). The occurrence of words, word forms, constructions, etc. can all be correlated with the usual social variables investigated by sociolinguists whenever corpora provide reliable information on the social categories of users. A number of studies of discourse phenomena ranging from intonation, pragmatic particles and discourse markers to conversational routines have been carried out using corpora (see Aijmer 2002; Aijmer/Stenström 2004; cf. article 49).

2. Corpora of particular interest to sociolinguists

The following list gives no more than a brief hint at some of the currently available corpora that might be of interest to sociolinguists, some of which will be used to illustrate the discussion of variables in this article (cf. articles 10, 11 and 20 for fuller lists).

2.1. The British National Corpus (BNC)

100 million words of written (90%) and spoken (10%) British English from the 1990s (http://www.natcorp.ox.ac.uk/). The corpus is annotated with metadata pertaining to demographic variables such as age, gender and social class, and textual features such as register, publication medium and domain. The spoken part includes informal, unscripted conversation by speakers of different ages, regions, and social classes, as well as spoken language from formal meetings, radio shows, phone-ins, and other situations (see Aston/Burnard 1998, chapter 6 for examples of how to use the corpora for analyzing social variables). The spoken texts in the corpus include both men and women from three geographic regions: south, midland, north. The speakers are further classified according to age (0−14; 15−24; 25−34; 35−44; 45−59; 60+) and social class. The BNC categorizes social class membership into four groups based on occupation, a commonly used indicator of socio-economic status. From highest to lowest ranked, these are: AB (top or middle management, administrative or professional), C1 (junior management, supervisory or clerical), C2 (skilled manual), DE (semi-skilled and unskilled manual). Unfortunately, this demographic information is not given for all speakers in all texts but is unevenly distributed across the corpus. This limits the use that can be made of the corpus and the conclusions that can be drawn about social variables. Only about 20% of the material in the spoken component is coded for the speaker's social class and education. The only speakers for whom the social class coding can be trusted are the recruited

respondents who were asked to record conversations. Similarly, one must be careful when using the corpus to look at regional variation because the corpus codes the region where the recording was made, not the variety used by the speakers.

2.2. Brown Corpus of American English (Brown)

1 million words of written American English from 1961. This corpus provided a model for a set of parallel corpora (LOB, Frown and FLOB), all of which contain a million words and are constructed in parallel fashion so that they contain 500 word samples from 15 genres of written text. Brown and LOB (Lancaster-Oslo/Bergen Corpus of British English) represent American and British English in 1961 (http://khnt.hit.uib.no/icame/manuals/lob/index.htm), while Frown (Freiburg Brown Corpus of American English) and FLOB (Freiburg LOB Corpus of British English) were compiled at the University of Freiburg as matching databases representing the state of the two varieties in 1992 and 1991 respectively. These and other widely used corpora are distributed by the International Computer Archive of Medieval and Modern English (ICAME). Further information and on-line versions of the manuals are available at http://nora.hd.uib.no/icame.html.

2.3. Australian Corpus of English (ACE)

1 million words of written Australian English compiled in 1986 as a parallel corpus to Brown.

2.4. Wellington Corpus of Written and Spoken New Zealand English (WCNZE)

1 million words of spoken and written New Zealand English compiled in 1986–1990 as a parallel corpus to LOB.

2.5. London-Lund Corpus of Spoken English (LLC)

1 million words comprising 200 samples of 5000 words of spoken and written English collected from 1959 to 1988. The spoken texts contain both dialogue and monologue. The written texts include not only printed and manuscript material but also examples of English read aloud, as in broadcast news and scripted speeches.

2.6. International Corpus of English (ICE)

In 1990 the International Corpus of English began to assemble parallel one million word corpora of spoken and written material from 20 major varieties of English spoken

around the world. Each corpus follows a standard design and grammatical annotation, thus permitting the examination of regional variation (http://www.ucl.ac.uk/english-usage/ice/).

2.7. American National Corpus (ANC)

In progress, 100 million words of spoken and written American English parallel to BNC (http://americannationalcorpus.org/).

2.8. Corpus of Spoken, Professional American-English (CSPA)

Short conversational interchanges recorded between 1994 and 1998 from ca. 400 speakers centered on professional activities broadly tied to academics and politics, including academic politics (http://www.athel.com/cpsa.html).

2.9. Corpus of London Teenage Language (COLT)

500,000 words of spontaneous conversations between 13 to 17 year old boys and girls from socially different school districts in London (http://torvald.aksis.uib.no/colt/). In 1994−95 the conversations were transcribed orthographically, and tagged for word-classes by a team at Lancaster University. In this form, COLT became part of BNC.

3. Investigating sociolinguistic variables using corpora

Variationist methodology came into prominence in the late 1960s primarily to fill perceived gaps in traditional studies of variability which for the most part were concerned with regional variation. Dialectologists in the 19th and early 20th centuries concentrated their efforts on documenting the rural dialects which they believed would soon disappear. A primary concern was to map the geographical distribution across regions of forms that were most often different words for the same thing, e. g. *dragon fly* v. *darning needle*; some phonological and grammatical features were also included. The results often took many years to appear in print and were generally displayed in linguistic atlases of maps showing the geographical boundaries between users of different forms (cf. articles 1 and 53). More recently, some of these projects have made some of their material available in electronic form for downloading and/or on-line searches. The website for the Linguistic Atlas Projects contains an overview of these projects and the materials collected in various regions of the United States (http://hyde.park.uga.edu/). In addition to regional variation, it is possible to use some of the data to analyze other kinds of variation of interest to sociolinguists. The informants for the various surveys were classified according to social criteria (degree of formal education, occupation, age, sex).

By contrast, sociolinguists turned their attention to the language of cities, where an increasing proportion of the world's population lives in modern times. Aided by the mass-production of recording equipment, sociolinguists collected spoken data that were transcribed and analyzed, paying attention to easily quantifiable linguistic features, e. g. post-vocalic /r/ in words such as *cart*, etc. Most of the variables studied in detail have tended to be phonological, and to a lesser extent grammatical, although in principle any instance of variation amenable to quantitative study can be analyzed in similar fashion. Counting variants of different kinds in tape-recorded interviews and comparing their incidence across different groups of speakers revealed that when variation in the speech of and between individuals was viewed against the background of the community as a whole, it was not random, but rather conditioned by social factors such as social class, age, sex and style in predictable ways. Thus, while idiolects (or the speech of individuals) considered in isolation might seem randomly variable, the speech community as a whole behaved regularly. Using these methods, one could predict, for example, that a person of a particular social class, age, sex, etc. would pronounce post-vocalic /r/ a certain percent of the time in certain situations. Some variables are unique to particular communities, while others are shared across the English-speaking world. The replication of a number of sociolinguistic patterns across many communities permits some generalizations about the relationship between linguistic variables and society (Romaine 2000).

3.1. Region

The so-called 'first generation' corpora (Brown, LOB etc.) along with ICE and BNC are ideal for comparing features across different varieties of English. They provide a rich source of information on lexical, spelling and grammatical differences among the major regional varieties of English. Table 6.1 compares the use of *film* vs. *movie* and *journey* vs. *trip* in Brown, Frown, LOB and FLOB. Results are given in terms of number of hits as well as in the form of a ratio calculated by dividing the number of hits for *movie*/*trip* by the number of hits for *film*/*journey* respectively. A ratio of more than 1.00 indicates that *film*/*trip* are more common than *movie*/*journey,* and a ratio of less than 1, that *film*/*journey* are more common.

The corpus results do not bear out the common assumption that *movie* is preferred over *film* in American English, either in 1961 (Brown) or 1991 (Frown). Although the rate of occurrence of *movie* increases in relation to that of *film* in Frown, *film* is still the

Tab. 6.1: Comparison of *film*/*movie* and *journey*/*trip* in four corpora

	N of hits			N of hits		
Corpus	*film*	*movie*	Ratio	*journey*	*trip*	Ratio
Brown	126	67	.53	30	109	3.63
Frown	178	119	.67	3.4	85	2.5
LOB	243	7	.03	69	45	.65
FLOB	119	41	.34	66	74	1.12

preferred term in both British and American English. Comparing LOB and FLOB, however, shows that *movie* is increasing at the expense of *film.* In the case of *journey/trip*, however, American usage favors *trip* in both Brown and Frown, while British English favors journey in LOB, but not FLOB, where *trip* is more common than *journey.* Data from BNC, however, suggest that *trip* is slightly favored over *journey* only in spoken but not written English. There are 236 hits for *journey* and 256 for *trip* in the spoken component. The word *film* is preferred over *movie* in both the spoken and written components.

The comparisons can be extended by considering ACE and WCNZE. In some instances Australian usage aligns itself with the norms of American English, preferring, for example, *movie* over *film* and *trip* over *journey*, but in other cases, with that of British English, favoring, for example, *holiday* over *vacation*. The use of *movie* is less common in New Zealand than Australia, while the preference for *trip* over *journey* is in line with the Australian tendency towards the American variant *trip*, as is the greater use of *holiday* over *vacation*. Australian English is also like American English in disfavoring the use of the suffix *-st* on *while* and *among*. With respect to spelling, there are also divergent tendencies, with <or> on the increase in Australian English, e. g. *color.* By mid-1985 six of Australia's major urban newspapers used the American <or> spellings, but when it comes to words with <re> instead of <er>, e. g. *theatre,* both Australia and New Zealand favor the British variant. Although most Australians have learned at school to take an anti-American stance in language, especially in spelling, it is not necessarily the case that Australian English is becoming unilaterally more Americanized (Peters 1998).

Similarly, in Britain departures from British spelling norms in favor of American ones have not been welcomed in all quarters and have attracted attention. When in 2000 it was suggested that Britain should adopt internationally standardized spellings of scientific terms, such as *fetus* and *sulfate* (instead of *foetus* and *sulphate*), there were complaints. Looking at BNC, it is evident that the American spelling *fetus*, has already made considerable inroads into written British English. Just over one third (36%) of the 353 examples follow the American spelling, and 64% use the traditional British spelling *foetus*. The trend for *sulphate/sulfate*, however, runs in favor of the British spelling *sulphate*; only 3% of the 410 occurrences of the word use the American spelling *sulfate.* In the case of other words such as *globalization/globalisation*, the American spelling predominates in 63% of the 64 occurrences, and the British variant, *globalisation* is in the minority with 37%.

Indeed, the very occurrence of this term can be used as an index of the spread of a new term and the process of globalization in world English. It is a truism that the history of words offers a window into the history of a language. A closer examination of the history of the word *globalization* and its spread is itself instructive of the forces that many now seek to understand. Corpora can be used to show how linguistic changes having their origin in social and cultural developments can be manifested in vocabulary. Neither Brown nor Frown contains any occurrences of the word; nor does LOB. FLOB contains only one example, but the BNC has 64.

These findings are interesting in the light of Giddens's (2000, 25) comment that the term *globalization* came seemingly from nowhere and now it is everywhere. Although the word *global* is over 400 years old, the terms *globalization* and *globalize* began to be used in the 1960s, and spread thereafter, especially in 1980s onwards. This is reflected in the corpus findings. It is also sometimes said that globalization is moving the world inexorably toward greater homogeneity in the direction of American language and cul-

ture, and that the normative basis for World English has shifted from British to American English. In this view the global village has become a homogenized McWorld, where everyone speaks English, drinks Coke, and eats at McDonalds. Although this is clearly an exaggerated view of the extent of (American) English influence, there is little evidence of a wholesale shift towards American norms. Global English is still best described as a 'pluricentric' language, i. e. one whose norms are focused in different local centers, capitals, centers of economy, publishing, education and political power.

3.2. Social class

In the mid 1950s Ross (1980) suggested that certain lexical and phonological differences in English could be classified as U (upper class) or non-U (lower class), e. g. *serviette* (non-U) vs. *table-napkin* (U), to take what was then one of the best known of all linguistic class-indicators of England. Other notable pairs he mentioned were *have one's bath* (U) vs. *take a bath* (non-U), *writing paper* vs. *note paper* (non-U), *pudding* (U) vs. *sweet* (non-U), or what would be called *dessert* in the US. Such claims can be tested against corpora such as BNC that include information about the social status of speakers. Compare the results in Table 6.2 for *settee/couch/sofa* and *lounge/living room/sitting room*. For each term, the number is bold-faced for the social group showing the highest usage.

Tab. 6.2: Social distribution of selected lexical items in BNC: hits/million words for *settee/sofa/couch* and *lounge/living room/sitting room*

Social class	*settee*	*sofa*	*couch*	*lounge*	living room	sitting room
AB	12.32	2.7	0	11.09	12.32	**13.55**
C1	18.02	2.57	5.5	32.18	9.01	9.01
C2	13.98	**8.39**	**8.39**	**48.93**	13.98	5.59
DE	**31.21**	4.46	0	8.92	**22.9**	8.92

p is less than or equal to 0.01; distribution is significant for *settee/sofa/couch*.
p is less than or equal to 0.001; distribution is significant for *lounge/living room/sitting room*.

Looking first at variation in terms for the item of furniture, all four social groups use both *settee* and *sofa*; the term *couch* does not occur for the highest and lowest social group. The lowest social group strongly favors the term *settee*; the highest social group uses that term least. The term *sofa* occurs most frequently among class C2 followed by DE, but is less often used by two highest classes AB and C1. As for the room where this item of furniture is found, all social groups use all three terms. The middle and lower middle classes (C1 and C2), however, are the greatest users of the term *lounge*. The highest group leads in the use of the term *sitting room*, and the lowest in the use of the term *living room*. Thus, the upper class displays a tendency to sit in the sitting room, while the working class is more likely to sit on a settee in the living room, and the middle class to sit either on a sofa or couch in the lounge.

3.3. Gender

In a pioneering work on the relationship between language and gender, Lakoff (1975) suggested that women made use of a larger color vocabulary than men. In particular, she noted that women were more likely to use non-basic color terms such as *mauve*, *beige*, etc. as well as more secondary color terms such as *sky blue*, *pale green*, *hot pink*, etc. She also said that women used a different set of evaluative adjectives she called 'empty adjectives' more frequently than men, including words such as *lovely*, *divine*, *adorable*, *sweet*, *cute*, etc. Her claims can readily be tested with corpora such as BNC that include information on the sex of the speaker/author. Table 6.3 shows the distribution of the color terms *mauve*, *beige*, *pink*, *maroon* and the use of the descriptive adjective *pale* followed by a color term, along with three evaluative adjectives (*lovely*, *nice* and *cute*). The results for color words are not statistically significant, probably because the number of occurrences is small; *mauve*, for instance occurred only 13 times, and *beige* only 9. Yet the general trends are still in line with Lakoff's suggestions. Women used *mauve* 11 times and men only 2. For *beige*, there were 18 occurrences split equally between men and women. The results for the use of the three adjectives are significant. Indeed, *lovely* and *nice* are among the 25 most frequently used words by women in the spoken BNC (Rayson/Leech/Hodges 1997). The word *adorable*, however, occurred only three times in the spoken corpus and all users were male.

Tab. 6.3: Frequency per million words of selected color terms and evaluative adjectives in spoken component of British National Corpus

	Female	Male
mauve	3.37	.41
beige	2.75	1.83
pink	59.68	25.61
maroon	3.37	.61
pale + color term	4.28	2.44
lovely	437.04	135.15
nice	998.33	445.87
cute	10.1	2.85

p is less than or equal to 0.01; distribution is significant for evaluative adjectives.

Lakoff, along with a number of researchers, suggested that women used more standard forms and that they avoided 'bad' and 'taboo' expressions. The swear words *fuck* and *fucking* are among the most 25 most frequent words used by men in the spoken component of the BNC (Rayson/Leech/Hodges 1997). Stenström (1991) found that in LLC women used proportionally more weaker expletives such as *heavens* than men, as indicated in Table 6.4.

Because situation is an important variable, it is crucial to compare only data collected in comparable communicative contexts, e. g. mixed sex groups vs. single sex groups, etc.

Tab. 6.4: Some swear words used by men and women in the London-Lund Corpus (adapted from Stenström 1991)

	Female	Male
heavens	7.35	4.67
damn	36.73	29.02
blimey	22.04	10.57
fuck	32.75	68.28

p is less than or equal to 0.001; distribution is significant.

(cf. articles 9 and 49). Talk between men in a pub, women in a kitchen, between a male interviewer and female interviewee, or among men watching a football match on TV represent instances of situations that may affect amount and type of data obtained.

3.4. Style

Style is a notoriously difficult term to define, but at its simplest, variation between genres, text types, etc. can be thought of as kinds of stylistic differences. One of the first observations made by early corpus linguists working with the first generation of computerized corpora was that syntactic constructions such as the passive were unevenly distributed across text types. Svartvik (1966, 155) found that their rate of occurrence in the Survey of English Usage comprising the written component of LLC ranged from a low of 3.2/1,000 words in advertisements to a high of 23.1/1,000 words in scientific texts. In the corpus as a whole they occurred at a rate of 11.3/1,000 words, as shown in Table 6.5.

Tab. 6.5: Passives per 1,000 words in the Survey of English Usage (adapted from Svartvik 1966, 155, Table 7.4)

Genre	Hits/1,000 words
Science	23.1
News	15.8
Arts	12.7
Speech	9.2
Sports	9.0
Novels	8.2
Plays	5.3
Advertising	3.2
Whole Corpus	11.3

Many studies have investigated differences between speech and writing, examining features such as negation, contraction, etc. Verb contraction is more frequent in speech than in writing, as can be seen in Table 6.6 comparing the frequency of contraction of *be* and *have* in written and spoken components of the BNC. The ratio is calculated by dividing the number of contracted forms by the number of uncontracted forms. A ratio of more than 1.00 indicates that the contracted form is more common than the uncontracted form. In the spoken texts all the ratios are higher than those for writing; three (*'m*, *'s*, *'s*) exceed 1.00. Even in the written texts the contracted first person singular form *I'm* for *I am* is more common than the uncontracted form.

Tab. 6.6: Ratio of contracted and uncontracted forms in the written and spoken components of the BNC (adapted from Leech/Rayson/Wilson 2001, 130)

	Speech			Writing		
	contracted	uncontracted	ratio	contracted	uncontracted	Ratio
'm:am	2512	252	9.97	443	250	1.77
're:are	4255	4663	.91	439	4712	.09
's:is	15818	10164	1.56	1729	1729	.17
'd:had	575	2835	.20	284	4639	.06
's:has	1844	1598	1.15	119	2708	.04
've:have	4637	7488	.62	440	4416	.10

Other more sophisticated analyses of vocabulary are possible, but as these go beyond simple word/phrase searches, they require more effort (see articles 38 and 50). One such study examined the density of Latinate diction as a stylistic index in the collected speeches, letters and internal monologues of the characters in Jane Austen's novels. The study required assembling an electronic corpus of Austen's work (relying on the Oxford Electronic Text Library Edition of *The Complete Works of Jane Austen*). Such corpora of the texts of individual authors can nowadays be easily assembled form a variety of text banks, databases and archives. The study also required a way of identifying and counting words of Latinate origin, e. g. *artist*, *deception*, etc. This was done by means of a program called JALATIN devised by the researchers, which revealed that overall just over 36% of the words used by Austen were of Latinate origin. There was, however, considerable variation among and within the novels.

Compare these two extracts from Austen's *Mansfield Park* (1814) contrasting the manor at Mansfield belonging to Fanny Price's uncle with her parents' house in Portsmouth.

- The elegance, propriety, regularity, harmony,- and perhaps, above all, the peace and tranquillity of Mansfield, were brought to her remembrance every hour of the day, by the prevalence of everything opposite to them here.
- Every body was noisy, every noise was loud. Whatever was wanted, was halloo'd for, and the servants halloo'd out their excuses from the kitchen. The doors were in constant banging, the stairs were never at rest, nothing was done without a clatter.

When Fanny Price is exiled to Portsmouth to live with her parents in a squalid noisy house, she pines there for her uncle's elegant manor. The Latinate words convey the stately atmosphere of the house, while the Germanic words suggest the chaos and squalor prevailing in her parents' home.

The study actually followed a long tradition of similar stylistic investigations done by earlier scholars who did not have the advantage of modern methods relying on computers and corpora of electronic texts, but who nevertheless examined the proportion of Germanic vs. Romance vocabulary used by influential authors such as Chaucer, who introduced many French words in his works. English has a long tradition of extending its lexical resources through borrowing words from other languages, particularly Latin and French. Historians of English have discussed the impact of these borrowings on English, both in terms of their tendency to cluster in certain semantic domains, e. g. science and technology, as well as in terms of the addition of new roots and their derivational system (cf. *happiness* and *felicity*). As soon as French and Latin words were borrowed, native prefixes and suffixes were added to them, and when a sufficient number of foreign words were borrowed for their word formation patterns to be transparent and isolable, they could be used productively with both native and newly borrowed foreign words. Pairs such *dine/eat*, *commence/begin,* etc. illustrate social and stylistic stratification. The native Germanic members of these doublets are in everyday use, while the borrowings represent a higher, more refined stylistic level. Such choices can then be used by speakers/writers as stylistic resources. Authors such as Chaucer experimented with competing forms such as *frailness* vs. *frailty, stableness/stability/mutability,* etc.

Austen's novel *Pride and Prejudice* (1813) features a range of characters, who differ in the extent to which they use Latinate vocabulary. Table 6.7 shows the percentage of Latinate vocabulary used by the narrator and three women in the Bennett family. Mary, for instance, is bookish and pretentious and, not surprisingly, has the highest index of Latinate words, or for that matter of any character in any Austen novel. Lydia and Kitty Bennett, on the other hand, do not speak like well-educated characters and are at the opposite end of the stylistic and social spectrum. A low index of Latinate vocabulary is an index of low educational level, or low birth or both.

Tab. 6.7: Percentage of Latinate words used by characters in Austen's *Pride and Prejudice* (adapted from DeForest/Johnson 2000, 25)

Character	% of Latinate words
Mary Bennett	33.8
Lydia Bennett	6.3
Kitty Bennett	4.3
Narrator	25.4
All females	19.3

3.5. Age

The age distribution of a variable may be an important clue to on-going change in a community (see article 52). Some patterns of 'age grading' (i. e. variation correlated to

age) may reflect a passing fad (e. g. teenage slang), or be repeated anew in each generation (e. g. swearing by young males) and not lead to long-term change in the community as a whole (see Stenström/Andersen/Hasund 2002). In other cases, however, age grading or change in apparent time may lead to change in real time (Bauer 2002). Once new variants spread, they often follow predictable paths through social and linguistic structures, as new members adopt an innovation.

As a simple example of age-grading, take the distribution of the word *wireless* 'radio' in the spoken component of the BNC shown in Table 6.8. It is used only by those over 25, and even then only infrequently (N = 14) at a rate of 2.37 times per million words. It is most frequent in the oldest age group comprising those over 60. The more frequent term for all age groups is *radio*, especially in the younger age groups. The slang term *tranny* for 'transistor radio' is nearly obsolete, occurring only 11 times in the whole corpus of 100 million words. The BNC is not recent enough to show many instances of the new meaning of *wireless* that has arisen to refer to a variety of new wireless mobile communication devices such as wireless internet service, etc. To document such new uses it would be profitable to use the Web itself as a corpus, but that method would not be able to uncover the age and social distribution of the users (see article 18).

Tab. 6.8: Occurrence of *radio* and *wireless* by age group in spoken component of the BNC

Age	Number of hits		Hits/million words	
	radio	*wireless*	*radio*	*wireless*
60+	57	7	50.47	6.2
45−59	87	2	53.55	1.23
35−44	94	2	88.09	1.87
25−34	58	3	52.19	2.7
15−24	23	0	38.97	0
0−14	34	0	88.72	0

p is less than or equal to 0.05; distribution is significant.

Table 6.9 shows a similar age-graded distribution for *movie.* As suggested in section 3.1., *movie* may be increasing at the expense of *film.* The most frequent users are under 25, and the word is especially common among the youngest age group of those 14 and under. To confirm this trend, one would need to monitor usage over the coming years.

Another word that shows an age graded distribution is *bollocks.* Indeed, it is one of the ten most frequently used 'dirty' words in COLT, with differences between boys (58 instances) and girls (32 instances) (Stenström/Andersen/Hasund 2002, 32). Rayson/Leech/Hodges (1997) also found that *fucking*/*fuck* were among the words more frequently used by those under 35 in the spoken component of BNC.

One can also use parallel corpora collected at different points in time such as Brown/Frown and LOB/FLOB to investigate change in real time (cf. article 52). Holmes (1999) compared these four corpora with the written component of the Wellington Corpus of New Zealand English to investigate Lakoff's (1975) claim that the term *lady* (which she considered a patronizing, trivializing, non-sexual, polite euphemism for *woman*), was in the process of replacing *woman*. Holmes found that references to adult females had more

Tab. 6.9: Distribution of *movie* by age in spoken component of the BNC

Age group	Hits/million words
0–14	33.92
15–24	8.47
25–34	4.5
35–44	4.69
45–59	3.08
60+	3.54

Tab. 6.10: Distribution of *bollocks* by age in the spoken component of the BNC

Age group	Hits/million words
0–14	60.02
15–24	91.48
25–34	13.5
35–44	3.75
45–59	1.85
60+	3.54

than doubled overall, but this increase was not due to a rise in the use of the term *lady*/*ladies*, whose number of occurrences had barely altered over the 30 years between the appearance of Brown/LOB and Frown/FLOB.

4. Correlations among variables and the social embedding of variation and change

Some of the same linguistic features figure in patterns of both regional and social dialect differentiation at the same time as they also display correlations with other social factors. Generally speaking, the use of non-standard forms increases, the less formal the style and the lower one's social status. All groups recognize the overt greater prestige of standard speech and shift towards it in more formal styles. Another sociolinguistic pattern is that women, regardless of other social characteristics such as class, age, etc., tend to use more standard forms than men.

Berglund (1999) found evidence of such classic sociolinguistic patterns in her study of variation in the BNC between the phonologically condensed form *gonna* and the full form *going to*. That is, the form *gonna* was more frequent in the spoken component, in informal contexts, among the youngest two age groups, and men. Table 6.11 shows the

Tab. 6.11: Percent of *gonna* in the BNC for men and women in formal and informal style (adapted from Berglund 1999)

Style	Men	Women
Informal	81	70
Formal	45	26

interaction between style and gender; *gonna* is most frequently used in male informal speech and least in female formal speech.

Similar patterns can be found for other variables in BNC. McEnery/Xiao (2004) examined the occurrence of one common swear word (and its morphological variants) within and across all the spoken and written registers in BNC. They found the use of the word *fuck* to be more frequent in speech than writing, among men than women, among young people and teenagers more than among those over 35, and among the two lower social classes. In addition, their findings for the spoken component suggest that swearing may be increasing among women compared to Stenström's (1991) finding for LLC (see Table 6.4).

5. Limitations of corpora for sociolinguistic research

Although the availability of public corpora greatly increases the range of variables that can be studied in English and other languages, corpora also severely limit the phenomena that can be investigated to those that are most easily retrievable (cf. article 33). There are two reasons why many large public corpora are not well suited to the kinds of analysis undertaken by sociolinguists. Firstly, most corpora are composed primarily of written material in standard English and other standardized language varieties and are best suited to the study of lexical and grammatical variation. Sociolinguists, however, have been concerned primarily with non-standard spoken varieties. Secondly, there is often little or no information on many of the social variables such as class, ethnicity, gender, age, etc. that sociolinguists are most interested in. Nevertheless, the increasing availability of corpora of spoken language, often enhanced with sound files, has opened up possibilities for sociolinguistic analysis (cf. article 11). Despite this, even where phonetically transcribed corpora exist, automatic search and retrieval of the kind of variables of interest to sociolinguists can be extremely difficult; each token of a variable may have innumerable variants and sound files may not always be available (cf. articles 11 and 53).

Studies that would once have taken many years to complete can now be conducted more rapidly and have opened up linguistic phenomena to empirical investigation on a scale previously unimaginable. This article has illustrated how corpora can be used to test hypotheses and to examine the occurrence of many variables in relation to the parameters encoded.

6. Literature

Aijmer, K. (2002), *English Discourse Particles. Evidence from a Corpus*. Amsterdam: John Benjamins.

Aijmer, K./Stenström, A.-B. (eds.) (2004), *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: John Benjamins.

Aston, G./Burnard, L. (1998), *The BNC Handbook − Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bauer, L. (2002), Inferring variation and change from public corpora. In: Chambers, J. K./Trudgill, P./Schilling-Estes, N. (eds.), *Handbook of Linguistic Variation and Change*. Oxford: Blackwell, 97−113.

Berglund, Y. (1999), *Gonna* and *Going to* in the Spoken Component of the British National Corpus. In: Mair, C./Hundt, M. (eds.), *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi, 35−51.

DeForest, M. M./Johnson, E. (2000), Computing Latinate Word Usage in Jane Austen's novels. In: *Computers and Texts* 18/19, 24−25.

Giddens, A. (2000), *Runaway World*. London: Routledge.

Hasund, I. K. (2002), 'Congratulations, like!' − 'Gratulerer, liksom!' Pragmatic Particles in English and Norwegian. In: Breivik, L. E./ Hasselgren, A. (eds.), *From the COLT's mouth … and others'. Language and Corpora Studies in Honour of Anna-Brita Stenström.* Amsterdam: Rodopi, 125−139.

Hasund, I. K./Stenström, A.-B. (2005), Conflict Talk: A Comparison of the Verbal Disputes between Adolescent Females in Two Corpora. In: Sampson, G./McCarthy D. (eds.), *Corpus Linguistics: Readings in a Widening Discipline.* London: Continuum, 326−334.

Holmes, J. (1999), *Ladies* and *Gentlemen*: Corpus Analysis and Linguistic Sexism. In: Mair, C./ Hundt, M. (eds.), *Corpus linguistics and Linguistic Theory*. Amsterdam: Rodopi, 141−155.

Labov, W. (1966), *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Lakoff, R. (1975), *Language and Woman's Place.* New York: Harper.

Leech, G./Rayson, P./Wilson, A. (2001), *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

McEnery, T./Xiao, Zh. (2004), Swearing in Modern British English: The Case of *Fuck* in the BNC. *Language and Literature* 13(3), 235−268.

Milroy, L./Gordon, M. (2003), *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.

Peters, P. (1998), Australian English. In: Bell, P./Bell, R. (eds.), *Americanisation and Australia*. Sydney: University of New South Wales Press, 32−44.

Pusch, C. D. (2002), A Survey of Spoken Language Corpora in Romance. In: Pusch, C. D./Raible, W. (eds.), *Romanistische Korpuslinguistik − Korpora und gesprochene Sprache. Romance Corpus Linguistics − Corpora and Spoken Language*. Tübingen: Gunter Narr Verlag, 245−264.

Rayson, P./Leech, G./Hodges, M. (1997), Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. In: *International Journal of Corpus Linguistics* 2(1), 133−152.

Romaine, S. (2000), *Language in Society. An Introduction to Sociolinguistics*. Oxford: Oxford University Press.

Ross, A. S. C. (1980), U and non-U. In: Mitford, N. (ed.), *Noblesse oblige.* London: Futura, 11−38.

Stenström, A.-B. (1991), Expletives in the London-Lund Corpus. In: Aijmer, K./Altenberg, B. (eds.), *English Corpus Linguistics in Honour of Jan Svartvik*. London: Longman, 230−253.

Stenström, A.-B./Andersen, G./Hasund, I. K. (2002), *Trends in Teenage Talk. Corpus Compilation, Analysis and Findings*. Amsterdam: John Benjamins.

Svartvik, J. (1966), *On Voice in the English Verb*. The Hague: Mouton.

Trudgill, P. (1974), *The Social Differentiation of English in Norwich.* Cambridge: Cambridge University Press.

Suzanne Romaine, Oxford (UK)