Jan Svartvik,
Department of English, Lund University, Sweden

Lexis in English language corpora

1. The second corpus generation

Many more years ago than I care to remember, on the occasion of my inaugural lecture at Lund University, I spoke with some enthusiasm about the bright future of corpus-based study of spoken language, what with tape-recorders getting smaller, and computers getting bigger. In 1992, at the Fifth Euralex Congress in Tampere, the future of corpus linguistics seems even brighter than on that previous occasion. Yet, while tape-recorders may indeed be a bit smaller (the stereo set, though, seems colossal compared to our gramophone), computers are actually getting smaller too: there has been a radical development from the mainframe to the micro, personal, desktop, laptop, palmtop and notebook. But not only are computers getting smaller but also faster and cheaper. This fantastic technological hardware development that we are witnessing is of course only one reason for my belief that the future of corpus linguistics is even brighter now than at the beginning of the seventies. The best part is that the hardware is also becoming well matched by software, and software development is indeed crucial if the corpus approach is going to fulfil its promise.

The meaning of "corpus" as given in most dictionaries is rather vague and gives little indication of bright prospects, for example:

- MACQUARIE DICTIONARY: "a body of data".
- COLLINS COBUILD DICTIONARY: "a large number of articles, books, magazines, etc that have been deliberately collected together for some purpose".
- LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH: "a collection .. of material or information for study" (New edition, 1987).
- LONGMAN DICTIONARY OF THE ENGLISH LANGUAGE (New edition, 1991) is more explicit: "a collection of spoken and/or written language for scientific study of word formation, sentence structure, sounds, etc".

COBUILD adds the warning: "a formal, technical word" (but, like LONGMAN, also gives the helpful hint that the plural can be either *corpora* or *corpuses*). All of the definitions in these recent works fail to specify "machine-readable", which is of course the current norm and also the topic of this paper, in particular electronic corpora of spoken English.[1] Only LONGMAN gives a clear indication that there are, and should be, corpora of speech – by far the most common use of language and the variety that has too long been neglected in both grammatical and lexicographical description.

It is not often that we can date the beginning of a new bud on the linguistic tree structure, but this is indeed possible with corpus linguistics, at least English corpus linguistics. It is now getting mature, just over 30 years of age. From the humble beginning engaging only a small number of linguists, corpora have become "the flavour of the

decade" (Sinclair 1992: 379). The beginning of this movement was the making of the Brown Corpus of written American English which set a pattern for the making of a host of corpora of representing other varieties of English (for descriptions of English language corpora, see Aijmer & Altenberg 1991: 315-318; Taylor, Leech & Fligelstone 1991). It was a typical feature of this first generation of corpora that they totalled one million words made up from 2000 or 5000 word-samples intended to be representative of some of the uses of the language, and were made available on computer tape for batch processing on mainframe machines located behind glass doors and operated by systems engineers in white coats.

We are now beginning to experience the second generation of corpora. They are characterized by larger size than those of the first generation: for example, the British National Corpus is planned to include 100 million words (Quirk 1992), and the corpus used by one group working on machine translation is reported to total 365,893,263 words (Brown et al 1991). Instead of the "representative", finite size corpus of the first generation we are likely to be seeing more typological variation, such as the "monitor" corpus where "sources of language text in electronic form would be fed on a daily basis across filters which retrieve evidence as necessary" (Sinclair 1991: 9). There is a movement in the direction of corpus pluralism: the index of the proceedings from a symposium on corpus linguistics, which took place in Stockholm a year ago, includes the following corpus types: core, dialect, expanded, grammatical, lexicographical, monitor, non-standard, regional, specialized, spoken, test and training corpus. Their days are by no means over, but "standard corpora" will probably serve more and more as stepping-stones to other, specific corpus types.

One obstacle to corpus use has been the lack of a standard encoding system, but this is now disappearing with the emergence of SGML (Standard Generalized Markup Language), which is likely to be in wide use. It is only to be hoped that SGML will also support a generalized system for prosodic transcription of spoken language (see Johansson 1991).

2. Why use a corpus in the first place?

Particularly in the last decade improved access to massive corpora, efficient machines and user-friendly programs has changed the working conditions of those linguists who use "real language data". Of course, not all linguists want to use corpora. In Chomsky's approach (1988: 45), "externalized language" (E-language) and "internalized language" (I- language) are separate entities, and it is I-language, ie the native speaker's mental competence, that is the primary subject of linguistics. This view is, however, not shared by linguists such as Chafe, Fillmore, Halliday and Leech (all 1992), who rather emphasize the interdependence of linguistic theory building and language data analysis. Yet, while many linguists value corpus data, the terms "corpus linguistics", and even more so "corpus linguist", are considered unfortunate by Wallace Chafe:

> "The term 'corpus linguist' puts the emphasis on one tie to reality that has been neglected by many contemporary linguists, I believe to the great detriment of the field: a tie that must be vigorously pursued if our understanding of language and the mind is to enjoy significant progress. But there is a complementary danger in implying that that is all a linguist should do, of pitting corpus linguists against introspective linguists or experimental linguists or computational linguists. I would like to see the day when we will all be more versatile in our

> methodologies, skilled at integrating all the techniques we will be able to discover for understanding this most basic, most fascinating, but also most elusive manifestation of the human mind" (Chafe 1992: 96).

Geoffrey Leech takes a more positive view and sees corpus linguistics as a new research paradigm:

> "computer corpus linguistics (CCL) defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject. The computer, as a uniquely powerful technological tool, has made this new kind of linguistics possible. So technology here (as for centuries in natural science) has taken a more important role than that of supporting and facilitating research: I see it as the essential means to a new kind of knowledge, and as an 'open sesame' to a new way of thinking about language" (Leech 1992: 106).

Whatever view we take of the advent of large computerized corpora, efficient and inexpensive machines and user-friendly software, it seems clear that they are here, not just to stay, but to transform the lives for most linguists interested in large collections of language in authentic use. This, I take it, will include the participants of the Euralex Congress. I suppose that we are mostly E-linguists here – possibly with the exception of those who have their minds on mental lexicons. Textual data have always been a basic tool for lexicographers who, with or without machines, have resorted to various strategies. Elisabeth Murray reports that, by the time the OED was completed in 1928, James Murray had over 4 million citation slips. Lacking a computer, he managed with manual labour: much of the work of alphabetizing and sorting the slips was done by Murray's many children (Murray 1977: 178-179).

I am speaking to this audience with some hesitation since I have to confess that I am no lexicographer. On the other hand, I have had a long – and, most of the time, friendly – association with corpus making and corpus use, chiefly for grammatical studies, and, like Michael Halliday, I believe in the interdependence of lexis and grammar:

> "grammar and vocabulary are not two different things; they are the same thing seen by different observers. There is only one phenomenon here, not two. But it is spread along a continuum. At one end are small, closed, often binary systems, of very general application, intersecting with each other but each having, in principle, its own distinct realization. ... At the other end are much more specific, loose, more shifting sets of features, realized not discretely but in bundles called 'words', like *bench* realizing 'for sitting on', 'backless', 'for more than one', 'hard surface'; the system networks formed by these features are local and transitory rather than being global and persistent" (1992: 63).

With the insights drawn from extensive corpus investigations there might indeed be "little or no need for a separate residual grammar or lexicon" (Sinclair 1991: 137). Words like *get, of, any* belong to a common ground of grammar and lexicon where corpora will be particularly helpful. Returning to the pre-computer generation of lexicographers, we find that Murray complains about the lack of data on these "little words": "no more important help", he says in 1882, "could now be rendered to the Dictionary than the collection of modern instances of all uses and constructions of these little words" (*TPS 1882-4, 7*).

I think corpora are likely to make a major impact in a number of linguistic research areas. They may well open up new research paradigms and originate new linguistic models, and will certainly offer a descriptive foundation of a kind that we have not had before, including the study of register and dialect variation and probability of textual

occurrence. Future descriptive grammars and dictionaries are hardly likely to be produced without recourse to authentic examples.

Furthermore, corpus work will no doubt make its mark in many other areas like historical and applied linguistics. The CD-ROM versions of such historical depositories as the *OED* and the Helsinki Corpus of English Texts (see Kytö 1991) are likely to open up new possibilities in the field of diachronic studies (as examples of what a historical corpus can offer, see work by Matti Rissanen and his group at Helsinki, such as Nevalainen 1991 and Raumolin-Brunberg 1991). The now easily retrievable historical data can shed new light on historical developments such as the influx of Romance lexical material and the influence of French on English but also on theoretical issues, for example the relation of grammar and lexis, as stated in a recent study of suffixal derivation in Middle English:

> "Interesting though they were, the results of the morphological analysis, were not always significant. In the end it became fairly clear that it was semantics which was the more powerful driving force behind the shifts and reshuffles in the Middle English derivational system. Potentially, this is a finding which could feed back into our understanding and theoretical conception of word-formation and its position in a model of grammar as it seems to me to underline the role of the lexicon" (Dalton-Puffer 1991:327).

In language teaching, assuming that both teaching methods and exposure to authentic language are important for language learning, there is naturally much to be learned from "real data", as opposed to the "concocted examples" often used in linguistic studies or the "pedagogical language" as commonly encountered in language learning textbooks. We all have some experience of students coming to university with a naive attitude to usage as being either correct or incorrect. For such students, a hands-on, self-access experience of real data in the classroom could provide a valuable eye-opener to the wider linguistic issues of frequency, acceptability, collocability and style in current usage (see Tribble & Jones 1990).

3. Corpora of spoken English

All handbooks in linguistics have long stressed the importance of the spoken language, and for some time now we have witnessed novel approaches to the study of spoken discourse. Our contribution at Lund University to this field was the launching of the Survey of Spoken English in the mid-seventies. Our first undertaking was to obtain suitable data. Having been an associate of the research team on Randolph Quirk's Survey of English Usage at University College London in the sixties, it was a natural step to make use of this corpus by computerizing the spoken component of the carefully transcribed material, then stored only on paper slips in Foster Court filing cabinets. Given the technology available to us at the time, computerization of such complicated data with its detailed prosodic transcription was by no means a simple task, but the operation was nevertheless considered essential for three main reasons. We wanted, first, to have easy access to the material at our Lund base; second, to make use of the computer's superb possibilities as a tool for retrieval, storage, classification, etc.; third, to be able to share the database with fellow researchers no matter where they happened to be working. The original version of the London-Lund Corpus of Spoken English, which was distributed on computer tape and included 87 texts, became available in 1980, when we also publish-

ed a printed book including conversations in the corpus (Svartvik & Quirk 1980). The complete version, including all 100 texts (see the description in Greenbaum & Svartvik 1990), totalling half a million words, recently appeared in a CD-ROM version together with other English language corpora, and all with retrieval tools (WordCruncher and TACT) included.[2] The majority of the texts in the London-Lund Corpus are conversations. One reason for this is that informal, spontaneous, interactive discourse is by far the most common form of language use, another that it has been an underresearched area of modern English; this was conspicuously so in the late fifties when the plans were drawn up for the London Survey (see Quirk 1960).

The chief aim of the Survey of English Usage was to create a basis for studying English grammar rather than its lexis. For general lexical work, such as dictionary-making, a corpus of one million words, half of them written, half spoken, is clearly inadequate. For comparison, Cobuild, which is a project dedicated to lexical computing, has a text corpus of general English which "stands at around 20 million words in daily use, backed up by a range of more specialised texts coming to a total of about another 20 million" (Sinclair 1987: vii). Yet, while the London-Lund Corpus has been used chiefly for studies of grammar and discourse (see Greenbaum & Svartvik 1990, Appendix 2), it can indeed be used also for lexical studies, particularly if we take the view that grammar and lexis form a continuum and focus on Murray's "little words".

I will now briefly survey some areas where lexical work has been done on corpus-based spoken English: statistical vocabulary studies, adverbials and prosody, discourse items, register variation, semantic fields, and collocation. Most of these areas fields also hold great promise for future research.

4. Statistical vocabulary studies

The aim of the first uses of corpora, including those B.C. (before computers), was chiefly lexico-statistical. The studies on English by Thorndike (1921), Fries & Traver (1940), Thorndike & Lorge 1944, and Bongers (1947) were closely connected with language teaching and the "vocabulary control movement". In his work on vocabulary Palmer included six thousand collocations which led him to suggest that even common collocations "exceed by far the popular estimate of the number of simple words contained in our everyday vocabulary", thus "throwing a new light on the nature of vocabulary" (1933: 7; for a useful survey of this field, see Kennedy 1992).

So far the most extensive dedicated pedagogical use of corpora has been to produce statistics on frequency of vocabulary items and structural patterns. One form of information derived from word frequency counts is that, in most texts, a small number of different words (ie types) account for a very large proportion of all word tokens: in most written texts 5,000 words will account for up to 95% of the tokens, and 1,000 words will account for 85%; in speech, 50 function words account for up to 60% of the tokens (cf Kennedy 1992: 339; for LOB analyses, see Johansson & Hofland 1989).

Recent approaches, such as the lexical syllabus (Sinclair & Renouf 1987), highlight the common uses of common words, stressing the importance of the good company of words rather than the large number of words. Hence the foremost task for language learners is not to learn as many words as possible but the highly frequent words in their customary environment (cf Sinclair 1987: 159):

> At present many learners avoid the common words as much as possible, and especially the idiomatic phrases. Instead they rely on larger, rarer and clumsier words which make their language sound stilted and awkward.

Within the current Lund project "Public Speaking", we are making a study of new set of material, the Spoken English Corpus (SEC), compiled at the University of Lancaster in conjunction with the Speech Research Group at the IBM UK Scientific Centre (see Taylor & Knowles 1988; Knowles 1990; Wichmann 1991).[3] SEC includes radio news broadcasts and radio commentaries, public lectures, religious programmes, recitations, etc. Unlike LLC with its focus on spontaneous interactive speech, SEC consists of planned monologue, but both are prosodically analysed.

I include two tables from research in progress (both from Ekedahl 1992). Table 1 is a rank list of the fifty most common words in the SEC with the corresponding ranks of the same words in LOB, Brown and LLC.[4] The result of a calculation of the rank differences between the corpora is shown in Table 2.[5]

As a brief characteristic we can say that, in terms of the most frequent lexical items (graphic words), spontaneous speech (LLC) is strikingly different from all the other three text types; the written English texts from the two major varieties, British (LOB) and American (Brown), are remarkably similar; planned monologue (SEC) is more like writing than spontaneous speech.

5. Adverbials and prosody

There are, Altenberg states, "two areas where I think contemporary dictionaries fail to give an adequate representation of speech: the use of intonation to differentiate adverbial functions and the treatment of certain speech specific discourse-items" (1990: 177-178). Adverbs like *frankly, literally, personally, clearly, naturally, superficially, ironically, happily* can have two grammatical functions: as manner adjunct, for example

> He asked me to tell him frankly what I wished to do.

and as a sentence adverbial (conjunct or disjunct), as in

> Frankly, this has come as a bit of a shock.

Altenberg finds that both COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY and LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH fail to make the full tie-in with grammar by stating the different regular positions in the sentence and, above all, fail to provide important prosodic information: "although adjuncts and disjuncts may occur in the same syntactic position, they are always prosodically distinct" (Altenberg 1990: 181). Compare these examples (with nuclei in bold) from Allerton & Cruttenden (1976: 48):

(1) Richard played **nàturally▪** (adjunct)

(2) Richard **plàyed▪ nàturally▪** (disjunct)

In addition to tone unit separation as in (2) and positional mobility, as in (2a):

(2a) Naturally Richard played

disjuncts often carry a falling-rising tone, instead of a falling tone.

Table 1. The 50 most frequent words in SEC and comparisons with LOB, Brown, and LLC

Word	SEC	LOB	Brown	LLC
the	1	1	1	1
of	2	2	2	5
and	3	3	3	3
to	4	4	4	4
a	5	5	5	6
in	6	6	6	9
that	7	7	7	8
was	8	9	9	13
for	9	11	11	20
it	10	10	12	10
he	11	12	10	18
is	12	8	8	11
on	13	16	16	16
as	14	13	14	29
at	15	19	18	26
his	16	18	15	85
with	17	14	13	32
I	18	17	20	2
but	19	24	25	15
by	20	20	19	65
's	21	–	–	–
this	22	22	21	14
be	23	15	17	21
one	24	38	32	36
you	25	32	33	7
from	26	25	26	53
they	27	33	30	24
have	28	26	28	19
we	29	40	41	23
an	30	34	29	81
are	31	27	24	42
were	32	35	34	64
all	33	39	36	33
not	34	23	23	35
which	35	28	31	43
there	36	36	38	38
had	37	21	22	55
their	38	41	40	–
been	39	37	43	68
n't	40	–	–	–
so	41	46	52	30
two	42	63	69	81
has	43	42	44	94
said	44	48	53	76
who	45	50	46	83
or	46	31	27	44
when	47	44	45	67
can	48	57	61	70
up	49	52	55	61
will	50	48	47	–

Table 2. Sums of rank differences for the 50 most common words in SEC.

SEC vs. LOB	200
SEC vs. Brown	211
SEC vs. LLC	675
LOB vs. Brown	93
LOB vs. LLC	706

Adverbials occupy an intermediate position on the grammar/lexis continuum: they have specific grammatical functions but form a large, open lexical class with a wide range of meanings. Clearly they must be properly covered in the dictionary. Grammatical tagging of entries in dictionaries is now fairly commonplace, at least in learners' dictionaries, but it is of course doubtful whether this type of information is properly used. My own experience is that it is not. There are several likely reasons for this: one is that so far there have been only weak or non-existent attempts on the part of lexicographers to establish a solid link between grammar, lexis and prosody; another that there is no universally accepted system of grammatical and prosodic categories; most importantly, once we leave the reasonably obvious lexical definition of the word and enter the nebulous realms of grammar and prosody, the level of linguistic abstraction makes definitions more complicated. The understanding of, and motivation to learn, terms like "disjunct", "falling-rising tone" and even "transitive" are bound to be limited among general dictionary users who are accustomed to look up words in a dictionary mainly to check spelling or meaning. Yet as dictionaries have become more and more specialized and geared to the needs of different user-categories, those users who are familiar with grammatical and prosodic terminology are likely to benefit from more complete information than is offered in general-purpose dictionaries. Although it carries meaning, prosody has been almost totally neglected in dictionaries.

6. Discourse items

In the word-class tagging of part of the spoken corpus that we undertook at Lund, it became clear that the set of traditional word-classes was inadequate. Hence we devised a new tagset consisting of over 200 categories. This is large in comparison with other similar sets: the tagged Brown Corpus uses 179 different wordtags, the LOB tagset comprises 132 tags, and the Leeds tagset 137 tags (for a description of the tagset, see Svartvik 1990: 94; for the implementation of probabilistic word-class tagging on LLC and the design of a model for morphological knowledge representation, see Eeg-Olofsson 1991). The types of problems we faced can be exemplified by *mm, you know* and *sort of thing*.

'Responses' transcribed as *m, mm* or *mhm* are usually not to be found in dictionaries; COBUILD seems to be an exception here:

> "*Mm* is used in writing to represent a sound that you make when someone is talking, to indicate that you are listening to them, that you agree with them, or that you are preparing to say something" (928).

The frequency list indicated that the verbs *know, think, mean, see* were extremely frequent in spoken as compared with written English. The reason is of course that a word-based frequency list fails to capture word combinations like *you know, you see* and *I mean* functioning as 'softeners', 'responses' such as *I see, that's right*, and 'hedges' such as *sort of thing*, which tend to find a place neither in dictionaries nor grammars. Yet in a sample of 50,000 words such 'discourse items' occupy fourth place, ahead of the well-established grammatical word-classes of prepositions, adverbs, conjunctions and adjectives.

'Discourse items' which are almost exclusively restricted to spoken discourse have been divided into groups (cf Nattinger 1988: 78-79; Stenström 1990: 144; Stenström forthcom-

ing) such as social interactions, necessary topics and discourse devices, including, for example:

greetings: *how are you doing*
closings: *be seeing you*
politeness routines: *if you don't mind*
refusing: *no way*
time: *how long ...*
space: *how far ...*
fluency devices: *you know*
sensory predicates: *it seems to me ...*
reinforcers: *OK, and then what happened*
hedges: *sort of thing*
responses: *fine, quite, right, sure thing, fair enough, uhuh*

One customer in spoken English is particularly slippery: it is very hard to adequately describe – let alone teach – *well*, as in these examples from the London-Lund Corpus:

and I I **sáid**▪ well I I don't **rèally** think▪ I could I **wrîte**▪ (S.1.3.6)

B: I I think they've got quite a good **opînion** of him▪
A: I well (m) :I :I have **tôo**▪ (S.1.3.38)

This innocent-looking four-letter word has rank 14 in our corpus of conversations, ie it is more common than central grammatical items like *this, we, on, for, if, do, which*. While *well* as a discourse device (as opposed to a manner adverb) is to be found in the Top 20 list in speech it is non-existent in writing and strikingly absent in most pedagogical handbooks. Clearly, an item with this kind of frequency in the conversation of native speakers has got to be important also to foreign students who want to manage conversations adequately.

7. Register variation

Probably the most comprehensive corpus-based study of linguistic variation in spoken and written English has been conducted by Douglas Biber. His multi-dimensional, statistical comparison of linguistic characteristics of 23 genres does not lead him to make an absolute, two-way distinction between spoken and written discourse: "... the variation among texts within speech and writing is often as great as the variation across the two modes" (1988: 24). Yet, face-to-face conversation is described as the prototypically oral genre and three dimensions in particular distinguish oral and literary discourse (162):

Informational versus Involved Production
Explicit versus Situation-Dependent Reference
Abstract versus Non-Abstract Information

Without questioning Biber's conclusions in this valuable study it seems clear that, to the participant – in particular the foreign language learner – the gap between the two modes of writing/reading, on the one hand, and speaking/listening, on the other, is actually wider than appears from his statement. The reason is that the linguist examines the end-product of a process, as evidenced in a corpus, while the learner is the actual perfor-

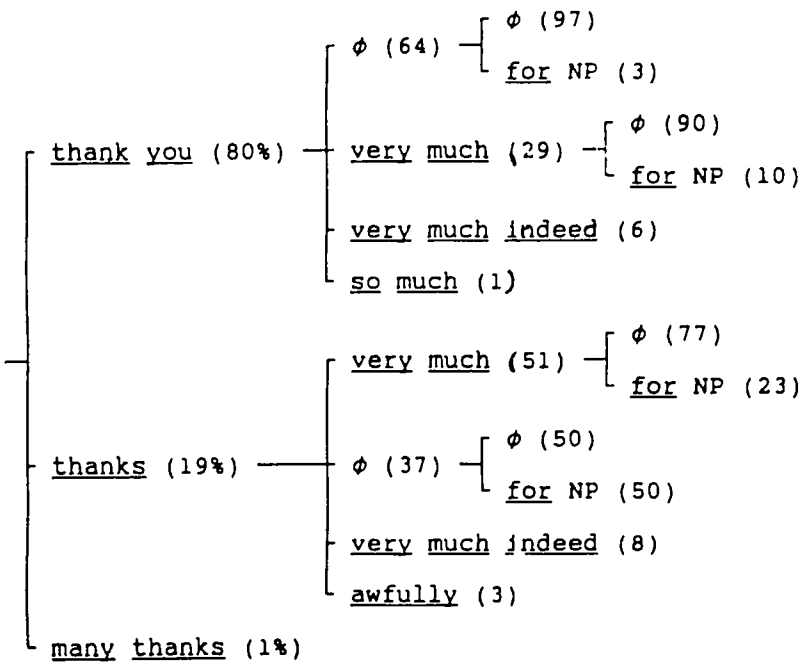
Table 3

you know	152
[m] [m]	128
yes yes	120
I think	106
sort of	100
you see	95
oh yes	94
isn't it	88
and then	82
which is	81
I mean	74
and he	73
and they	72
thank you	72
at all	65

Table 4

at the moment	203
for a moment	16
at this moment	12
in a moment	11
one moment	8
for the moment	6
just a moment	5
wait a moment	4
for one moment	4
a few moments	4
that moment	3
a moment ago	2
a moment please	2
any moment	2
at any given moment	2
dreadful moment	2
from the moment	2
of the moment	2
this moment	2
at this very moment	2
within a matter of moments	2

Table 5



mer/producer of the process, and the speech process is radically different from the writing process, in particular with its real-time constraint.

8. Semantic fields

What appears to be a most fruitful lexical use of corpora is the analysis of specific semantic fields and pragmatic categories. In his study of the expression of modality, Hermerén (1986) found, among other things, that verbs are used much more frequently than other word classes to express Obligation, Permission, Volition and their negated equivalents, yet "modal auxiliaries express these modalities less often than the exponents of other word classes put together", and modal nouns are generally more frequent in written than spoken English (90).

Similarly, in her study of epistemic modality as expressed in some ESL textbooks as compared with real corpus-data, Janet Holmes has shown that many textbook writers "devote an unjustifiably large amount of attention to modal verbs, neglecting alternative linguistic strategies for expressing doubt and certainty" (1988: 40). Such alternatives include lexical verbs (*appear, believe, doubt, seem, suggest,* etc), adverbials (*apparently, certainly, doubtless, inevitably, necessarily,* etc) and nouns (*belief, certainty, idea, opinion, possibility, tendency,* etc). The reason for the traditional emphasis on modal verbs to the exclusion of lexical verbs, adverbials and nouns can be traced to structural grammars where the morphological peculiarities of modal auxiliaries (lack of third-person-s, infinitive, and participle forms, etc) naturally place these auxiliaries high on the list of teaching items. Other semantically equivalent expressions (*suggest, apparently, belief,* etc) do not constitute any morphological problem and, consequently, have no place in a morphologically-biassed textbook.

Kennedy has studied the uses of certain lexical items such as *between* and *through.* While they are among the most frequent words in the English language there is neither descriptive nor pedagogical guidance about them. In addition to offering a statistical dimension to this area, Kennedy provides information about their occurrence: "like other structural words, [they] are learnt not as representatives of word classes or lexemes in isolation, but in association with other words" (1991: 110).

9. Collocation

Large collections of real data offer a rich, but as yet largely uncultivated, field for studying habitual cooccurrences of lexical items, whether they be called lexical phrases, collocations, prefabs or preassembled chunks. Some such multi-word items belong to the speech-specific categories already mentioned (*if you don't mind,* etc), but most types do not appear to be characteristic of either the spoken or written varieties. Yet there is a reason why such prefabs may be considered particularly relevant for the student of spoken discourse. Interactive speech takes place in real time which – unlike written discourse – offers no opportunity of resorting for help to a dictionary, a friend or an embassy. In the typical information structure of speech we speak in brief chunks (ie information units, tone units) which are often made up of habitual cooccurrences.

The study of recurrent lexical patterns in spontaneous speech is important for language teaching and speech recognition besides lexicography. Bengt Altenberg, my Lund colleague, has a large database containing some 200,000 recurrent examples (tokens) representing 68,000 different types of word combinations. Table 3 shows the most frequent two-word combinations, Table 4 shows the collocational tendencies of the word *moment*, and Table 5 shows the variant expressions of thanks (from Altenberg 1991).

With access to large corpora, spoken and written, we can now begin the serious study of collocation. The mastery of collocation is of course a real stumbling-block to the foreign learner:

> "The mental lexicon of any native speaker contains single-word units as well as phrasal units or collocations. Mastery of both types is an essential part of the linguistic equipment of the speaker or writer and enables him to move swiftly and with little effort through his exposition from one prefabricated structure to the next" (Kjellmer 1991: 125).

There is no dictionary I know of that clarifies the restrictions of *good, strong* and *high* in such collocations as the following (Bolinger 1975: 103-104):

good likelihood	*strong likelihood*	**high likelihood*
**good probability*	*strong probability*	*high probability*
good possibility	*strong possibility*	**high possibility*
good chance	**strong chance*	**high chance*

10. The electronic lexicon

Over the last two decades we have witnessed a rapid increase in the computerization of dictionaries, going from computerized type-setting via computerized lexical databases to fully electronic lexicons available on CD-ROM.

Electronic word tools can be very useful in the writing process. This is particularly true for an international language like English, with more non-native than native users. I would think that, today, it is impossible to sell a word-processing package that does not include a spelling-checker with a spelling-corrector. As yet, grammar-checkers, and certainly grammar-correctors are unsophisticated, and some barely tolerable (why does the passive voice seem to be hated by all of them?), but they will be making progress, particularly if there is better cooperation between software engineers and linguists (see Kucera 1992). Similarly, there are interesting developments in style and readability programs such as Corporate Voice (see Bohm 1992).

One of the great linguistic challenges of the nineties is of course machine translation. So far there has been surprisingly little use made of corpora in this field, but there is now a growing awareness that the analysis of large collections of real data are required for solving many of the problems at hand (cf Allén 1992: 1). After a bumpy ride over the last forty years, machine translation has now turned right, into a smoother road. However, what seems to be badly wanted – in addition to realistic goals and linguistic insights – to make the journey successful is sophisticated and comprehensive bilingual and multilingual electronic dictionaries. Research on parsing has been too much concerned with syntactic rules and too little aware of the importance of contrastive lexical, grammatical, pragmatic and stylistic knowledge which can best be derived from authentic language use as found in large and diverse corpora carefully analysed by linguists.

Notes

1 I want to thank Bengt Altenberg and Anne Wichmann for comments on a draft of this paper.

2 The title of the CD-ROM (ISBN 82-7283-064-7, December 1991) is "ICAME Collection of English Language Corpora". It includes the Brown, Helsinki, Kolhapur, LOB, and London-Lund corpora and is distributed by Norwegian Computing Centre for the Humanities, Bergen, Norway, P.O. Box 53, N-5027 Bergen, Norway.

3 The project "Public Speaking" is funded by the Swedish Council for Research in the Humanities and Social Sciences (HSFR).

4 From LLC only a list of 100 was available, hence the two missing words, *their* and *will*. The contractions *'s* and *n't* are defined as words only in SEC. "*Not* would have a rank of 15 in SEC if all the negations were counted together. The *'s* total comprises contractions of both *is* and *has*. If we add up all occurrences of *is*, we get the total of 619, which would have a rank of 7. Contracted forms have been counted as distinct words in the other corpora" (Ekedahl 1992).

5 The Ekedahl (1992) formula used was $\Sigma \mid R_{1i}-R_{2i} \mid$, where R_{1i} is the rank of the word number i in the first list, and R_{2i} is the rank of the same word in the second list; i is the number of the word in the SEC list and varies between 1 and 50. The two '|' mean that the value between them is always to be turned into a positive number.

References

Aijmer, Karin & Bengt Altenberg (eds.). 1991. *English corpus linguistics*. London: Longman.

Allén, Sture. 1992. "Opening address". In Svartvik (ed.), 1-3.

Allerton, D.J. & A. Cruttenden. 1976. "The intonation of medial and final sentence adverbials in British English". *Archivum Linguisticum* 7: 29-59.

Altenberg, Bengt. 1990. "Spoken English and the dictionary". In Svartvik (ed.), 177-191.

Altenberg, Bengt. 1991. "The London-Lund Corpus of Spoken English: Research and applications". Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text, 71-83. University of Waterloo, Waterloo, Ontario, Canada.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bohm, Cecilia. 1992. Readability analysis by computer. An evaluation of the readability programme Corporate Voice. [Research paper.] Department of English, Lund University.

Bolinger, Dwight. 1975. *Aspects of language*. New York: Harcourt Brace.

Bongers, H. 1947. *The history and principles of vocabulary control*. Woerden: Wocopi.

Brown, Peter F., Vincent J. Della Pietra, Peter V. de Souza, Jenifer C. Lai & Robert L. Mercer. 1991. "Class-based n-gram models of natural language. [Paper for the Pisa conference on European corpus resources, 24-26 January 1992.]

Chafe, Wallace. 1992. "The importance of corpus linguistics to understanding the nature of language". In Svartvik (ed.), 79-97.

Chomsky, Noam. 1988. *Generative grammar: Its basis, development and prospects*. Kyoto: Kyoto University of Foreign Studies.

Collins Cobuild English language dictionary. 1987. London: Collins.

Dalton-Puffer, Christiane. 1991. Suffixal derivation in Middle English. A corpus-based study. Ph.D. dissertation, Department of English, University of Vienna.

Eeg-Olofsson, Mats. 1991. *Word-class tagging. Some computational tools*. [Ph.D. Diss] Department of Computational Linguistics, University of Göteborg.

Ekedahl, Olof. 1992. Word and tag frequencies in SEC. [Research paper.] Department of English, Lund University.

Fillmore, Charles J. 1992. "'Corpus linguistics' or 'Computer-aided armchair linguistics'". In Svartvik (ed.), 35-60.

Fries, Charles C. & A. Aileen Traver. 1940. *English word lists. A study of their adaptability for instruction.* Washington: American Council on Education.

Greenbaum, Sidney & Jan Svartvik. 1990. "The London-Lund Corpus of Spoken English". In Svartvik (ed.) 1990, 11-45.

Halliday, M.A.K. 1992. "Language as system and language as instance: The corpus as a theoretical construct". In Svartvik (ed.), 61-78.

Hermerén, Lars. 1986. "Modalities in spoken and written English. An inventory of forms". In *English in speech and writing: A symposium* edited by Gunnel Tottie & Ingegerd Bäcklund, 57-91. Studia Anglistica Upsaliensia 60. Stockholm: Almqvist & Wiksell.

Holmes, Janet. 1988. "Doubt and certainty in ESL textbooks". *Applied Linguistics* 9: 21-44.

Johansson, Stig. 1991. "Some thoughts on the encoding of spoken texts in machine-readable form" [MS].

Johansson, S. & K. Hofland. 1989. *Frequency analysis of English vocabulary and grammar.* Oxford: Oxford University Press.

Kennedy, Graeme. 1991. *"Between and through:* The company they keep and the functions they serve". In Aijmer & Altenberg, 95-110.

Kennedy, Graeme. 1992. "Preferred ways of putting things with implications for language teaching". In Svartvik (ed.), 335-373.

Kjellmer, Göran. 1991. "A mint of phrases". In Aijmer & Altenberg (eds.), 111-127.

Knowles, Gerry. 1990. "The use of spoken and written corpora in the teaching of language and linguistics." *Literary and Linguistic Computing* 5: 45-48.

Kucera, Henry. 1992. "The odd couple: The linguist and the software engineer. The struggle for high quality computerized language aids". In Svartvik (ed.), 401-420.

Kytö, Merja. 1991. *Manual to the diachronic part of the Helsinki corpus of English texts. Coding conventions and lists of source texts.* Department of English, University of Helsinki.

Leech, Geoffrey. 1992. "Corpora and theories of linguistic performance". In Svartvik (ed.), 105-122.

Longman Dictionary of Contemporary English. 1987. New edition. London: Longman.

Longman Dictionary of the English Language. 1991. New edition. London: Longman.

Macquarie dictionary. 1991. Second edition. Macquarie University.

Murray, K.M. Elisabeth. 1977. *Caught in the web of words: James Murray and the Oxford English Dictionary.* New Haven and London: Yale University Press.

Nattinger, J. 1988. "Some current trends in vocabulary teaching". *Vocabulary and language teaching,* edited by Ronald Carter & M. McCarthy, 62-82. London: Longman.

Nevalainen, Terttu. 1991. *"But, only, just". Focusing adverbial change in Modern English 1500-1900.* Helsinki: Société Néophilologique.

Palmer, Harold E. 1933. *Second interim report on English collocations.* Tokyo: Institute for Research in English Teaching.

Quirk, Randolph. 1960. "Towards a description of English usage". *Transactions of the Philological Society 1960:* 40-61.

Quirk, Randolph. 1992. "On corpus principles and design". In Svartvik (ed.), 457-469.

Raumolin-Brunberg, Helena. 1991. *The noun phrase in early sixteenth- century English. A study based on Sir Thomas More's writings.* Helsinki: Société Néophilologique.

Sinclair, John. 1987. "The nature of the evidence". In *Looking up. An account of the COBUILD project in lexical computing,* edited by John Sinclair, 150-166. London: Collins.

Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, John. 1992. "The automatic analysis of corpora", in Svartvik (ed.), 379-397.

Sinclair, John & Antoinette J. Renouf. 1987. "A lexical syllabus for language learning". In *Vocabulary and language teaching*, edited by Ronald Carter & M. McCarthy, 140-160. London: Longman.

Stenström, Anna-Brita. 1990. "Lexical items peculiar to spoken discourse". Svartvik (ed.), 137-175.

Stenström, Anna-Brita. Forthcoming. An introduction to spoken interaction.

Svartvik, Jan (ed.). 1990. *The London-Lund Corpus of Spoken English: Description and research*. Lund: Lund University Press.

Svartvik, Jan. 1990. "Tagging and parsing on the TESS project". In Svartvik (ed.), 87-106.

Svartvik, Jan (ed.). 1992. *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991*. Berlin: Mouton de Gruyter.

Svartvik, Jan & Randolph Quirk. 1980. *A corpus of English conversation*. Lund Studies in English 56. Lund: Lund University Press.

Taylor, Lita J. & Gerry Knowles. 1988. *Manual of information to accompany the SEC corpus. The machine-readable corpus of spoken English*. Unit for Computer Research on the English Language, University of Lancaster.

Taylor, Lita J., Geoffrey Leech & Steven Fligelstone. 1991. "A survey of English machine-readable corpora". In *English computer corpora. Selected papers and research guide*, edited by Stig Johansson & Anna-Brita Stenström, 319-353. Berlin: Mouton de Gruyter.

Thorndike, Edward L. 1921. *Teacher's word book*. New York: Columbia Teachers College.

Thorndike, Edward L. & Irving Lorge. 1944. *A teacher's word book of 30,000 words*. New York: Columbia Teachers College.

Tribble, Chris & Glyn Jones. 1990. *Concordances in the classroom*. A resource book for teachers. London: Longman.

Wichmann, Anne. 1991. Beginnings, middles and ends. A study of initiality and finality in the Spoken English Corpus. Ph.D. thesis. Department of Linguistics and Modern English, University of Lancaster.