A method for finding and assessing differences in lexical clusters and cluster frequencies between spoken British and American English

Brigitta MITTMANN, Augsburg, Germany

Abstract

This paper is a report on research into phraseological differences between an American and a British corpus of spoken English. Recurrent combinations of two, three, or four words which occur significantly more often in one corpus than in the other are extracted and used as a basis for further and more detailed investigation. The results mainly show differences in conversational routines and prove once again that in many, if not most, cases no simple 1:1 equivalence between the two varieties can be established. While the results of this kind of research do not provide normative information for encoding written English, it is nevertheless argued that information about phraseological differences between spoken varieties of English is vital for teachers and students alike and should be retrievable from dictionaries.

1 Word combinations as a clue to a speaker's origin

When we hear a British or an American person speak we can usually tell which of the two countries he or she comes from. Most frequently, their pronunciation will be the give-away. There are also certain words or grammatical constructions which give hints as to whether somebody is likely to stem from Birmingham, Alabama or from Birmingham in the English West Midlands. Apart from this, however, there are – as I would like to argue – considerable differences between British and American speakers with respect to the word combinations which they typically use very frequently. This has important implications for applied linguistics, for fields as diverse as translation, teaching, and lexicography.

2 Previous research on phraseological or collocational differences between varieties of English

Standard general textbooks on varieties of English hardly offer any information on phraseological or collocational differences between American and British English. More usually, they contain lists of single lexical items which are typically more or less restricted to one of the two varieties. These lists often ignore further information on social or regional variation which would make the distinction less clear. They can seem rather impressionistic and subjective and typically also lack the corpus research backing that one would wish for nowadays.

There are also few specialised articles on phraseological or collocational differences between British and American English - notably the research on collocations in [Greenbaum 1988] and [Gläser 1991, 1992]. Some of their results have been called into question by insights gained through the use of corpora over the last few years[1].

Interestingly, monolingual learners' dictionaries still seem to be the best source for information about collocational and phraseological differences between British and American English. Another important source is the *BBI Combinatory Dictionary of English*. However, these dictionaries do not always agree on what should be labelled as *American English*/*especially American English* or *British English*/*especially British English*.

Moreover, there is as yet no systematic research (of the kind which a dictionary cannot present) into phraseological and collocational differences between British and American English. Neither is it clear - or easy to find out and quantify – how large these differences are.

3 A method for comparing corpora of spoken American and British English with respect to differences in word combinations

3.1 The corpora

The present article is in part an interim report on the research done for a dissertation about collocations and other phraseological units in spoken British and American English. This study is based upon two corpora of spoken English which both aim to be representative of natural, spontaneous, non-scripted spoken English: for British English, the 'spoken demographic' part of the *British National Corpus* (*BNCSD*); for American English: the *Longman Spoken American Corpus* (*LSAC*). Both corpora were transcribed orthographically, with a number of annotations relating to pauses, paralinguistic phenomena, etc. In both the *BNCSD* and the *LSAC* all age groups, all major social strata and different regions of the two countries are represented. The two corpora are similar in size – the *LSAC* having about 4.9 million and the *BNCSD* about 3.9 million words of running text.

The *BNCSD* and the *LSAC* are among the largest available corpora of conversational British and American English that are similar enough to be compared in this way. Nevertheless, they are rather small for serving as a basis for lexical research, and especially so when it comes to lexical combinatorics. This means that only the most frequent items can be considered for closer inspection.

There are also certain differences between the corpora which have to be taken into account when one tries to compare them. Firstly, despite their similarity in size the *BNCSD* is somewhat smaller than the *LSAC* which makes it necessary to normalize the figures for comparison. Therefore, all frequency counts will always be given both as raw figures and as *parts per million* (*ppm*). Other differences that have to be taken into consideration relate to the methods of transcription and mark-up that were used. While the *BNCSD* has *SGML/TEI*-conforming mark-up with initial and final tags bracketing stretches of the text, the *LSAC* has a completely different system of annotation which uses only initial tags in most cases. This means that corpus analysis programs have to be provided in two versions and that all results have to be double-checked for correctness. The fact that the *BNCSD* was transcribed using British English spelling, and the *LSAC* using American orthography, results in spelling variants. Some of these variants might be expected (e.g. *centre-center*) while others come unexpected (*a while* versus *awhile*; *'cause*

versus *cos*). With interjections, it is sometimes difficult, if not impossible, to decide whether one is dealing with mere spelling differences or actual differences of meaning and use. Some of the most frequent items in the two corpora – *er*, *erm*, *mm* (*BNCSD*) and *uh*, *um* (*LSAC*) – seem to have partly overlapping conditions of use, but there are also significant differences.

3.2 Finding clusters in corpora

A series of specially written programs[2] was used to extract the most frequent word combinations (hereafter called *clusters*) from each of the two corpora. These tools make it possible to filter all combinations of two, three, four, or more, words from a text and count their frequency of occurrence. The results are presented as lists. For example, the programs would output the following three-word-clusters for the sentence *Can I have a look at this?*: *CAN I HAVE – I HAVE A – HAVE A LOOK – A LOOK AT – LOOK AT THIS* (each with a raw frequency of one).

The programs were instructed to respect punctuation marks and speaker turn changes. This excludes many useless high-frequency clusters like *you don't you, isn't it yes, mm do you* which stretch over clauses or even turns of different speakers. Nevertheless, even with these restrictions, the output of the programs contains a number of clusters like *I have a* which are of little or no interest in terms of phraseology. On the other hand, many interesting phenomena can be found using this method, especially if long texts are being processed.

3.3 Notes on the size of the 'phrasicon'

One of the most interesting questions of phraseological research is just how much of language is prefabricated. There is reason to suppose that figures of recurrent clusters can give some hints towards the solution of this problem, bearing in mind that there is an underlying problem of drawing the line between prefabricated units and strings of words generated by syntactic rules.

In both corpora, about 35% of all two-word cluster types occur at least twice. For clusters involving more words, the rate is lower: around 20% for three-word clusters and around 10% for four-word clusters. It is obvious, however, that simple recurrence may be nothing but a product of chance. It is useful to compare this with the percentages of clusters occurring at least ten times: they are around 7% for two-word clusters, 2% for three-word clusters and only 0.5% for four-word clusters.

However, there are a number of difficulties connected with this approach. Firstly, as [Altenberg 1998] has pointed out, frequency of recurrence cannot be used as a criterion for phraseological status. Secondly, with this method, cluster frequencies (and with them, the size of the 'phrasicon') depend on the length of the strings of words that are analysed. The longer a sentence, clause, or other chunk of language delimited by punctuation marks, the more clusters it contains. Thus, in a way, cluster frequencies depend on the transcription – and as one can easily imagine, punctuation is a general problem in the transcription of any spoken text.

A third problem is that the cluster method blatantly ignores a large number of phraseological or semi-phraseological units. Amongst the items that fall through this net are, for example, many collocations of the types studied by [Hausmann 1984, 1989], [Cowie 1994], and [Howarth 1996] where as many as six or more words can intervene between the words that collocate. Other items

left out are so-called *formal idioms* [Moon 1998] which have a given syntactic frame which can be realized with entirely different words (e.g. the form 'interrogative + *or* + interrogative', as in '*Am I right or am I right?*').

For all these reasons, the figures for each individual corpus can only give a very rough idea of how much of language is phraseological. However, they indicate that there is indeed a remarkable number of prefabricated elements to be found in everyday conversation. If it can be proved that there are also significant cluster differences between corpora of different geographical varieties of English, then this will point even more clearly to the prefabricated – or at least semi-prefabricated – character of much of language.

3.4 Comparing cluster lists

But how does one find such significant cluster differences between corpora? One method which is frequently used for calculating the significance of frequency differences between corpora is the chi-square test. Chi-square values were calculated for all clusters in the corpora, since this makes it possible to produce a list ranking all clusters from those that are most clearly different to those that are most clearly similar. There were a large number of highly significant cluster differences to be found[3]. In order to limit the output to those cases where the difference was clearest and the evidence strongest, the final version of the lists produced by the programs only contains those combinations which have a minimum frequency of occurrence of 25 *ppm* in that corpus in which they appear more frequently (i.e. at least about 99 occurrences in the *BNCSD or* about 122 occurrences in the *LSAC* in raw frequencies).

4 Results

Naturally, the clusterlists do not just show purely phraseological phenomena. Many of the clusters show syntactic differences between the corpora, such as valency differences or the fact that the past tense is used more frequently in spoken American than in spoken British English. Another, very marked difference lies in the fact that the *BNCSD* contains many more occasions of *have got* (as in *I've got*) than the *LSAC*, which simply has the corresponding forms of *have*. This difference is well-documented in the literature on differences between British and American English [cf. e.g. Strevens 1978].

It is useful at this point to come back very briefly to the question of whether the two corpora are similar enough for comparison. A problem with using standard lists of British-American differences to see whether they are reflected in the corpora lies in the fact that lists of this kind are largely made up of phenomena that are fairly rare in the spoken language. It is more useful to compare the findings in the corpora with the labelling in modern dictionaries. Indeed, a number of them can be confirmed in this way, although there are others where the evidence from the corpora is at variance with the dictionaries (see below).

However, there is a fundamental problem with using the language material of the corpora as a basis for comparison. Instead, the similarity of two corpora should be determined primarily through the similarity of the sampling technique used, since using the data itself may prejudice results. Word combinations such as *I have a*, which at first glance look as if they might function

as standard for checking the similarity of the corpora, can turn out to be useless for this purpose. This does not preclude, however, that the material itself may point to certain imbalances in the composition of the corpora. For example, the high frequency of the clusters *may I help (you), (thank) you for calling* points to the fact that there are two texts in the *LSAC* (162201 and 162202) that were recorded by a receptionist at her workplace, while the figures for other clusters such as *good boy or good girl* indicate a larger amount of parent-child conversations in the *BNCSD*. There is therefore reason for caution in evaluating the cluster frequencies. On balance, however, the experience gained through working with the corpora has not only confirmed the similarity of the sampling technique, but also the impression that – apart from the problems cited – the material that they contain is indeed comparable.

4.1 Different types of phraseological differences highlighted by the cluster method

The lists showed clearly that there are a large number of phraseological differences between American and British English. The clusters point to phraseological material of very different kinds. Some of it has previously been mentioned as being more typical of one of the two varieties, some appears to be new. The material can be grouped into different kinds of semantic, pragmatic, or structural groups. Among them are (parts of) time adverbials (such as British *half past, quarter to, (quarter) of an hour*; *at the moment, in a minute*; American *right now, at this point, (every once) in a while, the whole time*), phrasal and prepositional verbs – e.g. British *get on with, go out with*; American *go ahead (and...), figure out, work out* ('exercise') – and a very large number of conversational routines (see 4.3). Other findings concern the collocational behaviour of individual words, like the fact that *a lot* collocates far more frequently with *quite* and *not* in British English than it does in American English.

4.2 Support verb constructions and the problem of equivalence

One traditional area of collocational research are support verb constructions, that is, combinations of delexicalised verb plus object like *take a shower*, *have a swim*, etc. Authors like [Benson et al. 1986] have pointed out that certain types of support verb constructions like *have a break* versus *take a break* are typically formed in British English using *have*, whereas American English uses *take*. With respect to the *BNCSD* and the *LSAC* this has been of particular interest with regard to the noun *look*. The figures for *have a look* and *take a look* are listed in table 1.

Cluster	**LSAC**	**LSAC ppm**	**BNCSD**	**BNCSD ppm**
have/has/had/having a look	17	3.48	909	229.87
take/takes/took/ taking/taken a look	100	20.47	13	3.29
let's see	989	202.46	180	45.52
let me see	417	85.36	77	19.47

Table 1: *Have a look* vs. *take a look*

Table 1 shows that it is indeed true that *take a look* is predominantly American English, and *have a look* predominantly British. However, it is not possible to conclude from the material that *take a look* is the American equivalent of British *have a look*. It is remarkable that *have a look* appears much more frequently in the *BNCSD* than *take a look* does in the *LSAC*. Apparently, structures other than support verb constructions take over its function in American English. One of these structures - and it is likely to be the most important one besides *let's see* - is *let me see*. *Let me see* appears more than four times as frequently in the *LSAC* than it does in the *BNCSD*. In this context, it is interesting to note in which structures *have a look* appears in the *BNCSD*. Almost 20% of the 909 occurrences of *have a look* are for*let's have a look*[4]. In certain cases, *let's have a look* is used synonymously with *let me have a look* – and thus also with*let me see*, as in the following example, in which a childminder talks to one of her young charges.

> <PS14B> [...] *you've got a filthy nose.* **Let's have a look.**
>
> (*BNCSD*: kb8/98.293)

This kind of synonymous use is a typical phenomenon. There are many cases in which a more fixed, more 'idiomatic' expression in one variety corresponds to another, less clearly 'phraseological' in the other. Equivalence depends to a significant degree on context. This means that, unfortunately, it will not be possible to present a list of simple British-American equivalents. Reality is more complex than this.

4.3 The largest group: conversational routines

Most of the highly frequent recurring word combinations are conversational routines or parts of them. They are of very different kinds, ranging from hedges like *kind of* (*LSAC*) and *sort of* (*BNCSD*) to whole responses like *here you are (BNCSD)* or *here you go (LSAC)*, items on the borderline between hedge and filler, such as *and stuff (like that), and shit (like that) (both LSAC)*, or expletives like *bloody hell*, *oh dear* (*BNCSD*) or *(oh) my gosh*, *oh boy*, *oh man*, *oh wow (LSAC).*

Sometimes it is possible to find groups of such routines with a similar function and internal structure, as in the example in table 2. The items *I guess*, *I figure/I figured*, and *I believe* are used more frequently in the *LSAC*, while *I suppose*, *I reckon*, and *I expect* can be found more frequently in the *BNCSD*. *I think* has similar *ppm* frequencies in both corpora. It is interesting to note the differences for the past tense forms.

Unfortunately, however, groups like this one can seldom be found easily. The primary explanation for this is that – as with the support verb constructions mentioned – phraseological items in one variety may have equivalents in the other that are far less clearly phraseological, or at least less compact and more variable in structure.

When it comes to closer study of groups of items with a similar pragmatic meaning, it may be interesting to include items whose frequencies lie below the threshold of 25 *ppm* in both corpora. One such group is given in the upper part of table 3, table 3a, which contains different 'checking' strategies which are used to ask for repetition or clarification.

Cluster	*LSAC* **raw**	*LSAC* **ppm**	*BNCSD* **raw**	*BNCSD* **ppm**
I guess	3613	739.62	32	8.09
I figure	88	18.01	2	0.51
I figured	243	49.74	2	0.51
I believe	297	60.8	86	21.75
I think	11941	2444.44	8909	2252.95
I thought	3367	689.26	4104	1037.84
I suppose	237	48.52	1307	330.52
I reckon	7	1.43	381	96.35
I expect	24	4.91	243	61.45

Table 2: Conversational Routines

	LSAC **raw**	*LSAC* **ppm**	*BNCSD* **raw**	*BNCSD* **ppm**
"Sorry?"[5]	13	2.66	245	61.96
I'm sorry?	24	4.91	3	0.76
beg (your) pardon?	6	1.23	33	8.35
pardon me?	40	8.19	3	0.76
pardon?	64	13.10	613	155.02
excuse me?	32	6.55	3	0.76
"What?"	2091	428.05	2803	708.84
"Huh?"	1349	276.15	40	10.12

Table 3: "Checking" strategies (upper part: Table 3a, lower part: Table 3b)

The decision to study the elements in table 3a stems largely from the wish to find out whether the 'check' *excuse me?* is responsible for the high frequency of the cluster *I'm sorry* in the *LSAC* (919 occurrences, i.e. 188.13 *ppm*; as against 320 occurrences or 80.92 *ppm* in the *BNCSD*). It is interesting to note that there is a slight tendency towards the two-word combinations (*pardon me?, I'm sorry?, excuse me?)* in the *LSAC* rather than the single *Sorry?* or *Pardon?* which can be found in the *BNCSD*. As the lower part of table 3, table 3b shows, however, the most frequently used 'checks' are the one-word checks *what?* (esp. *BNCSD*) and *huh?* (esp. *LSAC*).

Most, if not all of the prefabs and single words mentioned so far were fairly easy to classify into groups. Others present more of a challenge. Table 4 lists the figures for a few items which have not been studied in detail so far. They are included here for the purpose of giving the reader a few more examples of the wide range of prefabricated and semi-prefabricated units found in the clusterlists.

5 The cluster method: a brief appraisal

Naturally, even in combination with the thresholds mentioned above, studying phraseological differences between corpora of spoken English through clusters and with methods of the kind

Cluster	*LSAC* **raw**	*LSAC* **ppm**	*BNCSD* **raw**	*BNCSD* **ppm**
mind you	19	3.89	684	172.97
I should think	6	1.23	319	80.67
tell you what	236	48.31	478	120.88
the trouble is	16	3.28	127	32.12
could do with	9	1.84	103	26.05
you're welcome	255	52.2	23	5.82
how about	491	100.51	124	31.36
you're kidding	179	36.64	11	2.78
I appreciate	128	26.2	7	1.77
would love	123	25.18	25	6.32

Table 4: Prefabs so far unclassified

described above is not without limitations. First of all, it should be pointed out that the raw figures calculated by the programs have to be confirmed by reference to the concordances. Clusters can be ambiguous and their interpretation often depends on their individual contexts. The misleading character of some other clusters originates in the fact that in both corpora clauses are not always separated by punctuation. Secondly, while the programs respect existing punctuation and speaker turns, they can only count orthographic words, not lexemes. In many cases, however, this is irrelevant since many conversational routines have a fairly fixed form including inflections. A third point of criticism is that not all irrelevant combinations are removed from the output. This mainly concerns clusters containing interjections like *er er er*. Combinations of this kind will have to be discarded manually. Finally, as stated above, a substantial number of phraseological units are ignored by the procedure chosen here. Therefore, the method does not permit accurate statistic statements of phraseological density.

On the whole, however, the approach adopted here is very suitable for a pilot-study of the kind described here. Its open, non-prestructured character does not restrict or anticipate the results in any way. It might be argued that for lexicographic purposes it would be desirable to have more powerful software which can find more of these combinations in a more automatic way like the *CQP/Xkwic* tools for finding Verb+Object-collocations described by [Docherty/Heid 1998]. However, it is doubtful whether this will indeed be possible. Prefabricated units typically have no clearly defined boundaries. The syntactic structure of routine formulae is difficult to determine in advance. Moreover, large, non-POS-tagged spoken corpora like the *LSAC* would present substantial problems to automatic analysis. For a first investigation into phraseological and collocational differences between two spoken corpora the cluster method is more appropriate.

6 Relevance for lexicography

Both lexical combinatorics and the spoken language are currently of great interest in different fields of applied linguistics. Recent trends in language teaching – like the "Wortschatzwende" or the interest aroused by books such as *The Lexical Approach* [Lewis 1993] – favour an approach

to language teaching which gives priority to phraseological aspects of the language. In order to be able to teach successfully, teachers who are not native speakers of English need to have access to collections of typical phrases, lexicalised sentence stems and other types of fixed word combinations which belong to the variety of English which they themselves have learnt and are representing in class – not just for written, but also for oral communication.

In response to the growing emphasis on the spoken language in classrooms and exams in many countries [Summers 1999] advocates the coverage of spoken English in learners' dictionaries. She argues that this can counterbalance unnatural conversations in textbooks, but also show students when spoken language is not appropriate. [Moon 1998], on the other hand, doubts whether a normal conversation is the situation in which learners would use a learner's dictionary for their encoding needs. She argues that what learners want above all from a conventional monolingual dictionary is normative information about written English. Instead, she suggests that a lexicon offering information about the spoken language should have a very different structure altogether, with increased emphasis on phraseological and collocational as well as pragmatic matters and new patterns of description for various aspects.

It is interesting to note in this context that prefabricated items from spoken English are already common in learners' dictionaries. This does not mean, however, that differences between British and American spoken English are always indicated. If, for example, one looks up the items from table 2 in the four great monolingual learners' dictionaries issued in 1995, one will see that while *I guess*, *I figure* and *I expect* are also shown to be typical of only one variety in at least two dictionaries, only *LDOCE3* labels *I suppose* and *I reckon* as *especially BrE*. There may be some connection here with Longman activities in the field of spoken American corpora, but the sample is really too small to draw any general conclusions.

There are also a number of clusters which are significantly more frequent in one of the two corpora but which are not labelled for regional variety in any of the four learner's dictionaries, even though they may be highlighted as phrases. They are of very different types, from discourse markers to whole turns, hedges, and so forth. Amongst them are, for the *BNCSD*, *as I say*, *I must admit*, *go on then*, *it's no good* (+ V-*ing*), *might as well*. The *LSAC* offers, for example, *but see*, *sounds good*, *big deal*. This could mean that these differences between American and British English have hitherto not been discovered. Alternatively, it may mean that they have been discovered but that labelling (or even inclusion) was decided against. This is likely to be the case, for example, with *sounds good* or *big deal*. Lexicographers may feel that an item is not largely restricted to or more typical of one variety – despite clear differences in frequency between corpora.

As expected, working with the two clustering programs shows that almost all of the differences between British and American English do not reflect absolute presence or absence of prefabs or multi-word units in the two varieties, but differences in frequency. Even so, these differences can be very clear. If, however, the items found with the help of the clustering programs were to be described in a dictionary, the labelling would in most cases have to read *especially British* or *especially American*.

7 A basis for monitoring phraseological differences between American and British English in the future

[Trudgill 1998] holds that, whilst British and American English are slowly diverging from each other in pronunciation, they seem to be retaining their distinctive grammatical characteristics and are, in fact, converging in terms of vocabulary. While this may be true, the sub-field of phraseology or lexical combinatorics is not touched upon at all. The traditional lists of lexical items which make up most differences between American and British English hitherto described need to be supplemented by findings on prefabricated elements of the language. The research on the *BNCSD* and the *LSAC* suggests that there are still considerable differences between the two varieties, to the point that a speaker's origin may be recognizable from the prefabs that he or she uses. With the help of the clusterlists described above, it should also be possible to monitor future developments and to find out whether these phraseological differences between American and British English will still be as noticeable in the years to come.

Notes

[1]On [Greenbaum 1988], cf. [Bublitz 1998]. A comparison of Gläser's list of phraseological units with the corresponding entries in *OALD5* and *LDOCE3* (both corpus-based) shows that some of the items that she labels as *American* are in fact found in both varieties. Amongst these are *the daily grind, give somebody the ax* (British English spelling *axe*), *no sooner said than done* and *no kidding. Get the sack/give sb the sack* is labelled "BrE informal" in *LDOCE3*. On the whole, however, this confirms her impression that both varieties are getting more similar in their phraseology.

[2]I am very grateful to Florian Klämpfl for writing all programs described here. In some respects, these programs were inspired by a tool contained in Mike Scott's concordancing package *WordSmith*. The term *clusters* was also adopted from there. However, the tool in *WordSmith* seems to be designed primarily for small amounts of text. It cannot cope with large amounts of data such as a four million word corpus.

[3]A problem of using the chi-square-test in this context is that the total of the values entered into the calculation is very large. There are several hundred thousands of cluster-tokens, even if only the clusters above the threshold of 25 *ppm* are used as a basis and this raises – and thus skews – the chi-square values for each individual cluster. It seems useful, therefore, to consider only those cluster differences for research which either have a high ratio (if possible 1:2) or are very frequent.

[4]The figures for the most frequent constructions of *have a look* are as follows: *let's have a look* 176 (= 19%), *go and have a look* 83 (= 9%), *I'll have a look* 69 (= 8%), *let me have a look* 30 (= 3%).

[5]Since the figures for single-word items are only given as a rough indication for comparison, the figures for items in double quotation marks in this table are valid for the respective spelling variant only, including capitalization and punctuation.

8 References

8.1 Articles and Books

[Aijmer 1996] Karin Aijmer: *Conversational Routines in English. Convention and Creativity*, (London/New York: Longman) 1996.

[Benson et al. 1986] Morton Benson, Evelyn Benson and Robert Ilson (1986): *Lexicographic Description of English*, (Amsterdam: John Benjamins) 1986.

[Bublitz 1998] Wolfram Bublitz: "'I entirely dot dot dot': Copying Semantic Features in Collocations with Up-Scaling Intensifiers", in: Rainer Schulze (ed.): *Making Meaningful Choices in English. On Dimensions, Perspectives, Methodology and Evidence*, (Tübingen: Narr) 1998: 11-32.

[Cowie 1994] Anthony P. Cowie: "Phraseology", in: R. E. Asher (ed.): *The Encyclopaedia of Language and Linguistics, Volume 6*, (Oxford/New York: Pergamon) 1994: 3168-71.

[Docherty/Heid 1998] Vincent J. Docherty and Ulrich Heid: "Computational Metalexicography in Practice – Corpus-based support for the revision of a commercial dictionary", in: Thierry Fontenelle, Philippe Hiligsmann, Archibald Michiels, André Moulin and Siegfried Theissen (eds.): *EURALEX'98 Proceedings, Vol. II*, (Liège: University of Liège/English and Dutch Departments) 1998: 333-345.

[Gläser 1991] Rosemarie Gläser: "Phraseologismen im amerikanischen Englisch im Lichte der Varietätenproblematik und der Lexikographie", in: K. Hansen (ed.): *Studien zur Sprachvariation im Englischen und zur Textlinguistik (unter besonderer Berücksichtigung der Lexik)*, (Berlin: Humboldt-Universität) 1991: 156-166.

[Gläser 1992] Rosemarie Gläser: "Phraseologismen im amerikanischen Englisch", in: *Fremdsprachen lehren und lernen* 21, 1992: 88-103.

[Greenbaum 1988] Sidney Greenbaum: "Some verb-intensifier-collocations in American and British English", in: Greenbaum, Sidney: *Good English and the Grammarian*, (London: Longman) 1988: 113-124.

[Hausmann 1984] Franz Josef Hausmann: "Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen", in: *Praxis des neusprachlichen Unterrichts* 31, 1984: 395-406.

[Hausmann 1989] Franz Josef Hausmann: "Le dictionnaire de collocations", in: F. J. Hausmann/ O. Reichmann/H. E. Wiegand and L. Zgusta (eds.): *Wörterbücher. Dictionaries. Dictionnaires, Volume 1*, (Berlin/New York: Walter de Gruyter) 1989: 1010-1019.

[Howarth 1996] Peter Andrew Howarth: *Phraseology in English Academic Writing. Some Implications for language learning and dictionary making*, (Tübingen: Niemeyer) 1996.

[Lewis 1993] Michael Lewis: *The Lexical Approach. The State of ELT and a Way Forward*, (Hove: Language Teaching Publications) 1993, repr. 1996.

[Moon 1998] Rosamund Moon: "On using spoken data in corpus lexicography", in: Thierry Fontenelle, Philippe Hiligsmann, Archibald Michiels, André Moulin and Siegfried Theissen (eds.): *EURALEX'98 Proceedings, Vol. II*, (Liège: University of Liège/English and Dutch Departments) 1998: 347-355.

[Strevens 1978] Peter Strevens: *British and American English*, (New York: Cassell) 1978.

[Summers 1999] Della Summers: "Coverage of spoken English in relation to learners' dictionaries, especially the Longman Dictionary of Contemporary English, in: Thomas Herbst and Kerstin Popp (eds.): *The Perfect Learners' Dictionary (?)*, (Tübingen: Niemeyer) 1999: 257-264.

[Trudgill 1998] Peter Trudgill: "World Englishes: Convergence or divergence?", in: Hans Lindquist, Staffan Klintborg, Magnus Levin and Maria Estling (eds.): *The Major Varieties of English. Papers from MAVEN 97, Växjö 20-22 November 1997*, (Växjö: University) 1998: 29-34.

8.2 Dictionaries

[BBI] M. Benson et al. (eds.), *The BBI Combinatory Dictionary of English. A Guide to Word Combinations,* (Amsterdam: John Benjamins) 1986.

[CIDE] P. Procter (ed.), *Cambridge International Dictionary of English,* (Cambridge: Cambridge University Press) 1995.

[COBUILD 2] J. Sinclair et al. (eds.), *Collins Cobuild English Dictionary,* (London: Harper Collins) 2 1995.

[LDOCE 3] D. Summers (ed.), *Longman Dictionary of Contemporary English,* (London: Longman)3 1995.

[OALD 5] J. Crowther (ed.), *Oxford Advanced Learner's Dictionary of Current English,* (Oxford: Oxford University Press) 5 1995.

8.3 Corpora

The British National Corpus is a collaborative inititative carried out by Oxford University Press, Longman, Chambers Harrap, Oxford University Computing Services, Lancaster University's Unit for Computer Research in the English Language, and the British Library. The project received funding from the UK Department of Trade and Industry and the Science and Engineering Research Council and was supported by additional research grants from the British Academy and the British Library. For more details see http://info.ox.ac.uk/bnc/

The Longman Spoken American Corpus is copyright of Pearson Education Limited and was created by Longman Dictionaries. It is available for academic research purposes, and for further details see http://www.longman-elt.com/dictionaries/