JOANNA SZWABE

THE NOT SO SKEWED CORPORA. WHY USE A CORPUS FOR A LANGUAGE STUDY?

Introduction

»Having asked a question „Could I say so and so?", many of us have encountered the response „Sure, you could say that...But I never would."« - Wallace Chafe reflects on the days when elicitation was the only way of an analysis of a foreign speech (Chafe 1992:85).

What would a native speaker really say? One way to provide answers to this question has been trusted for a long time under the name of linguistic competence. It attracts by its directness but is founded on questionable supposition. On the other pole there is corpus linguistics, with verifiable but laboriously gathered evidence.

The issues of data-based versus theory-based approaches to linguistic analysis have been subject to heated discussion for the past few decades. Opinions have been divided since Noam Chomsky questioned the relevance of collecting evidence for linguistic analysis and thus cast doubt on the structuralist approach. It is not surprising that American Structuralists who were strongly influenced by a positivist and behaviourist view of empirical sciences, favored inductive methods in their studies of the language (Leech 1991:8). In the late 50s Noam Chomsky totally reversed the situation

by restoring introspection for linguistic methodology. In his view corpora were inadequate for the language study because, consisting of the finite number of sentences, they would never be capable of reflecting any more than a fraction of the infinite language phenomenon.

According to Chomsky: „any natural corpus will be skewed. Some sentences will not occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so widely skewed that the description would be no more than a mere list" (Chomsky 1962 in Leech 1991:8). The interests of linguists suddenly turned to the intuitions of a native speaker and corpus linguistics became marginal in most countries.

There are, however, serious reasons why not to fallow Chomsky in this respect. The purpose of the present paper is to prove that corpora are not only adequate but superior to any other method of linguistic study. I am strongly convinced that corpus based research is much less likely to distort a view of language than elicitation or introspection. But what is far more crucial, corpora reveal facts about language that are not visible to other methods and would most probably remain undiscovered. There are numerous instances of linguists reporting discoveries inspired by corpus studies. Others claim that testing their hypothesis against corpora led them to corrections of their view of certain language phenomena. I will refer to some of the cases later.

The term 'corpus' will be used to refer to a collection of naturally occurring written or spoken texts which is stored in a machine readable format for the purposes of linguistic description or as a means of verifying hypotheses about language. It is approved in the field of corpus linguistics to use the term 'text' in reference to both written and spoken language (Renouf 1987; Clear 1987; Sinclair 1987 & 1991; Burnard 1995). The primary tool for retrieving linguistic information from machine readable database are concordances produced by relevant corpus query software. The recorded data are transformed into a series of one-line extracts presenting a keyword in its immediate context (see a specimen on p. 10).

To my knowledge corpora remain virtually unknown outside linguistics. I assume this will require a brief introduction to basic concepts and applications of corpus research. Whoever is acquainted with corpus linguistics may skip sections 3 and 4.1 covering these issues.

1. Starting with the validity of methods

In the time Chomsky had an opportunity to judge the usefulness of corpus research, corpora were mostly of elicited type and were compiled mainly for the purposes of phonological research (Sebba & Fligelstone 1994:770). Typically, these were small collections, limited to a certain language variety and consisting of data acquired by interviewing a native informant, thus displaying little resemblance to today's exhaustive corpora, representative of language as a whole, and what is decisive here, containing only naturally occurring language material. Until the early sixties there was hardly any cooperation between linguistics and computer science and Chomsky, as many other

linguists, failed to predict how the potential of large collections of verifiable data may be exploited with the use of automated techniques for organizing and inquiring a body of language material.

Handling large corpora demanded being both computerate and competent in linguistics. No wonder nascent computational linguistics was limited to a handful of far-sighted enthusiast like Henry Kucera the coordinator of the Brown Corpus project. The first Brown Corpus computer in 1960 had less than 40 KB of core memory, an extremely modest capacity by today's standards, and the text was stored on 100,000 of punched cards. In a letter to Sonoda Katsuhide on the corpus linguistics internet discussion list Kucera writes: „The initial sort of the one million records of the Brown Corpus took 17 hours of uninterrupted processing when I had to reserve the machine for the entire weekend."[1]

Technical demands of corpora in an electronic format and also high costs of compiling a corpus, often involving sponsorship from commercial institutions, are responsible for restricted availability of corpora. The situation gradually improves as internet becomes more and more widespread - corpora can be accessed and interrogated from web-pages while the precious database remains behind the scenes. Carefully constructed, large and diverse corpora, though designed with linguistic purposes in mind, might be equally useful outside linguistics in psychology, sociolinguistics and cross-cultural studies, as they already are in artificial intelligence and speech recognition projects.

After all the significant changes that corpus linguistics has undergone over past four decades the disagreement parting intuition-based and data-based approaches still persists. It is so probably because it lies deeper than technical distinctions and is concerned with value of evidence introspection may provide.

First of all, there are certainly areas of research where introspection is excluded. Native speaker's intuition is clearly inadequate when it comes to study of historical language. Since Chaucer's contemporaries are no longer around we are bound to rely, this way or another, on the evidence of remaining texts. It obviously more beneficial to rely on a large and computer accessible collection such as, for instance, The Penn-Helsinki Parsed Corpus of Middle English created under the direction of Matti Rissanen at the University of Helsinki.

Secondly, when a researcher is not a native speaker of a language he wishes to examine, he must rely on elicitation (interviewing a native speaker informant) which is just another way of entrusting introspection but in an even less controlled way. John Sinclair points to another disadvantage of the elicitation method: „The informant will not be able to distinguish among various kinds of language patterning - psychological associations, semantic groupings, and so on. Actual usage plays a very minor role in one's consciousness of language and one would be recording largely ideas about language rather than facts of it" (Sinclair 1991:39). The difference between introspection and corpus study seems to be the difference between opinions and evidence.

[1] Source: http://nora.hd.uib.no/corpora/1996-3/0088.html

2. A Case Study

We used to treat grammatical correctness as strictly binding as opposed to appropriateness where choices are believed to be largely optional. When it comes to language habits we find ourselves unable to explain why we express things the way we do and also less sensitive to irregularities. However patterns found in native corpora as compared to learner language corpora such as the International Corpus of Learner English (ICLE) conducted by Professor Sylviane Granger at Centre d'Etudes Anglaises, Université Catholique de Louvain suggest that grammaticality extends its influence to the fuzzy area of appropriateness.[2] The rules guiding appropriateness escape our insight into language competence. Identifying those systematic patterns seems to be a suitable task for specific tools of corpus linguistics.

Examining selected sentences in respect to their grammaticality or lack thereof we often tend to ignore their naturalness. The naturalness, that is responsible for sounding native (the phenomenon which is currently being investigated by the ICLE project) can be best explored through 'prolonged exposure to corpora' as Wallace Chafe – a veteran corpus linguist - a bit ironically calls this laborious task.

A detrimental impact of the less humble, corpora-free, approach is explicitly exemplified in Wallace Chafe's sharp critique of invented examples being at the same time a case study in corpus research. Edward Sapir, by no means an armchair linguist, well known for his work on ethnic languages, becomes a target of Chafe's criticism for having entrusted the sufficiency of artificial examples for illustrating the functions of morphological and syntactic elements. An extraordinary sentence *The farmer kills the duckling,* Sapir used to illustrate how derivation, inflection, and word order contribute to the understanding of the sentence, is overtly unnatural even to a novice. The use of the present tense instead of the progressive aspect here conflicts with real life discourse habits. The other more likely expression would be **The farmer killed the duckling* but it would lose the *–s* ending, which was one of the points in Sapir's argument. It becomes even more problematic in the light of Chafe's findings in a conversational corpus, namely the „light subject constraint" and the „one new idea constraint" (Chafe 1992:87-95).

The light subject constraint says that a subject in conversational language cannot express new information. A subject of a clause is bound to be either given (i.e. assumed by the speaker to be already active in the consciousness of the addressee) or accessible (where the referent is presumed to be semiactive in the consciousness of the addressee), in other words 'not new'. Interlocutors simply do not say anything like: **A burglar stole my camera yesterday*, where *the burglar* remains to be important in the conversation. Misleadingly the sentence is acceptable for native speakers. Consequently, for a Sapir's example to be a realistic one, its subject should be either given or accessible. But if it was given it would not be repeated as *the farmer* but pronominalized into *He kills the*

[2] My observations are based on the contrastive studies of the LOB and the Polish subpart of the ICLE, available from the Institute of English Philology at the Adam Mickiewicz University, Poznan; a detailed analysis of which exceeds the span of this paper. There are, however, several other traces which may be interpreted in favour of this hypothesis and which are closely related to the subject of the present study also in other respects, namely that of Chafe's, Sinclair's and Halliday's. For now the question of appropriateness will be confined to the implications of these.

duckling. A more refined examination of the corpus, however, revealed a small residue of subjects (3%) that do express new information. Although it is not a serious counterexample to the light subject constraint as a strong tendency the finding intrigued Chafe and this way led to new studies with new discoveries. The characteristic feature of the exceptional subjects conveying new ideas is that they express referents of minimal importance in the discourse, thus decisively excluding a subject as location of both new and important information and shifting interest to a predicate.

Again, tested against the spoken corpora, new information has been found to be limited to no more than one idea that is activated in the current discourse for the first time. The remaining ideas must be either given or accessible. It is only through frequency analysis of a corpus that we can testify if one new idea constraint, can be treated as a rule. The answer from the corpus indicates that minor irregularities of this rule fall into two classes: one represents low content verbs spoken typically with secondary stress, as in *I just talked to Jim.* By contrast, sentences containing both high content verb and a new information as in **I just complimented Jim* do not occur in real language and it is their absence that supports one new idea constraint. Exceptions of the other kind are examples in which the entire verb-object phrase has been lexicalized, like in an idiomatic expression: *They were dragging their feet* where the idea of *dragging* cannot be activated separately from the idea of *the feet*. Not only Chafe's counterintuitive hypotheses have been verified by the corpus study but also the analysis involved provided a clearer understanding of related phenomena of low content verbs and lexicalization (Chafe 1992:88-95).

What we do not find in corpora are examples resembling **A burglar stole my camera yesterday* and **The farmer killed the duckling*. In the case of the latter, as *kill* is hardly a low content verb, the ideas of *killing* and that of *duckling* must be separate. On the contrary, normally the event of killing would be expressed in a context where the ideas of both the farmer and the duckling were given. What natives would most likely say is: *He killed it*. The original Sapir's sentence violates constrains of real conversation and therefore turns out to be irrelevant for proving facts about language. The inappropriateness of these invented examples, however, is invisible for introspection and elicitation – thus undermining the quality of these methods.

This is not to say that we should abandon any other method of linguistic research but a corpus. As Sinclair notes, the influence of personal intuition is in fact inevitable but it's place is in evaluating evidence rather than creating it (Sinclair 1991:39).

3. The microcosm of corpora

„A word must occur to remain in the language, and therefore to be the concern of lexicographers of the contemporary language." (Sinclair 1991:44)

One of the ways to avoid a corpus being skewed is to compile material on one hand large and on the other balanced enough to be representative of a language. Historically, the representative approach can be traced from the Brown Corpus, first made available in 1964 under the name of a Standard Sample of Present Day American English (Kucera

& Francis 1964), and the Lancaster-Oslo-Bergen (LOB) corpus of British texts, completed in 1978 (Johansson & Leech 1978). Those challenging pioneer projects, not only were launched in opposition to the mainstream theoretical linguistics of the times but also ventured adapting new computational techniques to language analysis, this way helping corpus linguistics become recognized and letting today's major projects such as the British National Corpus and the COBUILD come into being.

The largest Corpus of English language so far is the COBUILD (COllins Birmingham University International Language Database) - the latest release of the corpus in 2000 amounted to 415 million words sampled from a variety of texts and 20 million words of transcribed natural speech, and it still continues to grow.[3] COBUILD started in 1980 as a joint project between the University of Birmingham and Collins Publishers, specializing in the preparation of reference works for learners of English. The corpus has been designed to mirror the language that surrounds us but is organized in a more accessible form and equipped with a variety of fast tools for information retrieval.

What Chomsky viewed to be a drawback of a corpus - that some of the sentences do not occur in it - should be rather perceived as a part of information it provides. As Wallace Chafe clearly points out „this disadvantage is partially offset by the fact that the very infrequency of the occurrence of something is likely to be a relevant observation in itself" (Chafe 1992:84).

Secondly, if a given phrase structure, which our linguistic competence allows us to form, does not occur in a balanced corpus of over 400 million words - this should not be a fact to be ignored when considering it as an argument in discussion.

Of course, one might say that it is largely a matter of individual preferences on what a researcher wants to focus: either hypothetical or factual language. But it is certainly the latter that must be central to lexicographers who are responsible for providing a model of language to learners and should bear on their decisions on what is essential and what is peripheral in language teaching.

Apart from the requirement of being exhaustive the emphasis is placed on balance. Probably the best illustration of a balanced corpus is the British National Corpus (BNC), completed in 1994 by an industrial-academic consortium lead by Oxford University Press. Similarly as the COBUILD, the BNC is over 100 million words monolingual, synchronic corpus, both spoken and written.

It requires a handful of details to make it apparent how carefully modern corpora are constructed. The written component of the BNC is composed of current imaginative and informative texts samples, generally no longer than 45,000 words to avoid over-representing idiosyncrasies. As a general-purpose corpus it must take account of both perspectives: production and reception. Therefore, unpublished letters, memos, reports, essays, written-to-be-spoken materials (e.g. play scripts) were also included (Burnard 1995:5-11).

Both for written and spoken parts representativeness was achieved on one hand by demographic sampling in terms of age, gender, social group, and region, and the context-governed part on the other (Burnard 1995:22-23). A corpus constructed in this way reflects current English in its entirety and additionally serves as a testbed for contrastive studies of various text types.

[3] Data acquired from COBUILD website http://titania.cobuild.collins.co.uk

As for spoken corpus, practical difficulties of transcribing sufficiently large quantities of text have prevented the construction of a spoken component of over four million words. Unquestionably, there is a greater need for reliable data on speech than it is for a written word, for it is only in natural, spontaneous interaction, that lexicogrammatical potential of the system is brought into play. M. A. K. Halliday vividly reports: „If you listen grammatically, you will hear sentences of far greater complexity than can ever be found in writing - sentences which prove barely intelligible when written down, yet were beautifully constructed and had been processed without any conscious attention when they occurred in natural speech. I had heard verbal groups like had been going to have been paying and will have been going to have been being tested tripping off the tongue, at a time when structuralist grammarians were seriously wondering whether something as forbiddingly complex as 'have been being eaten could ever actually be said!" (Halliday 1991:62)

The spontaneous and temporal character of speech acts can only be frozen by recording a speaker unaware of being recorded. Undoubtedly, the above method is far superior to elicitation which creates an unnatural laboratory-like situation. It is an aspect of notorious observer's paradox problem that informants change their verbal behaviour as soon as they realize they are being recorded. The naturalness of data collected in this way is questionable, and so are the results of the study based on the observations that are distorted from the very start.

During the BNC project around 700 hours of recordings were gathered and over four million words conversational English were transcribed. It is hard to realize what a challenging project it was. Selected individuals used a portable tape recorder to record language people use in everyday conversation (Burnard 1995:19-21). In order to be consistent with one's right for privacy all participants were told they had been recorded and explained why. If requested, the recording was erased.

Especially promising corpora are those consisting entirely of recorded (not transcribed) speech. Such corpora are already being constructed, some even spontaneously by researchers who interchange recorded data via internet. To name a few: projects found in CHILDES[4] database for child language research, i.e. Neonatal and Infant Cry archive[5], or Ann Peters study of fillers, unglossable syllables that children produce as they move from 'one word' stage to 'two words' stage in language production[6]. Child speech contains properties not subject to standardized description and thus if transcribed may be distorted or certain features may be omitted. Internet database of recordings solves these problems by making the material accessible in an original form to everyone, and not restricted exclusively to a researcher who has collected it.

[4] CHILDES http://childes.psy.cmu.edu

[5] Neonatal and Infant Cry Archive http://www.siu.edu/departments/coe/comdis/special.html

[6] Ann Peters „Filler Syllables: What Is Their Status In Emerging Grammar" unpublished manuscript available from http://www2.hawaii.edu/~ann/filler.pdf

4. What is it like to interview a corpus?

4.1. Annotation

It is apparent that the quality of the evidence corpora can provide depends firstly, on the selection of material to be concorded and various criteria for this selection; and secondly, on the processing of the material gathered, with a view of attaining maximum effectiveness of the retrieval techniques. The latter is most often referred to as corpus annotation.

Concordance programs sort and count objects they find in a corpus – which, in what is called a 'raw' corpus, are strings of characters between spaces. A lot can be done with raw text but in order to exploit the potential of corpus data to the full the text must be annotated with a number of features including part-of-speech tagging, marking paragraphs, sentence boundaries, and headings, etc. in written texts, and speech turns, pausing, para-linguistic features such as laughter for spoken texts, all encoded in the Standard Generalized Markup Language (Johansson 1991:308-309). Some problems posed by assigning an item to a single word-class by a machine are illustrated by a piece of the BNC text, raw and annotated with an automatic stochastic tagger, which is given in the annex.

Once a corpus has been tagged, it becomes a rich source of data for further research, e.g. the study of word-class combinations (Johansson & Hofland 1989:1). An annotated input enables a concordance program to search for grammatical information, such as instances of the passive voice, of the progressive aspect, of a noun-noun sequences, etc. (Leech 1991:19).

While automatic part-of-speech tagging has been a success in computational linguistics, automation of syntactic, and especially semantic, annotation techniques is still problematic. The process of building a 'treebank' by a computer (parsing) still requires human assistance (Leech 1991:20). However, there are already parsers which are able to perform some manipulation on the material given, e.g. change questions to statements, active voice to passive and vice versa.[7] The remaining problem is the extent of the accuracy of the operation which so far has not been satisfactory.

There have also been some attempts to introduce semantic annotation to computerized corpora like, for example, Douglas Biber's (1991) study of referential strategies with the combined method of hand editing and computer processing.

4.2. Concordance

One of the tenets of data-driven approach is that a expression gains its meaning from the context. In corpus linguistics the context is provided by concordances.

Although concordance is a central notion and an essential tool in corpus lexicography its purpose was not originally linguistic.[8] The most widely used format of computer

[7] Cf. Ergo Linguistic Technologies Parser developed by Philip.A Bralich,. & Derek Bickerton in 1997, demonstration version, www.ergo-link.com.

[8] It was the conviction that the parts of the Bible are consistent with each other, as parts of a divine revelation, that made the exegetes embark upon a task of compiling concordances (Ashmore

concordancing is an effective convention known as Key Word In Context (KWIC). Its distinctive feature is that it is focused mainly on the immediate context of a word or, in other words, its co-text.[9] The idea is emphasized by a characteristic display arrangement as in a specimen presented below.

A KWIC concordance for 'decline' from the Bank of English acquired through COBUILD*Direct* Corpus Sampler via Internet.

```
Communist Party showed a further decline and our correspondent says perhaps
     you've already started to decline. Anyway, you'll never be this
Scottish stout brewing went into decline due to Government policy during
Stott superbly portrays MacBryde's decline from energetic mento to wino and
in May and twenty six in April. The decline has been more noticable in the
Minister, responding to the reported decline in cycling mileage (down 19 per
factories in 1915 had resulted in a decline in the quality of the shells used.
   society and the corresponding decline in the influence of the common
       [p] In the long term the decline in the number of traffic-
have to do [p] Wilson has suffered a decline in fortunes over 1,500 metres
   Many of the reasons for this decline in interest were apparent from
```

Working with concordancing programs many lexicographers (Krishnamurthy 1987; Renouf 1987; Sinclair 1991) notice that concordance is a phenomenon of a two-fold nature, namely the nature of a graphic display and that of the linguistic information which it contains. It should be emphasized that the very form in which a citation is presented to a researcher contributes to the specific nature of concordance's efficiency and usefulness. In concordance lines words following the headword may be arranged in the alphabetical order which often makes collocational patterns immediately apparent. For instance, a preposition which frequently occurs right after the headword indicates a syntactic requirement. Similarly, a noun dominating a post-headword position may suggest a nominal compound (Krishnamurthy 1987:75).

John Sinclair (1987,1991), who has been advocating data-driven lexicography for the past twenty years, often reports on concordance as being superior to any other method for it not only brings reliable data, but what is more, frequently uncovers unexpected facts about language (Sinclair 1991:42). Having a large database at his disposal Sinclair gives a powerful example right in the Introduction to *Looking up* - an account of the COBUILD Project: „we think of verbs like *see*, *give* or *keep*, as having each a basic meaning; we would probably expect those meanings to be the commonest. However, the

1963:261). In Medieval Latin already the notion of 'concordantia', understood as a parallel use (of a word), came into use. It was usually used in reference to the Bibliae concordantiae prepared to reveal scriptural relationships which otherwise would be hardly discernible (Stanley 1996:682).

Concordance was soon found helpful in interpretation and explanation of non-biblical texts as well, and books of concordances for secular literature were prepared early.

Since the computers made concordances easy to compile there has been a great expansion in concordancing. Today literary works, linguistic corpora, and still the Bible may be concorded via Internet without the actual texts being published online (Foster 1998:2).

[9] The co-text of a selected word or phrase consists of the other words on either side of it. This is a more precise term than 'context' which may mean either immediate lexical surrounding of a given word or the general, non-linguistic environment of any language activity such as the sociocultural background (Sinclair 1991:171-172).

database tells us that *see* is commonest in uses like *I see*, *you see*; *give* in uses like *give a talk* and *keep* in uses like *keep warm*" (Sinclair 1987:vii).

On the grounds of his experience in the field of corpus lexicography Sinclair strongly opposes relying on introspection when making statements about the nature of lexical behaviour. „It is clear - he writes - that the early stages of computer processing give results which conflict with our intuitions. Current work in lexicography shows that, for many common words, the most frequent meaning is not the one that first comes to mind and takes pride of place in most dictionaries" (Sinclair 1991:36).[10] Just examining frequency listings we encounter facts that once found may not seem totally unexplainable but which otherwise would probably never come to mind. It is striking, for example, that in the LOB Corpus most modal verbs are not among first fifty commonest words (cf. frequency tables in Frequency analysis of English Vocabulary and Grammar, Johansson & Hofland 1989:19-20). Which words are usually at the top of the list? The modest ones - those which convey much less semantic information than, for instance, nouns or verbs. *The*, *of*, *to*, *and*, *a* are the highest ranking words in corpora as diverse as the native English LOB and the learner English ICLE.

A possible drawback of the computer method might be that concordancing programs allow identification of words by spelling only resulting in inclusion of homographs and omission of plural forms, simple (*heart*, *hearts*) and irregular (*child*, *children*), and also verb forms (*find*, *finds*, etc.). That disadvantage can be avoided by prior lemmatizing of a corpus. Lemmatization is the process of arranging the composite set of word-forms called lemmas or lemmata (Sinclair 1991:41). Corpora may be variously lemmatized, which has an impact on concordances based on them.

Lemmatization itself poses an interesting problem, for it is far from being obvious how meanings are distributed among different word-forms. Let me refer to an example: in the case of *decline* the study of concordance lines in relation to shades of meaning shows that its nominal usage tends towards *deteriorate*, while verbal and adjectival use shows the opposite bias and the trace of the *deteriorate* sense entirely disappears in technical terms. The other main sense, that of *refuse*, is verbal, associated particularly with the form *declined* (Sinclair 1991:51). Further analysis of the COBUILD Sample Corpus of 7.3 million words shows that *decline* in its uninflected form, which appears in dictionaries as a headword, does not follow the pattern of the verb forms, but overwhelmingly is used as a noun (14 instances of verbal use as opposed to 108 of nominal use), while *declining* is used more often in adjectival and not verbal sense (Sinclair 1991:46). It should not be neglected that learners and translators face real, not idealized, language which means they have roughly seven times as much chance of encountering *decline* as a noun than as a verb. Dictionaries suggest different picture of how the word should be used: for instance in the Collins English Dictionary (the one Harper Collins had issued before they compiled the COBUILD) the most often used nominal form is given as sixth sense of a headword. Instead, the CED gives a place to some of the rarely

[10] It is a layman's idea that it is enough for a person who intends to work in an English-speaking country to know five words in English, these include: *can*, *make*, *take*, *give* and *come*. However simplistic this might be the prevailing view suggest the expectations about the frequency and probably reflects intuitions shared not only by laymen. Surprisingly, in the LOB corpus of native English *can* is not in the first fifty commonest words, *come* occupies 150th position and *give*, *take*, *make* are even less frequent (cf. frequency tables in Johansson & Hofland 1989:19-20).

encountered: *declinometer*, *declensionally*, *decliner*. So what should be the reasons behind the choice of some forms and omission of others? A definite answer comes from Sinclair, a leader of the COBUILD lexicographic team: „word-formation rules are highly productive, and only the evidence of text is likely to control what is otherwise a monstrous list of forms" (Sinclair 1991:45). Recently there have been some attempts to give an account of how the resources of the lexicon are expressed in actual usage: there is a relatively small number of words which occur over and over again, while the remaining rest are not frequently engaged in discourse. The idea is to concentrate on the language that students do need to command as opposed to its theoretical models that aim at describing language in its entirety.[11]

It is often reported by teachers that when learners of English of an upper-intermediate level use words looked up in a dictionary in their writing they are almost never used appropriately. In fact, the essays are better when written without assistance of dictionaries. Why standard dictionaries fail to explain meanings of words? It is obvious that grasping the point does not automatically lead to appropriate usage. One can say that word's meaning cannot be separated from its particular syntactical behaviour. It seems that it is so because meanings cannot be fully described by other words that seem to be *synonymous* - they are better explained by the words *with* which they occur. Recently it has been acknowledged by corpus lexicographers that context not only disambiguates but also reveals word's typical syntactical and collocational patterns. Made-up or edited examples in dictionary entries, blamed for creating models of English which encourage unnatural expression, have been substituted with exhaustive authentic instances of occurrence of a given word in the corpus (Fox 1987:147-149).

5. Lexicogrammar continuum - a view from a corpus

Perhaps the most significant of corpus inspirations is the impression of continuity characterizing language as it occurs as opposed to dichotomous distinction of semantic and syntactic dimensions in terms of which we used to describe language. With that view in mind corpus linguists M. A. K. Halliday and John Sinclair approach language continuum from two different conventional perspectives of grammar and lexicon.

We persist to describe vocabulary as a finite set of fairly definable units which can be fitted interchangeably into a stable framework of grammar, this way giving rise to an

[11] An interesting application of corpus-driven lexicography is the Collins COBUILD Student's Dictionary - the prototype of an online dictionary including over fifty hours of authentic speech. The publishers claim that - based on the statistical approach - the dictionary accounts for ninety per cent of the English language that is written and spoken (see http://talisker.linguistics.ruhr-uni-bochum.de).

Another attempt to bring authenticity into teaching materials is the COBUILD machine readable dictionary (1995) containing a 5mln word corpus which enables advanced learners to carry out their own searches.

Furthermore, selected concordances for classroom teaching, focused on a particular area of English grammar or vocabulary, have been issued (Goodale 1995:5). Recently there have been also more and more attempts to introduce concordances to classroom teaching (Johns 1998:1) and it is not very unlikely that they will become more widely used in the future.

infinite stream of unique sentences. This is the account which is both intuitive and having an established methodological tradition, thus not likely to be easily abandoned. Halliday, a grammarian by origin attempts to overcome the discrepancy posed by the very nature of syntax and semantics by incorporating the two dimensions into a unified model of lexicogrammar. It is possible, he claims, to interrogate the whole system grammatically as well as lexically: „The amount of effort required to get grammar-like answers increases, and the payoff goes down, as you move towards the lexical end" and vise versa. This does not mean that grammaticality and lexicality are spread evenly over the language continuum. Some areas, such as English circumstantial system, systems of modality and temporality are penetrable more or less equally well from either pole, while phenomena at the extremes are qualitatively different. Furthermore, the intertwined relations between a lexis and its neighbourhood point to local and transitive character of networks rather than global and persistent nature often ascribed to language as a system (Halliday 1991:62-65).

Sinclair, who was once Halliday's co-worker at the project of collection of spontaneous English conversation corpus in the early 1960s, used to be „tunneling through the system" from the other pole. In the course of his lexical investigations of various corpora he found himself moving further and further towards the grammatical end. From this perspective, current semantic models seem to be wrong focusing exclusively on the paradigmatic dimension. Meaning is believed to be created through choice, a choice, that is made within fixed frames provided by the syntagmatic dimension. However, there are strong patterns found in large corpora suggesting that a change of choice induces a change in environment. Involved in collocational patterns the word „is not normally the unit of meaning, because it has too little freedom of choice to create much meaning" (Sinclair 1997:119). It seems that the only way to grasp meaning is to step down between the words. A corpus-based analysis offers not previously encountered possibility to derive the stable elements of a word's meaning without actual isolating it from its original contexts just by the frequency of their occurrence and the regularity of their distribution.

Again, for Sinclair it is possible to reconcile the paradigmatic and the syntagmatic within a framework that distributes meaningfulness correctly within both dimensions (Sinclair 1997:119-121). These inspirations are still far from being systematized but it has already become clear that aiming at an adequate linguistic description we have to adjust our received perceptions of language.

Annex

A sample below indicates some difficulties with attaching a word to a single category, that arise during automatic tagging (see items *lurk* and *Sinkport*). The computer programme used for the demonstration was CLAWS - a stochastic tagger developed by Roger Garside used for annotation of the LOB corpus and also, in its later version, the British National Corpus (BNC). This automatic procedure has an error rate of around 1.7% (Garside 1997:3).

Little does he realise what villainy and treachery lurk in the little town of Sinkport, or what a hideous fate may await him there.

Little&DT0; does&VDZ; he&PNP; realise&VVI; what&DTQ; villainy&NN1; and&CJC; treachery&NN1; lurk&NN1-VVB; in&PRP; the&AT0; little&AJ0; town&NN1; of&PRF; Sinkport&NN1-NP0; &PUN; or&CJC; what&DTQ; a&AT0; hideous&AJ0; fate&NN1; may&VM0; await&VVI; him&PNP; there&AV0; &PUN;

The symbols of the CLAWS tagset which were used in the above fragment are explained below.

AJ0 adjective (unmarked) (e.g. GOOD, OLD)
AT0 article (e.g. THE, A, AN)
AV0 adverb (unmarked) (e.g. OFTEN, WELL)
CJC coordinating conjunction (e.g. AND, OR)
DT0 general determiner (e.g. THESE, SOME)
DTQ wh-determiner (e.g. WHOSE, WHICH)
NN1 singular noun (e.g. PENCIL, GOOSE)
NP0 proper noun (e.g. LONDON, MICHAEL)
PNP personal pronoun (e.g. YOU, THEM)
PRF the preposition OF
PUN punctuation - general mark(i.e. . ! , : ; - ? ...)
VDZ -s form of the verb „DO", i.e. DOES
VM0 modal auxiliary verb (e.g. CAN, 'LL)
VVB base form of lexical verb (except the infinitive)(e.g. TAKE)
VVI infinitive of lexical verb

References

Ashmore, Harry S. (Ed. In Chief
1963 Encyclopaedia Britannica Chicago: Encyclopaedia Britannica Inc. William Benton Publisher, Volume 6

Biber, Douglas
1991 „Using computer-based text corpora to analyze the referential strategies of spoken and written texts" in Jan Svartvik (ed.) *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82. Stockholm.1991. Berlin: Mounton de Gruyter; pp. 213-252

Burnard, Lou (ed.)
1995 *British National Corpus Users Reference Guide* Version 1.0, Oxford: Oxford University Computing Services.

Chafe, Wallace
1992 „The importance of corpus linguistics to understanding the nature of language" in Jan Svartvik (ed.) 1992 *Directions in Corpus Linguistics*: Proceedings of Nobel symposium 82. Berlin: Mouton de Gruyter; pp.79-97

Clear, Jeremy
1987 „Computing" in John Sinclair (ed.) 1987. *Looking up*. London: Collins; pp.41-61

Renouf, Antoinette
1987 „Corpus Development" in John Sinclair (ed.) 1987. *Looking up*. London: Collins; pp. 1-42

Foster, Greg
1998 TSEbase: The Online Concordance to T. S. Eliot's Collected Poems. http://www.missouri.edu/~enggf/tsebinfo.html

Fox, Gwyneth
1987 „The Case for Examples" in John Sinclair (ed.) 1987 *Looking up*. London:Collins; pp. 147-149

Garside, Roger
1997 „Using CLAWS to Annotate the British National Corpus". http://info.ox.ac.uk/bnc/what/garside_allc.html

Goodale, Malcolm
1995 *Collins COBUILD Concordance Samplers 2: Phrasal verbs*. Musselburgh: Harper-Collins Publishers.

Halliday, M.A.K
1991 „Corpus studies and probabilistic grammar" in Karin Aijmer, Bengt Altenberg (eds.) 1991 *English Corpus Linguistics*. New York: Longman; pp. 61-77

Johansson, Stig & Leech, Geoffrey
1978 *Manual of Information to Accompany the Lancaster-Oslo-Bergen Corpus of British English, for Use with Digital Computers*. Bergen: ICAME, The Norwegian Computing Centre for the Humanities.

Johansson, Stig & Hofland, Knut
1989 *Frequency analysis of English Vocabulary and Grammar*. Oxford: Clarendon Press.

Johansson, Stig
1991 „Times change, and so do corpora" in Karin Aijmer & Bengt Altenberg (eds.) 1991. *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman Group UK; pp. 305-314.

Johns, Tim
1998 „Data-driven Learning". http://sun1.bham.ac.uk/johnstf/timconc.htm.

Leech, Geoffrey
1991 „The State of the Art in Corpus Linguistics" in Karin Aijmer & Bengt Altenberg (eds.) 1991. *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman Group UK; pp. 8-29.

Krishnamurthy, Ramesh
1987 „The Process of Compilation" in John Sinclair (ed.) 1987. *Looking up*. London: Collins; pp. 62-85

Kucera, Henry & Francis, W. Nelson
1964 *Manual of Information to Accompany a Standard Sample of Present Day Edited American English, for Use with Digital Computers*. Brown University: http://khut.hit.uib.no/icame/manuals/brown/index.htm

Sebba, M. & Fligelstone S.D
1994 R.E. Asher Editor-in-Chief 1996 The Encyclopedia of Language and Linguistics. Pergamon Press Ltd., Volume 2; pp. 769-772

Sinclair, John
1987 „Grammar in the Dictionary" in John Sinclair (ed.) 1987. *Looking up*. London: Collins; pp. 104-115

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

Sinclair, John
1997 „The Problem of Meaning" in Ruta Marcinkeviciene, Norbert Voltz (eds.) 1997. *TERLI Proceedings of the Second European Seminar*: TERLI; pp.119-121

Stanley, E.G.
1996 in R.E. Asher Editor-in-Chief 1996 The Encyclopedia of Language and Linguistics. Pergamon Press Ltd. Volume 2, pp. 681-682.