JONĖ GRIGALIŪNIENĖ

CORPORA IN THE CLASSROOM

VILNIUS UNIVERSITY
2013

Apsvarstė ir rekomendavo išleisti
Vilniaus universiteto Filologijos fakulteto taryba
(2012 m. kovo 30 d.; protokolas Nr. 6)

Recenzentės:
doc. **Nijolė Bražėnienė**
doc. Dr. **Nijolė Maskoliūnienė**

ISBN 978-609-459-150-1

1 INTRODUCTION TO CORPUS LINGUISTICS. A BRIEF HISTORY OF CORPUS LINGUISTICS. BEST-KNOWN CORPORA.

Corpus linguistics can be described as the study of language based on text corpora. What is a corpus? A *corpus* is a fashionable word today. Everything that used to be called *data* a few years ago is now a corpus. It should be noted, however, that not every haphazard collection of texts is a corpus. Most linguists (Kennedy1998, Aston and Burnard 1998, McEnery 2006, Sinclair1991, Leech and Fligelstone1992) make a distinction between a corpus and an archive, the latter being defined as an opportunistic collection of texts. The term *corpus* in modern linguistics is used to refer to a collection of sampled texts, both written and spoken, in a machine-readable form. There are many ways to define a corpus, but most scholars agree that a corpus is a collection of machine-readable, authentic texts, chosen to characterize or represent a state or variety of a language. The issue of what makes a corpus representative is rather contentious. What does it mean to represent a language? According to Leech (1991:27), a corpus is thought to be representative of the language or its variety if the findings based on its contents can be generalized to the language as a whole or a specified part of it. However, as G. Kennedy rightly points out (1998:62), the issue of representativeness is in fact "representative of what"? Can a sample of texts represent a language or a variety as such? And yet, as Kennedy observes (1998:62) that:

> It remains a legitimate goal for the compilation of a corpus to be representative of a language. After all, generalizations are an essential part of science and we have no difficulty accepting generalizations about the human body in the diagrams in an anatomy text even when we know that every person's body is different from those diagrams.

Representativeness of most corpora, as it is often maintained (McEnery 2006:13), is determined by two factors: balance (i.e. a range of genres included in the corpus) and sampling (i.e. the selection of texts). A balanced corpus should cover a wide range of text categories which are supposed to be representative of the language or a variety under consideration (McEnery 2006:16). One should bear in mind, however, that at present there is no reliable scientific

measure of corpus balance, therefore, the issue of a balanced corpus is more a matter of faith than a statement of fact (McEnery 2006: 16). The question of text selection is equally intractable. Summers (1991) presents a number of possible approaches to text selection: an 'elitist' approach based on literary and academic merit or 'influentialness', random selection; 'currency', or the extent to which the text is read, subjective judgement of 'typicalness'; availability of text in archives; demographic sampling of reading habits, etc. A pragmatic approach would be to use a combination of these approaches to select text types and sources, taking into account 'currency' and 'influentialness'. (See *Corpus Creation* section).

A brief history of corpus linguistics

Although the use of authentic examples from selected texts has a long tradition in English studies, there has been a rapid expansion of corpus linguistics in the last five decades. This development, as it is often maintained, stems from two important events that took place around 1960. One was Randolph Quirk's launching of the Survey of English Usage (SEU) with the aim of collecting a large and stylistically varied corpus as the basis for a systematic description of spoken and written English. The other was the advent of computers which made it possible to store, scan and classify large masses of material. The first machine-readable corpus was compiled by Nelson Francis and Henry Kučera at Brown University in the early 1960s. It was soon followed by others, such as the Lancaster-Oslo/Bergen (LOB) Corpus, which used the same format as the Brown Corpus and made it possible to compare different varieties of English. The corpora were rather small by today's standards – just a million words. G.Leech (1991:10) referred to them as the first generation corpora. The second generation corpora, according Leech, were much bigger and benefitted from the newer technology – KDEM character recognition devices which saved the compilers from a great deal of manual input and enabled them to collect big amounts of text quickly. The second generation corpora are represented by John Sinclair's Birmingham Collection of English Texts (the Cobuild project), the Longman/Lancaster English Language Corpus, the British National Corpus (BNC), the International Corpus of English (ICE), etc.

The third generation corpora can be measured in hundreds of millions of words, many of them are in commercial hands, using the technologies of computer text processing (for more information on text corpora see: O'Keefe, et al 2007: 284-296).

The importance of corpora has not always been as widely accepted as it is nowadays. When in the early 1960s Nelson Francis was asked what he was up to at the time, and Francis replied that he had a grant to compile a computerised corpus of English, he was asked "Why in the world are you doing that?" Francis replied that he wanted to uncover the true facts of English grammar. Then a person who asked him this looked at him in amazement and exclaimed:

"That is a complete waste of your time and government's money. You are a native speaker of English, in 10 minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text" (Francis 1982: 7-8).

Such a viewpoint is not at all surprising as the "dominant source of data in the investigation of linguistic theory has been the introspective powers of individual linguists, supplemented by questions asked of native speakers concerning the grammaticality judgements of 'linguistically interesting' sentences.

"The prevalent linguistic fashions of the early 1960s were hardly favourable to any enterprise that included examination and analysis of actual language data. The goal then was "to capture", to use the favourite verb of that age, various profound generalizations about the competence of an ideal speaker-listener who, we are instructed, knew his or her language perfectly, had no memory limitations, including demands of style or effective communication; all of this inquiry was to be pursued with the ultimate aim, achieved only perhaps in the following millennium, of discovering the basis of a universal grammar by the application of superior reasoning. Collecting empirical data was thus not considered a worthwhile enterprise, as it was believed that a native speaker of English could provide the linguist in five minutes with a much greater amount of useful information than even a corpus of billion words could. Henry Kučera admitted that he had in his files a letter from a very well-known linguist of those days who, with something much less than good taste, paraphrased the well-known saying of Hermann Goering: "Whenever I hear the word computer, I reach for my gun". There were many members of the humanistic world in various academic institutions, including Brown University, who had a predictable fear of the new "calculating machines" and little more than contempt for those who dared to commit the treason

of joining the scientists' camp of vacuum tubes, relays and binary numbers" (Kučera 1991: 402-403).

There was a time when members of the linguistic corps regarded corpus as a corpse.

History, as the last several years so amply demonstrated, has the unpleasant habit of not being particularly kind to self-righteous prophets.

In the fifty years since 1961, Corpus Linguistics has gradually extended its scope and influence, so that, as far as natural language processing is concerned, it has almost become a mainstream in itself. It has not revived the American structural linguist's claim of the all-sufficient corpus, but the value of the corpus as a source of systematically retrievable data, and as a testbed for linguistic hypotheses, has become widely recognized and exploited. More important, perhaps, has been the discovery that the computer corpus offers a new methodology for building robust natural language processing systems.

Best-known corpora

The Birmingham Collection of English Texts
Compiled in collaboration with Collins Publishers by a research team under the direction of John Sinclair at the Research and Development Unit for English Studies, University of Birmingham.

The British National Corpus
100 million words of written (90 mln) and spoken (10 mln) English.

The Brown Corpus
Compiled by W.Nelson Francis and henry Kučera, Brown University, Providence, RI. It contains one million words of American English texts printed in 1961.

The Helsinki Corpus of English Texts: Diachronic and Dialectal

Compiled by a research team led by M. Rissanen, O. Ihalainen and M. Kyto at the Department of English, University of Helsinki. The diachronic corpus contains 1.6 million of British English texts from 850 – 1720; the dialectal includes over a million words of contemporary British dialects.

The International Corpus of English (ICE)

The ICE began in 1990 with the primary aim of collecting material for comparative studies of English worldwide. Compiled by national groups (including Australia, Canada, East Africa, India, Jamaica, New Zealand, Nigeria, Philippines, UK, USA).

The Lancaster/IBM Spoken English Corpus (SEC)

Compiled at the Unit for the Computer Research on the English language (UCREL), University of Lancaster, and the IBM UK Scientific Centre, Winchester. It contains 52,000 words of spoken (broadcast) British English.

The Lancaster-Oslo/Bergen Corpus (LOB)

Compiled and computerized by research teams at Lancaster (G. Leech), Oslo (S. Johansson) and Bergen (K.Hofland). It is modelled on the Brown Corpus and contains one million words of British English texts printed in 1961.

The London-Lund Corpus of Spoken English (LLC)

The spoken part of the Survey of English Usage Corpus, computerized at the Survey of Spoken English, Lund University under the direction of J. Svartvik. It consists of 500,000 words of spoken British English recorded from 1953 to 1987.

For more information on corpora see: A. O'Keefe et al. 2007.

Discussion and research points

Discuss the issue of the status of corpus linguistics. Is corpus linguistics a methodology or a theory?
Comment on the difference between corpus-based and corpus-driven approaches to language study.

Further reading
For the discussion of the status of corpus linguistics see: McEnery and Wilson 1997; McEnery et al. 2006; Marcinkevičienė 2000.

2 CHOMSKY CRITICISES CORPUS LINGUISTICS

Corpora (though not always called that) were widely used in traditional linguistics: the great dictionaries of the 18th century (Samuel Johnson‘s dictionary and the OED) were compiled on the basis of large collections of words, the grammars were also constructed using authentic language data (Poutsma and Kruisinga provided copious illustrative examples in their grammars), other language documenters who work in the field of oral histories or other texts also used similar methods.

Chomsky in a series of publications (1957, 1965) managed to change the direction of linguistics away from empiricism towards rationalism. (Rationalism is an approach to a subject – in our case linguistics - which is based on introspection rather than external data analysis. Empiricism is an approach to a subject – in our case linguistics – which is based on the analysis of external data, such as texts and corpora). Chomsky was and still is an enormously influential figure in linguistics. Pinker points out (1994:23) that Chomsky “is among the ten most-cited writers in all of the humanities (beating out Hegel and Cicero and trailing only Marx, Lenin, Shakespeare, the Bible, Aristotle, Plato, and Freud) and the only living member of the top ten.“

The dispute between rationalism and empiricism concerns the extent to which we are dependent upon sense experience in our effort to gain knowledge (Stanford Encyclopedia of Philosophy). Rationalists claim that our concepts and knowledge can be gained independently of sense experience. In language a rationalist theory is a theory based on artificial behavioural data, and conscious introspective judgements. Rationalist theories are based on the development of a theory of mind in the case of linguistics, and have as a fundamental goal cognitive plausibility. The aim is to develop a theory of language that not only emulates the external effects of human language processing, but actively seeks to make the claim that it represents how the processing is

actually undertaken. On the other hand, empiricists claim that sense experience is the main source of all our concepts and knowledge. An empiricist approach to language is dominated by the observation of naturally occurring data, typically through the medium of the corpus. In this case, we may decide to determine whether sentence x is a valid sentence of language y by looking in a corpus of the language in question, and gathering evidence for the grammaticality, or otherwise, of the sentence.

There are advantages and disadvantages to both approaches, but for the moment we will use this characterisation of empiricism and rationalism without exploring the concepts further.

Chomsky suggested that corpus investigations address performance rather than competence, which, according to Chomsky should be the linguist‘s main concern. Competence is best described as our tacit, internalised knowledge of a language. Performance, on the other hand, is external evidence of language competence, and its usage on particular occasions when, crucially, factors other than our linguistic competence may affect its form. It is competence which both explains and characterises a speaker’s knowledge of the language. Performance, it was argued, is a poor mirror of competence. Performance may be influenced by factors other than our competence. For instance, factors as diverse as short-term memory limitations and whether or not we have been drinking can alter how we speak on any particular occasion. (see: McEnery et al. 1997)

Another of Chomsky‘s criticisms was connected with the fact that a corpus is finite while language is infinite. The assumption that the sentences of a natural language can be collected and enumerated, just like blades of grass on a lawn, if a linguist is patient and industrious enough, was connected with the view held by some of the early corpus linguists who considered the corpus as the sole source of evidence in the formation of linguistic theory. Such a view was very attractive as it allowed to set linguistics up alongside other empirical sciences and make language description more objective. Unfortunately, this assumption was false and, as it is well known, the number of sentences in a natural language is infinite. A corpus can never be the sole explicandum of natural language (see Leech 1991:8).

Chomsky also argued that corpora would always be 'skewed'. Some sentences are in the corpus because they are frequent constructions, some by sheer chance. To quote Chomsky (1958:159) on the matter:

Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description (based upon it) would be no more than a mere list.

This is an accurate observation by Chomsky. Corpora are partial in the sense that they are incomplete. They will contain some, but not all of the valid sentences of a natural language. They are also partial in the sense that they are skewed, because frequency of a feature in the language is a significant determiner of inclusion. As Chomsky himself stated so amusingly, the sentence *I live in New York* is fundamentally more likely than *I live in Dayton Ohio* purely by virtue of the fact that there are more people likely to say the former than the latter. This partially was seen by Chomsky as a major failing of early linguistics.

One more criticism made by Chomsky is connected with the corpus methodology as such. Why bother waiting for the sentences of a language to enumerate themselves, when by the process of introspection we can delve into our minds and examine our own linguistic competence? The corpus research is slow, limited and the corpus had cast the linguist in a somewhat passive, and often frustrating mode. Fillmore (1992:35) comments most amusingly on this. He satirises the corpus linguist thus:

He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech.

Fillmore's depiction of the corpus linguist is undoubtedly ironic and exaggerated. Simlarly, he ridicules the so-called armchair linguist who: provides a caricature of the armchair linguist as well: But the real question is: why should we look through a corpus of millions of words when we can get examples via introspection, consulting native speakers?

Fillmore (1992:35) also ridicules the so-called armchair linguist who:

... sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!" grabs his pencil and writes sth. down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

Fillmore's idea is to „marry" the two types of linguists, "because the two kinds need each other"(1992:39).

Chomsky's criticisms did not stop the development of corpus linguistics, his critiques were not invalidated and they helped the corpus linguistics of the day improve.

Even if we assume a performance-competence distinction (which we might not), performance is still an inherently valid object of study. Entire fields of science and research use exclusively or almost exclusively observational data: astronomy, archeology, paleontology, biology, etc. In these fields we observe, build models, make predictions, and collect more observational data. Naturally-occurring data can be collected, studied, analysed, commented and referred to. Corpus-based observations are more verifiable than introspectively based statements.

Frequency lists compiled objectively from corpora have shown that human intuition about language is very specific and far from being a reliable source. Word frequency is also a good reason to use very large and well-balanced corpora. Corpora nowadays are collected in extremely systematic and controlled ways.

The finite-infinite is not a big issue, since in many other fields we also have an infinite number of possible examples, but it does not stop us from studying them (cf. an infinite number of possible songs does not stop us from studying music). It is true that we cannot expect that a corpus will ever cover every possible utterance in a language, but a a big enough corpus (such as

a 100 million word British National Corpus) will provide a lot of utterances one is likely to encounter in language.

Despite Chomsky‘s critique, the development of corpus linguistics did not stop and today we corpus linguistics is mainstream linguistics.

Discussion and research points

Which of the critiques were particularly valid and helped corpus linguistics to improve?

Further reading

For a more detailed discussion read McEnery and Wilson 1997: 5-13.

3 WHY USE CORPORA?

The use of corpora nowadays is no longer an activity interesting only to a small group of linguistis – corpus linguistics has firmly established itself in mainstream linguistics and is taken for granted. There is every reason to believe that corpus linguistics is going to develop even further and impact every aspect of the way languages are taught, learnt and researched. What can corpora offer to language research, learning and teaching?

- Authenticity
- Objectivity
- Verifiability
- Exposure to large amounts of language
- New insights into language studies
- Enhance learner motivation

Authenticity. The key notion in the field of corpus work is that of authenticity[1]. It is certainly reasonable to take a look at real manifestations of language when discussing linguistic problems. When you examine authentic texts you will often be surprised at what you actually find. There is no reason or motivation to invent an example when you are knee-deep in actual instances. "One does not study all of botany by artificial flowers" (Sinclair 1991:24).

Objectivity. When you look at a corpus, you get a more objective picture, since there is no prior selection of data. Paper slips could provide useful information on features that struck the excerpter as interesting, odd, but they are not necessarily the most typical examples. They may be idiosyncrasies of various authors. As Jespersen writes (1995: 213):

I am above all an observer; I quite simply cannot help making linguistic observations. In conversations at home and abroad, in railway compartments, when passing people in streets and on roads, I am constantly noticing oddities of pronunciation, forms and sentence constructions.

Most of the reference books, grammars and dictionaries are also only secondary sources: they present somebody's selection, interpretation of the primary facts, while the greatest advantage of corpora is the authenticity of the language. There is no prior selection – we have the language the way it is used in reality.

Empirical research has shown that the structures many current textbooks teach for certain functions are either never used or used infrequently while quite unexpected structures are the ones that actually occur. In a study of the language of meetings, for example, Williams (1988) finds that many structures for functions taught by business English texts were almost never used in recorded transcripts of business meetings. The structures actually used resembled lexical phrases rather than traditional sentences: they were prefabricated chunks, seldom complete sentences, and were almost always sequences as part of discourse. The structures taught, however, were just the contrary: they were complete sentences, which were not sequenced or considered in combination with other utterances.

[1] Although the term *authenticity* is a controversial concept in linguistics, especially language teaching, and may mean different things to different people, in the context of corpus linguistics *authentic texts* are defined as those that are used for a genuine communicative purpose rather than written specially for teaching purposes.

For example, learners of English were taught **to disagree with sb**. saying *I disagree with you.* Real data, McCarthy argues (1998:19), show speech acts to be far more indirect and subtle in their unfolding. In the CANCODE (Cambridge and Nottingham Corpus of Discourse English), a five million word corpus of spoken English, there were only eight occasions where someone says *I disagree,* and none where *with you* follows. All eight occurrences have some sort of modification which suggests a reluctance on the part of the speaker to utter such a bald statement; these include *I just disagree, I beg to disagree, you see now I do disagree, I'm bound to disagree.* Where the verb form *disagree* occurs, the contexts mostly either 'report' disagreement, or disagree with ideas and propositions, rather than people.

It's possible to study real language with corpora.

Verifiability. Verifiability is a normal requirement in scientific research, thereore, the science of language – linguistics -- (which is often claimed to be the scientific study of language) should not be exempt from this standard mode of research procedure (cf. Leech 1991:112).

New insights into language. Corpus Linguistics is associated with a new view of language. Sinclair noted (1991:1) that *traditionally "linguistics has been limited to what a single individual could experience and remember… Starved of adequate data, linguistics languished – indeed it became totally introverted. It became fashionable to look inwards to the mind rather than outwards to society. Intuition was the key, and similarity of language structure to various formal models was emphasized. The communicative role of language was hardly referred to…. Students of linguistics over many years have been urged to rely heavily on their intuition and to prefer their intuitions to actual text where there is some discrepancy. Their study has, therefore, been more about intuition than about language".*

Corpus linguists do not deny the role of intuition in language research, but the descriptions of language based on introspection and intuition differ considerably from those based on evidence from the corpora. It led Sinclair to suppose that "human intuition about language is highly specific and not at all a good guide to what actually happens when the same people actually use the language" (1991: 4).

Many subtle observations have been made using corpora.

Corpora may help learners discover new meanings of the words they already know:

wa heavy, paintless oak door, then into a room **dimly** lit by a window that looked on to the covered way. The
heavily insulated against all outside noise and **dimly** lit by subdued yellow lighting. There were three display
spread well back from the front door, a long **dimly** lit store that sold everything from flour and salt to patent It
brought me back too precipitately into the small, **dimly** lit world of the hotel bedroom. Masha reached to answer
I admit, as it is now night-time. Some streets are **dimly** lit by smoking torches, but the houses have only the shado
And for a while time stood still. " Hello? " Maggie **dimly** became aware that she was being spoken to. " Hello? Are
ed a foot on the first stone step. She was only **dimly** aware of the approach of the two boys who were walking
lovely sensations would begin again. She was also **dimly** aware that they had passed the point of no return -- now sh
was trying to wrestle with a new idea. It was very **dimly** aware that it needed a new type of thought. There had been
ed up in the half light into the eyes of someone he **dimly** recognized. Then he saw that he was looking at the puz
Earth-digested, come to dust. Someone, I thought **dimly** , was waiting to see if I moved: and if I moved there wo
cause our experience falls far short of what we **dimly** perceive in the pages of Scripture. Much as we love our ut
John Makepeace's Hooke Park College. I can **dimly** recall the challenges of testing a router photograph, and ivi
through l ove... Here speaks what women have **dimly** felt and uncertainly expressed. 10 Critics who wrote rev
rather than any kind of weird visual intuition. I **dimly** understood that by holding out to me this realm of materia
ingthat she did not love him, nor ever could. She **dimly** perceived the terrible conflict that went on in Johnny's min
took on words. " Oh Benny, Benny, Benny...! " I **dimly** discerned a large dog spread over the knees of four small ch
reading had fully revealed a fact which she had **dimly** apprehended before but lacked the courage to confront until
realize that he is part of Tite's plans -- as I now **dimly** begin to perceive -- to revive the flagging fortunes of the
her hand was pressed violently to her mouth, she **dimly** realised, and she was biting her knuckle so hard that she'd a
still practise today, though their meaning is only **dimly** remembered. There is the old English custom of dancing ar
wake Doyle. I thought you'd want to hear this. " **Dimly** he saw that she was holding something, but he did she soke
agnhat was he about to do to her? she wondered **dimly** . For suddenly she knew he was about to do som such close
scrutiny in a public place the two of them **dimly** sensed at one tissue-thin layer within the oldest parts

Source:BNC

J. Sinclair has developed his own understanding of meaning. According to him, "every distinct sense of a word is associated with a distinction in form" (1996:89). This principle was applied in practice: in compiling the Cobuild dictionary, where every sense of a word is presented in the most typical pattern, structure or model. Thus, Sinclair noted that the word *glare* (n), which according to most dictionaries, has two distinct meanings, e.g. (LDELC 1992: 551) 1. an angry look or stare; 2. a hard unpleasant effect given by a strong light, is also associated with two distinct patterns: when the word is used in the sense of 'angry look' it is used with the indefinite article, while the 'bright light' meaning – with the definite article:

of tea and sat down. Gradually my eyes got used to **the glare** and I was able to make sense of my surroundings.
ness of her extremities. Her eyes were closed against **the glare** of lights overhead, but still their dazzle came through
What do you mean? " Jack looked at Tina's face in **the glare** of the cars' headlights. He remembered how strong
At the major shows, winners are often exposed to **the glare** of television lights, and expected to parade for the
position well. A couple of silhouettes emerged from **the glare** of light. Jezrael stopped on the top step, yawning and
opened the door. For a moment she was blinded by **the glare** of headlights before she could pick out the white mini

of my surroundings I was temporarily dazzled in **the glare** of sunlight. My glance fell upon a stricken shrub and

be on the receiving end of **a** truly hostile **glare** from those eyes. " The worst thing that this
the door-knob, her face twisted in **a** ferocious **glare.**'What are you waiting for?"
Silas said gruffly, while directing **a** cool **glare** towards his uncle. The older man stood up.
suddenly became **a** „don't-you-dare" **glare.** „Your relationship with.."
He shot a suspicious glare towards me, **a glare** of hostility.
grinning after she gave him **a** warning **glare**. It was unfair, he said
the sheriff gave her **a** withering **glare** and leaned forward. I think we can

Source:BNC

Similarly, the word *budge* in the LDELC (1992: 151) is defined as'to (cause to) move a little'. If we look at the corpus data, however, we will see that the language does not talk about moving. In fact, all the occurrences are either grammatically or lexically negative: (more on that below: Understanding of meaning).

Enhance learner motivation

Another argument in favour of using corpora is what has been noted by Leech (1997: 2), who wrote that "corpus as an information source fits in very well with the dominant trend in university teaching philosophy over the past 20 years, which is the trend from *teaching as imparting knowledge to teaching as mediated learning".* In this context, there is no longer a gulf between research and teaching, since the student is placed in a position similar to that of a researcher, investigating and imaginatively making sense of the data available through observation of the corpus.

McCarthy (1998: 67-68) argues that the traditional 3 P's approach methodology – Presentation – Practice – Production should be supplemented by the 3 I's method: Illustration – Interaction – Induction. Illustration means looking at real data where possible, Interaction means talking among learners and teachers about language, showing and forming views, hypotheses. Induction means drawing conclusions about certain linguistic phenomena and their use. In this way the students "discover" language themselves, and this "discovery" feeling has a huge motivating effect on the learner. What is more, the 3 P's approach is congenial for students of all levels as it is a bottom-up study of the language that requires very little learned expertise. The students need

only basic reference categories of linguistic description as the starting point of their research is the observation and interpretation of language evidence. This observation then leads to the formulation of a hypothesis to account for the data observed and the generalizations made on the basis of the repeated patterns in the concordances.

A corpus, Leech argues (1997:3) "is itself a rich resource of authentic data containing structures, patterns and predictable features that are waiting to be unlocked by the human intelligence". In this respect, a corpus-based and corpus-driven activity could be compared to what happens in the scientific laboratory, or in fieldwork. A student working on a relatively small corpus assignment comes up with his/her own original observations and discoveries which have probably never been brought to notice before, and this proves extremely rewarding for the student.

This is a student-centred paradigm of 'discovery learning' Johns (1991) claimed that "the task of the learner is to discover the foreign language, and the task of the language teacher is to provide a context in which the learner can develop strategies for discovery – strategies through which he can learn how to learn".

Despite the cost of making and using concordances, their potential value in foreign language teaching is considerable for at least 2 reasons:

- The first is the Hawthorne effect – a well-known principle according to which any new tool or method tends to stimulate the actors of a pedagogic act and to improve the results more than the mere continuance of trite procedures.

- The second reason is less superficial: it has to do with the laws of memory. There is now evidence enough in support of the thesis that memory is conditioned by an active cognition of the past. In other words, we may safely assume that recognising and recalling any fragment of one's past – a fact, an emotion, or even a word – are in the long run much easier if the mind, at the very moment of the input, has actively associated the fragment with circumstances of that input. What sort of circumstances is the mind submitted to? Among others, to those caused by the learner's willful mental activity carried out while trying to get a grip on all the relevant holds to be found in the textual environment of any problem.

Third, it is agreed that exposure to large amounts of language nurtures a "feel of language", develops an understanding of what is natural in a language. If you want to learn foreign words you will need to learn them in 'living' contexts: newspapers, magazines, books, the radio - the

more contexts, the better. It is only by observing a word in many 'living' contexts that we can master its meaning. We learn our native language so thoroughly and accurately because during our lifetime we are exposed to many different linguistic contexts, different uses and meanings of words.

The computer corpus has been described by Barnbrook as " a tireless native-speaker informant, with rather greater potential knowledge of the language than the average native speaker" (1996: 140).

Are there any potential hazards or disadvantages in using corpora in ELT?

There are a couple of things that have to be taken into account when talking about using corpora in ELT.

Chomsky's criticism represented an extreme argument against using corpora in linguistic research and language teaching. Nowadays the situation is different and nobody categorically denies the importance of corpora and corpora evidence in language teaching. There are, however, some potential hazards embedded in overdependency on corpora data.

A corpus is not an infallible source of all linguistic information about language – there can be some unique instances, which have no statistical significance and which do not represent 'real' language. On the other hand, corpora users should not think that if some linguistic item cannot be found in a corpus, it does not exist at all. This overdependence and overreliance upon corpora can be an inhibiting dogma.

Another danger lies in an attempt to replace a laborious hands-on analysis by a rapid automatic processing. A careful manual analysis based on empirical data and intuition cannot be dispensed with in linguistics.

There are also some reservations expressed regarding the use of corpora in the classroom (see: Widdowson 2000). He argues that corpus linguistics as the quantitative analysis of text by

computer reveals facts about actual language behaviour which are not, or at least not immediately, accessible to intuition (2000: 6). He distinguishes three types of data: third-person observations, second-person elicitations, and first-person intuitions. The excerpt below, cited from Widdowson (2000), focuses on the limitations of corpus linguistics.

"There are frequencies of occurrence of words and regular patterns of collocational co-occurrence, which users are unaware of, though they must be part of their competence in a procedural sense since they would not otherwise be attested. They are third person observed data ('When do they use the word X?) which are different from the first person data of introspection (When do I use the word X?) and the second person data of elicitation (When do you use the word X?). Corpus analysis reveals textual facts, fascinating profiles of produced language, and its concordances are always springing surprises. They do reveal a reality about language usage which was hitherto not evident to its users. But this achievement of the corpus analysis at the same time necessarily defines its limitations. For one thing, since what is revealed is contrary to intuition then it cannot represent the reality of first person awareness. We get third person facts of what people do, but not the facts of what people know, not what they think they do; they do come from the perspective of the observer looking on, not the introspective of the observer. In ethnomethodogical terms we do not get member categories of description. Furthermore, it can only be one aspect of what they do that is captured by such quantitative analysis. For obviously enough, the computer can only cope with the material products of what people do when they use language. It can only analyse the textual traces of the processes whereby meaning is achieved: it cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted. It cannot produce ethnographic descriptions of language use. In reference to Hyme's components of communicative competence, we can say that corpus analysis deals with the textually attested, but not with the encoded possible, nor the contextually appropriate.

To point out these rather obvious limitations is not to undervalue corpus analysis but to define more clearly where its value lies. What it can do is reveal the properties of text, and that is impressive enough. But it is necessarily only a partial account of real language. For there are certain aspects of linguistic reality that it cannot reveal at all. In this respect, the linguistics of the attested is just as partial as the linguistics of the possible".

What is not taken into account is the pedagogic perspective, the contextual conditions that have to be met in the classroom for language to be a reality for the learners.

Discussion and research points

What cannot corpora tell us?

Further reading

McEnery et al. 2006: 120

Widdowson 2000: 3-25

Stubbs 2001: 149-172

4 CORPUS CREATION

The issues in corpus design and compilation are directly related to the validity and reliability of the research based on a particular corpus (Kennedy 1998: 60). Sinclair (1991: 13) claimed that "the decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently. The results are only as good as the corpus".

The issues to be considered include the size of a corpus, the type of a corpus (sample, monitor, general, special), the types of texts that should go into a corpus and the size of text samples.

Nowadays there are many ready-made corpora which can be accessed free, for a symbolic fee (when used for research purposes), or purchased. Many researchers find it necessary to compile their own corpora to address a particular research question.

Corpus design

Corpus design outline is a simple matter: corpora builders should decide upon the type of corpus, the size of it and then choose the texts for inclusion. The whole process of a corpus building is unfortunately much more complicated.

One of the biggest problems all corpora builders encounter is copyright.

Getting permissions

This is a very sensitive area of law and, although many publishers and rights holders understand why their texts are wanted, the fear of piracy and exploitation of materials for profit put additional strain on corpora builders. The issue of getting permissions has often been addressed by corpora linguists, unfortunately there is yet no solution to the problem of copyright in corpora building and "the labour of keeping a large corpus in good legal health is enormous" (Sinclair 1991:15). Corpora builders should always seek permission to include a text in a corpus they are building and using copyrighted material without the permission of the copyright holders would be a grave violation of copyright and may get corpus builders into trouble.

The whole business of getting permissions is further aggreviated by the variation in copyright law – different countries have different laws. Copyright problems should be solved internationally. Until a satisfactory solution is found, corpus projects should be designed with this in mind as a potential shadow over the enterprise (Sinclair 1991:15).

Discussion and research points.

Research the copyright laws of Lithuania and find out what restrictions govern the production of an electronic copy of copyrighted material for research purposes. Contact one or more publishers to find out about their policy and practice in assisting researchers to build corpora.

Further reading

McEnery et al. 2006: 77-79

Design

The design of a corpus is dependent upon the type of a corpus and purpose for which the corpus is to be used. The builder of a corpus should have an idea of the kind of analyses that could be undertaken. If a corpus is compiled in order to investigate some linguistic features that characterise a particular type of text, then a compiler will build a specialist corpus, if, however, a corpus is meant for the study of a particular language in general, then a collection of different types of texts will be needed.

TYPES OF CORPORA

The purpose of the compilation influences the design, size and type of a corpus. There are different types of corpora: sample, monitor, general, special, spoken, written, learner, etc.

SAMPLE CORPORA

A sample corpus is a static collection of texts (samples of texts) selected according to some strict criteria and intended to be typical of the whole language or an aspect of the language at a particular period of time. The first-generation corpora were like this. Thus, the Brown corpus is a sample of American printed English of the year 1961, while its British counterpart LOB (Lancaster-Oslo-Bergen) corpus represents British English of the year 1961. Their validity lies in the clarity of the internal structure and the criteria of the text selection. Both Brown and LOB corpora consist of a large number (500) short extracts (2000 words), randomly selected from within 15 genres of printed texts. With these dimensions of extracts, and their relationships - fairly regular and known - a great amount of useful information can be extracted with ease from these corpora. Biber argued (1990) that text samples of 2000-5000 words are big enough to represent their text categories. Such corpora have their our limitations and are inappropriate for the study of discourse, infrequent words, text cohesion, etc.

MONITOR CORPORA

Monitor corpora are text corpora that represent a dynamic, changing picture of a language. Such a dynamic collection of texts is constantly growing and changing with the addition of new text samples. Texts are collected over a period of time. Sinclair (1991:25) described the notion of a monitor corpus as holding the state of a language:

> It is now possible to create a new kind of corpus, one which has no final extent because, like the language itself, it keeps on developing. Most of the material will come in from machine-readable sources, and will be examined for the purposes of making routine records. Gradually, it will get too large for any practicable handling, and will be effectively discarded. The focus of attention will be on what information can be gleaned from the text as it passes through a set of filters which will be designed to reflect the concerns of researchers.

A monitor corpus will have a large and up-to-date selection of current English available; it will have a historical dimension, and it will have a comprehensive word list because of its elaborate record-keeping.

GENERAL CORPORA

They are assembled to serve as a reference base for unspecified linguistic research (Kennedy 1998:19). The linguists may use them to answer particular questions about the vocabulary, grammar or discourse of a language. To study features of the language in general, independently of the styles of particular types of text, you need a general corpus, a collection of texts of as many different types as possible.

Size

The issue of the size of a corpus is closely related to the issue of representativeness and balance – a corpus has to be big to be representative. This claim is based on the pattern of word occurrence in texts, first pointed out by Zipf (1935). There is a huge imbalance in the frequency of the words. Nowadays with a very large collections of texts stored and searched by computers it is possible to determine the frequencies of words by using fairly trivial computer programs (see Table 1 in the Appendix). According to Sinclair (1991: 18), most of any text is taken up by words like *of*, *is*, *up*, and *by*; rather less by *like*, *taken*, *any*, and *most*; still less by *words*, less again by *text* (the example words are the first ten words of this sentence). About half of the vocabulary of a text - even a very long text - consists of words that have occurred once only in that text.

As a general rule, the bigger a corpus is the richer and more interesting the output from a concordancing program will be, and the more likely to represent accurately features of the language.

On the other hand, as Leech argues (1991:10-12), to focus merely on size, would be naive - for four reasons.

Firstly, a collection of machine-readable text does not make a corpus (cf. the difference between a corpus and an archive). The third generation corpora have been collected very often according to what sources of material were made available and therefore are haphazard collections of texts. There are initiatives to assemble archives comparable to in scope to that of national libraries. Such collections will be archives and will differ from carefully designed corpora meant to perform a particular 'representative' function.

Secondly, all very large collections of texts have been in the medium of written language - we do not have reliable speech recognition devices, which could facilitate the whole process of spoken data collection. "Until speech-recognition devices have developed the automatic input of spoken language to the level of present OCR (optical character-recognition) devices for written language, the collection of spoken discourse on the same scale as written language, will remain a dream of the future" (Leech 1991: 11)..

Thirdly, as Leech claims (1991: 11), technology advances quickly, while human institutions evolve slowly. This applies to the legal systems and copyright issues discussed above, in particular. Copyright holders are unlikely to grant permissions freely and willingly.

Fourthly, as is well known, "hardware technology advances by leaps and bounds, software technology lags like a crawling snail behind it" (Leech 1991: 12). A corpus is a collection of texts which is made useful for a researcher only with the help of software. Although some good concordancing programs are available nowadays, more sophisticated search and retrieval packages are needed to make corpus analysis linguistically more interesting.

In practice the size of your corpus is likely to be limited by technical constraints. An in-memory concordancer imposes an absolute restriction on the quantity of text which can be analyzed at one time. With other types of software there may be no fixed maximum, but if your corpus keeps expanding it will reach a size where it takes too long for the program to scan it, or it occupies so much space as to be unwieldy, the output may become unmanageable. .

Spoken and written language

Another issue that should be considered is whether a general corpus should include spoken language.

Many language scholars and teachers believe that the spoken form of the language is a better guide to the fundamental organization of the language than the written form and that it should occupy as large a portion of a general corpus as is possible. Spoken language could provide: a rich source of data for all those interested in the nature of spoken language and language in general, as spoken language is primary and all the changes start there. Besides, spoken language is not that well researched and most of current understandings of language rely too much on written language. Spoken language can also prove valuable for the studies of differences between speech and writing and contribute to the understanding of how to facilitate the learner's transition from accomplished speaker to accomplished writer. It can constitute a source of information for those involved in second language learning and teaching and in the larger time frame could provide an invaluable source of information on how the language was spoken colloquially at a certain period of its development.

Many scholars agree that an ideal general corpus would contain a high proportion of transcribed spoken language. Unfortunately this is not so easy to achieve in practice.

Firstly, transcribing recorded speech is a very tedious and time-consuming process and we cannot speed this process up until we have reliable speech recognition devices.

Secondly, spoken discourse it is quite difficult to obtain. Interviews, debates and discussions on the radio and TV constitute only a small portion of the uses of spoken language. Everyday conversation, on the other hand, is very difficult to record. Even if participants give their permission, there are few situations in which they will speak naturally and spontaneously in the presence of a microphone, recording people without their permission is an unjustifiable invasion of privacy.

Film scripts, drama texts, etc., are of little value in a general corpus, because they are 'considered' language, written to simulate speech in artificial settings, since they do not reflect

natural conversation, which for many people is the typical example of the spoken language.

On the whole, the spoken and written forms of a language are so different that any corpus which contains examples of both has to be balanced in this respect. Just a small amount of spoken language in an otherwise written corpus might yield very peculiar results.

Discussion and research points

Discuss the approach to spoken corpus design used by the British National Corpus project.

Further reading

Crowdy 1993: 259-265

The composition of the BNC:

A1 WRITTEN LANGUAGE COMPONENT: INFORMATIVE
PRIMARY SUBJECT FIELD (or DOMAINS)
- Natural and pure science
- Social science
- Commerce and finance
- Belief and thought
- Biography
- Applied science
- World affairs
- Arts
- Leisure

GENRE
- Books
- Periodicals
- Written to be spoken
- Miscellaneous (published)
- Miscellaneous (unpublished)

LEVEL
- Specialist
- Lay
- Popular

DATE: 1975-PRESENT

A2 WRITTEN COMPONENT: IMAGINATIVE
GENRE

Narrative fiction
Essay
Playscript
Poetry

LEVEL

Literary
Middle
Popular

DATE: 1950-PRESENT

B1 SPOKEN COMPONENT: DEMOGRAPHIC SAMPLING

Selection of 100-200 "subjects" who are native speakers of British English, sampled across:

- Region
- Age
- Occupation
- Educational/social background

B2 SPOKEN COMPONENT: LAYERED SAMPLING

Sampling across a range of discourse types:

Dialogue

Private

Face-to-face: structured
Face-to-face: unstructured
Distanced
Classroom interaction

Public

Broadcast discussion/debate
Legal proceedings

Monologue

Commentaries
Lectures/speeches
Demonstrations
Sermons

Source: Leech 1992: 5-6

Summing up, compilers of general corpora might use the following as guidelines:

1. Texts should be authentic. The main advantage of a corpus is that it gives users direct access to genuine, authentic language, not artificial texts, concocted for the use of learners.

2. Use Contemporary texts rather than ancient literature, old enough to contain some archaic, unusual linguistic forms and patterns.
3. Beware of dialects, which may include odd forms and spellings.
4. Stick to prose. Verse achieves many of its effects by deliberately violating the normal patterns the language.
5. Include highly technical material only in very small doses.

Discussion and research points
Study and report on the composition of the Corpus of the Contemporary Lithuanian Language.

Further reading
Kennedy 1998; Hunston 2002; Meyer 2002, McEnery 2006 provide further information and discussion of corpus design issues in general.

5 CORPORA AND LEARNER LANGUAGE

Learner corpora are defined as electronic collections of authentic texts produced by foreign or second language learners (Granger 2003). Learner corpora are a recent phenomenon, although, according to Granger (1998:5), learner corpora can be traced back to the Error Analysis era. The early learner corpora differed a great deal from contemporary corpora in that they served as depositories of errors, they were smaller, heterogeneous and not computerised. Current learner corpora are much bigger in size, they are more sophisticated and varied, their design criteria are much stricter and they lend themselves to the analysis of most languages.

The first computerised learner corpora were collected in the 1990s when several learner corpora projects were launched: the Longman Learners' Corpus, the Cambridge Learner Corpus, the Hong Kong University Learner Corpus and the International Corpus of Learner English (ICLE).

The Longman Learners' Corpus contains ten million words of text written by learners of English of different levels of proficiency and from twenty different L1 backgrounds. The texts include

in-class essays written with and without a help of dictionaries, timed examination papers and other types of written assignment. Each essay is coded by L1 background and proficiency level. The corpus is partly error-tagged manually. The corpus offers invaluable information about learners‘ mistakes and is a useful resource for textbook and coursebook writers.

The Cambridge Learner Corpus is a large collection of written texts from learners of English all over the world. The texts are exam papers of learners taking Cambridge ESOL English examinations. The corpus contains over 25 million words and includes over 85 000 scripts from 180 countries (100 different backgrounds). Each paper is coded with information about the student‘s first language, nationality, level of English and age. Over eight million words have been coded for errors.

There are also a number of learner corpora which cover only one L1 background. The HKUST Corpus of Learner English is a ten-million word corpus which contains written essays and examination scripts of Chinese learners of English at the University of Hong Kong. The JEFLL (Japanese EFL Learner) corpus is a one-million corpus containing 10 000 sample essays written by Japanese learners of English. The JPU (Janus Pannonius University) learner corpus contains 400 000 words of essays written by the advanced level Hungarian university students. The USE (Uppsala Student English) corpus contains one –million words of essays written by advanced learners of English at Uppsala University. The Polish English Learner Corpus is a half-million word corpus of written learner language produced by Polish learners of English of different proficiency levels.

The International Corpus of Learner English (ICLE) is the best-known learner corpus which provides a collection of essays written by advanced learners of English (third and fourth year university students) from different native language backgrounds. The International Corpus of Learner English project was launched in 1990 by S. Granger at the University of Louvain in Belgium. The International Corpus of Learner English (Version 2) contains 3.7 million words of EFL writing from learners representing 16 mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana). The main aim of the project was to collect a corpus of objective

data for the description of learner language. The primary goal of ICLE was to investigate the interlanguage of the foreign language learner. The research goals of the ICLE project were twofold. On the one hand the project sought to collect reliable data on learners‘ errors and to compare them cross-linguistically in order to decide whether they are universal or language specific. On the other hand, ICLE aimed to research aspects of foreign-soundedness in non-native essays which are revealed through the uveruse or underuse of words or structures with respect to the target language norm.

Learner corpus research, according to Granger (2009:13), lies at the crossroads between four major disciplines: corpus linguistics, linguistic theory, second language acquisition and foreign language teaching.

Learner corpora can be analysed in many different aspects which describe peculiarities of learner language, i.e. interlanguage (Lauridsen 1996: 55; Granger 2002). Findings from research into learner corpora can be applied in materials design and development (see Kaszubski 1998:172-185), they can translate into classroom practice and inform the teachers of the typical learner error patterns (see Dagneaux et al. 1996; De Cock et al. 1998), underuse and overuse of particular linguistic features (Altenberg 1998; Altenberg et al. 2001, De Cock et al. 1998). At the discourse level, complete essay texts allow to analyse learners‘ discourse competence and their ability to create coherent and cohesive texts. At the sentence level, a corpus can be investigated in terms of specific features of vocabulary or grammar (O‘Keefe et al. 2007).

Learner corpora and Second Language Acquisition

Language acquisition is a mental process, which we can observe only through its product, i.e. the data the learner produces. One of the main problems with the SLA research is a narrow basis of empirical data. Thus, Gass and Selinker (2001:31) pointed out that” it is difficult to know with any degree of certainty whether the results obtained are applicable only to the one or two learners studied, or whether they are indeed characteristic of a wide range of subjects”. Learner corpora can provide a wider empirical basis on which many hypotheses can be tested and the principles that govern the process of learning a foreign language uncovered.

Learner corpora and language teaching

Although learner corpora can provide a great amount of useful information on many aspects of learning and teaching a foreign language, the introduction of corpora in the classroom might mean a tough job of changing attitudes of teachers and learners. Both teachers and students are more used to traditional methods and may sometimes find using corpora in the classroom quite challenging. The problem, as Aijmer writes (2009:1) is to find out the ways to reach students and teachers with information about corpora and what they can do. This on the one hand implies educating teachers and spreading the word about corpora and on the other helping students with the search options, search interface and the analysis of corpus output. Using corpora in the classroom changes the student's role. With a corpus and the appropriate tool kit, "the student can actually test the conventional wisdom of the textbooks and find out what really happens in connected texts. In this way the distinction between teaching and research becomes blurred and irrelevant" (Knowles 1990). Using learner corpora in the classroom is still a very new thing and before it becomes a standard practice it has a long way to go. However, "the exploration of learner corpora by learners themselves will motivate many more learners to reflect on their language use and thus raise language awareness" (Mukherjee and Rohrbach 2006: 228).

The direct exploration of corpora integrated into university courses for learners of English is still a rare phenomenon. Students need to be trained how to use corpora, they would not automatically catch on as corpora do not provide straightforward answers. Students should be familiarized with inductive methods, otherwise they will find corpora boring and difficult. Teachers should introduce students to corpus analysis by specially prepared exercises. This will give students insights into what they can learn from corpora and how they can use corpora. The students can also do various corpus studies in morphology, phraseology, syntax, etc.

Over the last few decades, native English corpora have increasingly been used in EFL materials design. The Collins Cobuild project set this trend and the belief that better descriptions of authentic native English would lead to better EFL tools and indeed, studies which have compared materials based on authentic data with traditional intuition-based materials have found this to be true.

However much of an advance they were, native corpora cannot ensure fully effective EFL learning and teaching, mainly because they contain no indication of the degree of difficulty of words or structures for learners. There is no doubt that the efficiency of EFL tools could be improved if materials designers had access not only to authentic native data but also to authentic learner data, with the native speaker (NS) data giving information about what is typical in English, and non-native speaker (NNS) data highlighting what is difficult for learners in general and for specific groups of learners. As a result, a new generation of EFL tools is beginning to emerge.

Despite many advantages and the great potential of learner corpora, there are some limitations as well. As noted by Nesselhauf (2004: 131), the receptive abilities of learners cannot be investigated, i.e. such questions as, for example, how certain are learners about the acceptability of what they are producing cannot be answered. Besides, if a word or a phrase does not occur in the text produced by the learner, there is no way of finding out whether the learner knows it or not. Therefore, very rare phenomena can only be investigated experimentally.

Discussion and research points

With the learner language corpora (ICLE and LICLE) we may look for the answers to many research questions, including the following taken from Leech (1998: xiv):

- What linguistic features in the target language do the learners in question use significantly more often ("overuse") or less often ("underuse") than native speakers do?
- How far is the target language behaviour of the learners influenced by their native language (NL transfer)?
- In which areas do they tend to use "avoidance strategies", failing to exploit the full range of the target language's expressive possibilities?
- In which areas do they appear to achieve native-like or non-native like performance?
- What (in the order of frequency) are the chief areas of non-native like linguistic performance which learners in country A suffer from and need particular help with?

Further reading

Granger, S. (ed). 1998. *Learner English on Computer*. London: Longman.

Aijmer, K. (ed). 2009. *Corpora and Language Teaching.* Amsterdam/Philadelphia: J. Benjamins.

6 CORPORA AND LANGUAGE RESEARCH. UNDERSTANDING OF MEANING IN CORPUS LINGUISTICS

Every distinct sense of a word is associated with a distinction in form. (J.M. Sinclair).

How can Corpus Linguistics contribute to the understanding of language?

The problem of defining corpus linguistics and whether as a theory or a methodology has been debated from different standpoints. It has been argued that corpus linguistics is not really a domain of research but only a methodological basis for studying language (Leech 1991). However, many linguists working with a corpus now agree that corpus linguistics goes well beyond this purely methodological role. Halliday, for instance, points out that corpus linguistics re-unites the activities of data gathering and theorizing and argues that the potential for quantitative research thus opened up is leading to a qualitative change in our understanding of language. The linguist who has, more than anyone else, opened our eyes to the new types of insights that corpus evidence has to offer is Sinclair. What we are witnessing is the fact that corpus linguistics has become a new research enterprise and a new philosophical approach to the subject, to put it in Leech's words "a new way of thinking about language" (1992: 106).

How does the language create meaning? What are the means by which language creates meaning?

Traditionally, we talk of the basic distinction between grammar and lexis (sometimes the terms *syntax* or *structure* and instead of *lexis* - *semantics* or *vocabulary* are used) . But there is always this basic distinction between patterns of organization and items that fill places in the patterns.

Sinclair argues that "recent research into the features of language corpora give us reason to believe that the fundamental distinction between grammar, on the one hand, and lexis, on the

other hand, is not as fundamental as it is usually held to be; it is worth considering how far, using modern techniques, we can get in describing a language without resorting to such a distinction" (2000:191). This distinction between grammar and lexis, as Sinclai claims, is a very basic model of language and there should be very strong arguments and new evidence to make us reconsider it. According to Sinclair, such a model became so well established because "before the computer age linguists were unable to describe all the complexity of language at once" (2000:192) since they had nothing but their own "five senses, memory and internal awareness, it was difficult to analyse such a complex matter as language" (ibid: 192). Grammatical patterns are easy to observe therefore grammars usually have very elaborate systems, ranks, hierarchies, categories and other systems of description, while lexical patterns are difficult to observe since they are realized through a vocabulary of infrequent words and are not easily discovered. With large corpora and sophisticated software we can work out and describe the recurrent patterns in lexis. Data coming from corpus research will impact and change the way lexical information is presented in dictionaries.

Corpus Linguistics and the understanding of meaning

In Corpus Linguistics, the role of context is crucial: it disambiguates. In continuous discourse, whether written or spoken, true ambiguity occurs rarely, except where a writer or speaker deliberately wants to be ambiguous – for example when punning or telling jokes. A whole battery of given and shared information means that a particular word is unlikely to be ambiguous at the moment of utterance, irrespective of how many different senses for it are recorded in a dictionary.

Meaning is the product of context.

J. Sinclair in identifying and defining the meaning of words takes into account on the one hand their contextual associations and on the other their pragmatic function. A word or expression will be defined, therefore, in terms of the grammatical and collocational patterning it entertains in its context, and the focus of attention will be on the pragmatic implications of its use. Sinclair (1996a) uses the term of *extended unit of meaning* and proposes the following methodological steps to define it:

- identify **collocational profile** (lexical realizations)
- identify **colligational patterns** (lexico-grammatical realizations)
- consider common semantic field (**semantic preference**)
- consider pragmatic realisations (**semantic prosody**)

Collocation is the occurrence of words with no more than four intervening words.

The term collocation was first used by Firth (1957). According to Firth (1968: 181), "collocations of a given word are statements of the habitual or customary places of that word". Firth's notion of collocation is essentially quantitative (see Krishnamurthy 2000: 32). The statistical approach to collocation is accepted by many corpus linguists (see McEnery et al 2006: 82), who argue that collocation refers to the characteristic co-occurrence of patterns of words. The task of determining frequency of co-occurrence of patterns manually is a daunting task, but in the age of the computer the calculation of collocation statistics is a relatively trivial task given suitable software. Computerized corpora and relevant software have freed linguists from overreliance on intuition. Intuition as Krishnamurthy (2000: 32-33) argues is a poor guide to collocation, "because each of us has only a partial knowledge of the language, we have prejudices and preferences, our memory is weak, our imagination is powerful (so we can conceive of possible contexts for the most implausible utterances), and we tend to notice unusual words or structures but often overlook ordinary ones".

Colligation is the co-occurrence of grammatical phenomena, and on the syntagmatic axis our descriptive techniques at present confine us to the co-occurrence of a member of a grammatical class – say a word class- with a word or phrase.

Sinclair refers to colligation as the co-occurrence of grammatical items with a specified node. For instance, he notes that the node *true feelings* has a strong colligation with a possessive adjective (Sinclair 1996: 86). Other kinds of colligation might be a preference for a particular verb tense, negative particles, modal verbs, participles, that-clauses, and so on.

Semantic preference is the restriction of regular co-occurrence to items which share a semantic feature, for example that they are all about say, sport or suffering. Semantic preference is a semantic field a word's collocates predominantly belong to.

Semantic prosody is attitudinal, and on the pragmatic side of the semantics/pragmatics continuum. Semantic prosody was introduced by Louw (1993), but developed by Sinclair (1991, 2004). The term has also been used by Stubbs (1996, 2001), Tognini-Bonelli (2001), Partington (1998, 2004), Philip (2011) and many others. Semantic prosody is a problematic concept, mainly because it has been used to describe such things as connotation, evaluation, appraisal, pragmatic force etc. and is differently understood by different authors. Partington associates semantic prosody with a binary distinction between positive and negative attitudinal meanings. Semantic prosody describes the way in which certain seemingly neutral words can be perceived with positive or negative associations through frequent occurrences with particular collocations. Thus, such verbs as **set in** (rot, decay, ill-will, decadence, infection, prejudice, etc.), **cause** (cancer, crisis, accident, delay, death, damage, trouble, etc.)**, commit** (crime, offences, foul etc.), **rife** (crime, diseases, misery, corruption, speculation, etc.), often have negative semantic prosody, while such words as **impressive** will occur with lexical items such as **dignity, talent, gains, achievement**, etc. will have positive prosody.
Semantic prosody, however, cannot be reduced to a simple positive or negative' evaluation. Sinclair uses the term 'semantic prosody' in a more subtle way, referring not to simple co-occurrence but to consistent discourse function form by a series of co-occurrences: the'unit of meaning' (Hunston 2007: 257).

Sinclair thus examined the word *budge* (2004: 142-147) and showed that the concept of semantic prosody is much more complex than a simple positive or negative evaluation. If we look up the owrd *budge* in a dictionary, we can find the following definition:
*To (cause to) move a little (*Longman Dictionary of Contemporary English)
The point is, as Sinclair claims, that English does not talk much about budging at all, but about not budging. The two examples that follow the definition are indeed both negative, but the entry reads as if the lexicographers had not noticed this primary fact of usage. (*We tried to lift the rock*

but it wouldn't budge/ we could't budge it. (fig.) She wouldn't budge from her opinions. LDCE 1992:151).

Studying the concordances from the corpora (see: concordances from the BNC) it would be difficult to find a context where this word would be semantically positive.

If we pursue the environment of any word, we will get this data on language:

Discussion and research points

Study the article "Corpus Classroom Currency" by E. Tognini Bonelli (2000:205-243).

Study the example of the analysis of the phrase *the naked eye* presented below (J. Sinclair. 1996. The Search for Units of Meaning. *Textus* vol. ix, no. 1, p. 75-106.

Study the example of the analysis of the word *budge* presented below Sinclair, J. 1997. The Lexical Item. In *Contrastive Lexical Semantics.* Weingand, E. (ed). Amsterdam/Philadelphia: J. Benjamins. 1-25.

In this article Sinclair demonstrates the inadequacy of the traditional assumption that a word is an autonomous, meaningful unit and that an autonomous lexicon is composed of such words. He points out that lexical semantics has traditionally been over-interested in paradigmatic relations to the neglect of syntagmatic patterns. Certainly, the latter that is more relevant to any theory that purports to model sentence production. He looked at the negative prosody of 'budge'. Sinclair does not provide a new model of the lexicon, but suggests that thinking of words as lexical items unto themselves may not be the right approach. He concludes that the way to better understanding of words' syntagmatic patterns is through computational analysis of large text corpora. Sinclair, here and elsewhere, has proved the strength of corpus investigations as discovery tools.

Example

The word ***budge*** **i**n English poses a problem for dictionaries

*To (cause to) move a little (*Longman Dictionary of Contemporary English)

The point is that English does not talk much about budging at all, but about not budging. The two examples that follow the definition are indeed both negative, but the entry reads as if the lexicographers had not noticed this primary fact of usage. (*We tried to lift the rock but it wouldn't budge/ we could't budge it. (fig.) She wouldn't budge from her opinions.* LDCE 1992:151).

Studying the concordances from the corpora (see: concordances from the BNC), it would be difficult to find an instance of this word which is semantically positive. Most of the indications of colligation with a negative are to be found to the left of the central, or node word; immediately to the left we find instances of words ending in *n't* and *not* – together making slightly over half the total. Most of the others show the word *to* in this position, and by examining the word previous to that, there is a strong collocation with forms of the lemma *refuse.* Although not a grammatical negative, *refuse* can reasonably be considered as a lexicalisation of the kind of non-positive meaning that characterizes *budge.*

There are, then, just a few remaining instances that do not follow one of the three prominent ways of expressing negativity. One line has a double negative in an extended verbal group, another has *determined not to*, and there is one which has a *neither/nor* construction.

The negative quality of the phrase centred around *budge* is thus expressed in different ways, but with a predominance of collocations *refuse to* (and inflections), *wouldn't, didn't, couldn't.* Colligation is with verbs, with modals (including *able to*) accounting for half the 30 instances.

The distinction between *won't* and *can't* draws attention to two different reasons why people or things do not budge, refusal or inability.

We may wonder why people use this word, why they do not just use the common verb *move*, with which any use of *budge* can be replaced. Something does not budge when it does not move despite attempts to move it. From the perspective of the person who wants something moved, this is frustrating and irritating, and these emotions may find expression, because this is the semantic prosody of the use of *budge.*

The semantic prosody of an item is the reason why it is chosen, over and above the semantic preferences that also characterize it. It is a subtle element of attitudinal, often pragmatic meaning and there is often no word in the language that can be used as a descriptive label for it.

***Budge** if someone will not **budge** on a matter, they refuse to change their mind or to compromise;*

*If something or someone will not **budge** or if you cannot budge them, they will not move at all from a particular place or position.*

Discussion and research points

Study the example of the analysis of the phrase *naked eye* presented below (J. Sinclair. 1996. The Search for Units of Meaning. *Textus* vol. ix, no. 1, p. 75-106.

Sinclair, J. 2000. Lexical Grammar. *Darbai ir Dienos*, t. 24, 191-203.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Sinclair, J. 1999. The lexical item. In Weingand, E. (ed). *Contrastive Lexical Semantics.* Amsterdam/Philadelphia: John Benjamins. P. 1-25.

Tognini-Bonelli, E. 2000. Corpus Classroom Currency. *Darbai ir Dienos* t. 24, 205-243.

Hunston, S. 2007. 'Semantic prosody revisited'. *International Journal of Corpus Linguistics* 12:2, 249-268.

Louw, B. 1993. 'Irony in the text or insincerity in the writer?'The Diagnostic potential of semantic prosodies' In M. Baker, G. Francis & E. Tognini-Bonelli (eds) *Text and Technology. In Honour of John Sinclair.* Amsterdam: Benjamins, p. 157-176.

Research points: mini-project 1.

In groups of 2-6 choose a group of synonymous words and carry out a research project using J. Sinclair's understanding of meaning.

A mini-research project 2.

Study collocations of a word of your choice, using Sinclair's seven-stage procedure (Sinclair 2003:xvi-xvii):

1. Initiate – Search for patterns to the right and left of the node.
2. Interpret – Form a hypothesis that may link these patterns.
3. Consolidate – Look further away from the node to determine if there are variations in the patterns found or additional patterns.
4. Report – Write out your hypothesis to use it for further searches.
5. Recycle – Search again the extended content of the node to find further examples.
6. Result – Record the results for further studies.
7. Repeat – Repeat the process with more data.

Mini-research project 3.

Using Hunston's (2010: 163) 'accumulative collocation' technique, conduct a study of the collocational patterns of a word of your choice. This technique can be used to perform recursive searches that gradually refine what is observed. For example: "The most adjacent word-collocate of *distinguishing* is *between*, so the string *distinguishing between* is then taken as the starting point for further search. The most frequent adjacent collocate of *distinguishing between* is *of*. Taking *of distinguishing between* as the node, the words which most frequently precede this string are: *way, capable, importance, difficult, means, incapable,* and *ways. task, point, method* and *ways*. (Hunston 2010: 163)

7 CORPORA APPLICATION

Corpora in teaching
Corpora in learning
Corpora in research

Corpora in teaching and learning

I. Collective nouns

1. Investigate variation in the verb form used with collective nouns: ***aristocracy, army, audience, cast, committee, community, company, council, crew, data, family, government, group, jury, media, navy, nobility, opposition, press, public, staff, team.***

nen 1982: 140 argues that most ethnographic **data are** conversation-based). As one further measure of
Moreover, the results depend totally on what **data are** put in, and the methodology of necessity limits data
er , in many developing countries geological **data are** often incomplete. And because drilling is expensive,
or bloc towards more monopolization.The **data are** patchy, but indicate a steady, if unspectacular, rise in
a ken of the animals' life styles. The same **data are** plotted in b but the species are categorised into three
in ife ept. And for all of us, until contrary **data are** received, our perception of reality is " true ".

ce the extent to which confidential personal **data is** sold. But as those who want the information get mo
more than an access terminal and important **data is** stored on a central file server, is also changing buying
ter data. At the moment, under English law, **data is** not property, and damage or theft has to relate to the
pupil " succeeds " and which does not. This **data is** incorporated in publicizing of the unit and its work; in
used for. Thus we have a situation where **data is** not transformed into information, ie what staff need

2. Conventional collective noun phrases. Using the BNC, complete the following:

a ____ of chickens
a ____ of partridges
a ____ of ants
a ____ of cattle
a ____ of birds
a ____ of sheep
a ____ of geese
a ____ of deer
a ____ of cattle
a ____ of puppies
a ____ of ants
a ____ of hounds
a ____ of wolves
a ____ of bees
a ____ of locusts
a ____ of fish
a ____ of porpoises
a ____ of whales
a ____ of grapes
a ____ of keys
a ____ of flowers

a ____ of flowers
a ____ of sticks
a ____ of hay
a ____ of mountains
a ____ of trees
a ____ of stars
a ____ of stairs/steps
a ____ of thieves/robbers
a ____ of islands
a ____ of people
a ____ of stones
a ____ of sand
a ____ of shoes
a ____ of books
a ____ of hills
a ____ of events
a ____ of clothes

Words: bouquet, brood, bunch, bundle, chain, clump, cluster, colony, covey, drove, flight, flock, gang, gaggle, group, heap, herd, litter, nest, pack, pair, pile, range, series, shoal, school, suit, swarm.

II. Countable v. Uncountable nouns
Definite v. Zero Article

1. There are in English a number of countable/uncountable pairs of words. Study the examples given, try to work out the difference in meaning between the noun as countable and as uncountable and then try to complete the 'gapped' citations to see if you have formed the correct hypothesis.

Language

" Guidelines tell us we can not transmit an undue amount of bad **language** . " This film's a bit tough. Of course the primary years mark a time when children must master **language** .	The Académie Française, that illustrious guardian of the **French language** , set up by Richelieu in 1635, last week in effect buried the very reform for which it had voted unanimously nine months earlier. The British Council acts as a cultural ambassador for Britain and for the **English language** .

1. This course has similar entrance qualifications to 1A1/1A2. For students taking two modern European languages in their second year, there is also the opportunity to study Russian nineteenth- and twentieth-century history or the history of ___ **Russian language** .
2. She knew she would never be able to master ___ **Greek language** .
3. It is this richness of ___ **scientific language** which I feel has been lost.
4. Some features of Richards's theory may now seem rather out of date: his notion that ___ **poetic language** is purely emotive, his materialistic conception of literary value, his view of the author -- text -- reader relationship.
5. As it happens, in the early days of computers it was thought that it would be only a few years before computers would be able to understand ___ **natural language.**
6. While every effort has been made to express the ideas in ___ **everyday language** , it has been impossible to dispense with some scientific terms.
7. A popular arrangement involved four areas, devoted to reading, art, maths and ___ **language** .

Society

The offices of the Research Defence **Society** were originally located in the premises of the Medical **Society** of London, near Harley Street, where they remained until 1981.	He was angered, and for a moment was tempted to reply that Louise herself had managed to fit in to **French society**, despite her origins and her antecedents, but he curbed himself.
In many of her novels May Sinclair was concerned with her characters' struggle for individuality in a suppressive environment, which was frequently identified with the values of **the Victorian society** the author herself grew up in.	Inevitably the boundaries of what was and was not considered permissible in village life were much clearer in the nineteenth century, as they were in **Victorian society** generally.
A Flocks, elected chair, explained that he had for sometime been anxious as to where **the society** was drifting.	What the media should not do is cause friction and division within **society** and especially it should not encourage opposition or resistance to government decrees.

1. In spite of his years, Sir Richard remains very active and will deliver one of the principal addresses at today's conference at the offices of ___ **Royal Society** in London.
2. The Gordon Riots (1780, described by Dickens in Barnaby Rudge), in which the London mob sacked the houses of Catholics and released the inhabitants of Bedlam, frightened all sections of ___ **English society**.
3. There are intransigent problems about the place of the very old in ___ **modern society**.

4. Norman Tebbit's Disraeli lecture in 1985 spelt out his distaste for the " valueless values of ___ **permissive society** ", of the 1960s and 1970s -- represented by legalized abortion and homosexuality, fewer constraints on what is portrayed in the media and theatre, and growing disrespect for authority.
5. Overall, the Census data suggest not only that the standard of living for those already on state benefits in 1971 has fallen further behind, " but that many more of the residents have become dependent upon benefits and have so little disposable income as to be unable to participate in ___ **consumer society** at all.
6. This power derives, not from any superior individual or institutional competence, but from the strategically important role which these interests have been able to mark out for themselves in ___ **American society**.

Literature

The unwillingness of some English teachers to teach **literature** stems from their convictions about the neglected richness of working-class culture. Graduating MA with first-class honours in **classical literature** in 1869, he spent one year as a private tutor and then from 1871 to 1872 he was an assistant in the department of humanities at King's College, Aberdeen. Forsyth is a top-selling and stylish author -- decidedly not a purveyor of **great literature** , but a man writing for men, with thrills guaranteed.	Pivotal to medical scientific progress is the **scientific literature** , with the discipline imposed by writing and the reasoned critical argument in which the strengths and weaknesses of the scientific case are stated. Much of the **technical literature** on the subject seems to confuse the two sets of questions distinguished in this section.

1. This lasting antipathy coexisted in his mind with a rare mastery of philosophical debate and ___ **classical literature** .
2. for the upper class the belief in the educative values of ___ classical **English literature** was still strong.
3. Their failure to make the most of ___ **scientific literature** seems to begin early in their academic careers.
4. We could make a similar point about ___ **psychological literature** explaining subculture as resistance to parental norms.
5. The one fact which does stand out is this: in the creation of the Victorian town, just as in the life of the ancient universities, in the spread of learning and in the writing and publishing of ___ **Victorian literature** , Nonconformists were a vital element in English life wielding an importance far beyond their numbers.
6. They rarely publish their arguments in ___ **technical literature** ; when they do, the arguments usually fare poorly.

7. All her writings are characterized by an outstanding clarity and vigour of presentation, qualities which were a reflection of her keen interest in ___ **English literature** .
8. Report has it that they have now virtually committed racial suicide, declining to accept the deplorable standards of ___ **modern literature** and paper.

III. Phrasal Verbs

Phrasal verbs are known to be notoriously difficult for language learners.
The main problems:

1. Avoidance.
2. Style deficiency.
3. Semantic confusion.
4. Lack of collocational awareness.
5. Using 'idiosyncratic' phrasal verbs.
6. Syntactic errors.

Choose for each sentence the verb that in your opinion best fits the context and fill in that verb. Assume that these sentences have been written in normal, colloquial English.

1. As we all thought that my uncle had left the country we were surprised to see him _____ at my mother's birthday party.
   A claim B appear C look up D turn up

2. After having failed to have a decent conversation with a German couple I had met in the pub, I decided that it was time to ____ my German.

   A calm down B improve C abolish D brush up

3. We were really astonished when John did not keep his promise: we hadn't thought that he would ever _____ his friends.

   A let down B solve C disappoint D carry on

4. When you are a chain-smoker it is incredibly difficult to _____ smoking.

   A fall down B stop C give up D elect

5. I spent one hour trying to ring my mother from a phone booth but didn't manage to _____ her.

   A earn B get through to C reach D mix up

6. When the weather is nice I love to _____ early.

   A release B look after C get up D rise

7. "Don't you think it's a good idea to have a break now and to _____ playing after lunch?" my hungry bridge-partner asked me.

   A cheer up B continue C flush D go on

8. When the war was just about to _____, in 1940, my father must have been about 15-years- old.

   A break out B look down on C start D satisfy

9. Luckily there would be no one in the embassy-building when the bomb was to

   A go off B explode C tune in D reply

10. According to my grandfather it is very difficult, nowadays, to _____ one's children well.

    A listen B raise C bring up D come across

11. "Hello Suzy? How nice of you to call me! But someone has just rung the doorbell: could you ____ a second?"

    A capture B hang on C wait D fall down

12. She did it again! She always forgets to _____ the fire when she leaves!

    A put out B foresee C extinguish D break into

13. When Jack was late for his date, he knew his girlfriend would be furious, so he had to _____ a story about a traffic-jam.

    A make up B follow C lie down D invent

14. The fight between Robert and Paul stopped when Paul twisted his ankle and had to _____ .

    A realize B surrender C look up to D give in

15. When my aunt had just opened the shop, she was forced to _____ several interesting business-offers, because she was simply short of time.

    A offend B turn down C cheer up D refuse

Using the data from the BNC choose a group of phrasal verbs:

back away back down back off back out back up	pass around pass away pass down pass off pass on pass out pass over pass round
break away break down break in/into break off break out break through break up break with	pay back pay off pay out pay up
catch on catch out catch up catch up with	put about put across put around put away put down put forward put off put on put out put through put together put up
come about come across come along come apart come away come back come between come down come for come forward come from	set about set apart set aside set back set down set forth set in set off set on set to set up

come in come on come off come out come round come to come up	
fall apart fall away fall back fall behind fall out fall over fall through	sit around sit back sit by sit down sit in on sit sit on sit out
get about get across get ahead get after get along get around get away get back get by get down get in get off get on get up	stand back stand by stand down stand for stand out stand up
give away give back give in give off give out give over give up	step aside step back step down step in step on step up
go ahead go along go around go away go back go down go on	take aback take after take against take apart take away take back take down

go out go under go up	take in take off take on take out take over take to take up take up on take upon take up with
hand down hand in hand on hand out hand over hand round	think back think out think over think through think up
keep away keep back keep down keep in keep off keep on keep out keep to keep up keep under	turn against turn around turn back turn down turn off turn on turn out turn over turn round turn up
lay aside lay by lay down lay in lay on lay out	wear away wear down wear off wear on wear out wear through
leave behind leave off leave out	work in work off work on work out work over work up

IV. Prepositions

Study the concordances of *above* and *over* and work out the similiarities and differences between them:

Flopping Central banks from round the world were drafted in yesterday to stop the pound's slide. But they couldn't stop it flopping to 3.79 German marks -- just a fraction **above** the lowest permitted level.
Leading libel lawyer Brian Hepworth confirmed that Diana could be called as a witness. He said: " Only the Queen is **above** the law and could not be subpoenaed.
Mind you, you have to pay for your sound equipment over and **above** the list price.
The parents of 26 children refused to send them to the designated school and instead made arrangements for tuition to be given to them by, inter alios, a volunteer retired teacher, in rooms **above** a public house.
Overall illiteracy rates among the black population are still thought to stand **above** 50 per cent, and schooling is not accessible to many black children.
The imposition of a curriculum from **above** will not mean, if assurance given by politicians is to be believed, that teachers will be prevented from delivering it in the way they think most appropriate.
The result is that all Home Secretaries are grossly over-worked, although most will have found their own ways of keeping their heads **above** water.
I find myself on a small brick platform about twenty feet **above** a man-made, well-bricked channel which follows a straight course through the factories and warehouses. The water looks clear and has long green weeds waving in it.
On Thursday, the unemployment figures for December may show that the number of people out of work and claiming benefit in Britain has risen **above** three million.
DOVER CASTLE, THE KEY OF ENGLAND One of Western Europe's most impressive medieval fortresses, Dover Castle is strategically positioned high **above** the White Cliffs of Dover.

To enter you must be **over** 18 and answer this simple question: What type of video games console is a Mega Drive. Is it a) 8-bit; b) 16-bit; c) 32-bit?
But now, as British airmen are back on active service in the skies **over** Iraq, the secrets of the SAS are coming out.
Profits jumped by a third to **over** £6 million in six months.
They found one badly-injured woman in a toilet. As dense black smoke swirled **over** the town, residents were told to stay indoors.
But you don't play **over** 500 games with three big clubs, have a couple of big moves and become a bad player overnight.
There are **over** 300 horses chasing each other round three Flat meetings today -- not much death there.
Clutching a Union Jack, three-year-old Louis hurtled into the arms of his dad, judo silver medallist Ray Stevens, with joy written all **over** his face.
This combination means tides could leap from a level of nearly five metres to **over** six metres and flooding could occur.
And you don't ever have to bank with Barclays to apply. Accepted in **over** 7 million places worldwide, you can use it wherever you see the familiar MasterCard, Eurocard or Access signs.
Try our magic carpet **over** the desert to the 11 leading business centres of the Middle East.

Using the BNC, study the differences and similiarities between *across* and *over*; *under* and *below; in* and *at;*

V. Idioms.

Since computers do not know what an idiom is, automatic retrieval of idioms using conventional software is only partially possible (for a discussion see: O'Keefe et al. 2007: 80-99). O'Keefe et al. (2007) speak about the 'paradox' of idiomaticity "the very thing which, for native speakers, promotes ease of processing and fluent production seems to present non-native users with an insurmountable obstacle". Firstly, because of their varying degrees of syntactic and lexical flexibility, and because of their often specialized pragmatic attributes, idioms are, simply, difficult to get right. Secondly, idioms, even when correctly produced, can sound strange on the lips of non-native speakers. Often one hesitates to use idioms in a foreign language even if one knows them; it is as if one is claiming a cultural membership and identity one has no right to or does not wish to lay claim to. Thirdly, idioms do not just 'pop up' in native-speaker speech; rather they occur as part of. Native speakers are generally not *taught* the appropriate use of idioms; it is a long-term 'priming' (Hoey 2005) of the items which builds in the native user over many years.

There are several possible pedagogical conclusions which might be drawn from these facts:

The first conclusion might be not to bother with idioms at all, since they are simply too much of a formal obstacle and it may be better to focus on learning and using the many thousands of single words which can largely do the same job.

A second option is to question the dominance of utilitarian approaches to language learning and introduce more traditional, colourful, cultural aspects of language learning.

A third recourse is to engage in the teaching of idioms based on sets of relatively more frequent ones, ones which non-native speakers are likely to hear and see when confronted with native-speaker data, whether it be printed or electronic media, or films, TV and popular music, etc.

Research suggests that the speech of native speakers can be distinguished from the speech of advanced non-native successful users of English by the presence or absence of common chunks. (For more on that see: O'Keefe 2007).

Research points:

Use the BNC and the Corpus of Contemporary Lithuanian to analyse idioms contrastively.

Idioms test (Mackin 1978)
One word required, except in number 16.

1. She's got him eating out of her ___.
2. I don't understand him. He's talking over my ____.
3. It's high ____ he started working seriously.
4. A rag and ____ man.
5. For old times' ___.
6. It goes against the ____.
7. She gave him the cold ____.
8. He pulled a ____ one on me.
9. She went off the deep ____.
10. They hated each other like ____.
11. That should bring him to his ____.
12. He ploughs a lone ____.
13. It's time he learnt the facts of ____.
14. It's all in the melting ____.
15. He prided ____ on his ability to make people laugh.
16. Lucky in ____, unlucky at ____.
17. He was very quick at putting two and two ____.
18. He can't write for ____.
19. You must keep a cool ____.
20. That story's as old as the ____.
21. It's time he ____ his ways.
22. The new plane is ____ to none in the world.
23. He ____ to his guns.
24. It isn't all ____ and skittles.
25. We ought to ____ our blessings.
26. He ____ no bones about it.
27. He ____ his mother a dance.
28. It ____ to reason.
29. The ____ of the morning to you!
30. I did it on the spur of the ____.
31. The wrong ____ of the stick.

32. He soon ____ his tune.
33. He made an honest ____ of her.
34. They'll ____ a rat.
35. Who ____ the beans?
36. Cast your ____ wide.
37. He ____ his spite on her.
38. He's got the ____ of the gab.
39. If we stand ____, we shall be all right.
40. He couldn't do ____ to his food.
41. I've been in some tight ____ in my time.
42. They're trying to keep up with the ____.
43. He's a ____ bore.
44. The wrong ____ of the blanket.
45. I haven't the ____ idea.
46. He has a ____ of his own.
47. ____ and truly.
48. He's a ____ ass.
49. The ____ and the sheep.
50. To his ____ content.

Two or more words required, except in no.68, one word for each dash.

51. Let them ____ in their own ____.
52. It was all he ____ ____ to bring himself to say 'Thank you'.
53. Penny ____, ____ foolish.
54. He told her off in ____ ____ terms.
55. You can't have your ____ ____ eat ____.
56. His knowledge of Greek ____ him in good ____.
57. That's right! You've ____ ____ nail on the ____.
58. Don't take it at its ____ ____.
59. You must sometimes be ____ to be ____.
60. Every time she opens her ____, she ____ foot ____.
61. He hasn't got the ____ of his ____.
62. I've got a ____ to ____ with you.
63. We could do it at the ____ of a ____.
64. It's only fair, when all is ____ and ____.
65. Don't make a ____ out of a ____.
66. It made my ____ run ____.
67. We're all in ____ same ____.
68. Six of the ____.
69. ____ ____ favour.
70. All dressed ____, and ____ to go.
71. Don't ____ my poverty ____ ____ face.
72. It's enough to make him ____ in his ____.
73. He's like a ____ out of ____.
74. He ____ a ____ furrow.
75. He was revealed in his ____ ____.
76. Any man who's ____ his ____ would have done the same.

77. His ____ has ____ of clay.
78. Things have come to a ____ ____.
79. She spent a ____ ____ on cigarettes.
80. I shall ____ my ___ of the whole business.
81. He knows which side ____ bread ____ ____.
82. It's like trying to ____ yourself ____ by your own ____ straps.
83. You must learn to take the ____ with the ____.
84. It was a ____ ____ for him to swallow.
85. She always has the ____ ____.
86. He's a ____ peg in a ____ hole.
87. He went in ____ and ____ of his teacher.
88. He got off on the ____ ____.
89. She ____ ____ murder.
90. They are all ____ in ____ about it.
91. He couldn't for the ____ of ____ understand her.
92. The whole ____ of ____.
93. More ____ to his ____ .
94. Her money's ____ to her ____.
95. He can't ____ beyond the ____ of ____nose.
96. He doesn't let the ____ grow under ____ feet.
97. ____ breeds _____.
98. Don't ____ ____ ____ ____ in public.
99. He's always got ____ ____ in a book.
100. ____ is the best ____.

(Taken from Mackin 1978).

Idioms of comparison

Complete the following:

as black as **...**
as blind as …
as bold as …
as brave as …
as bright as …
as brittle as ….
as brown as ….
as busy as …
as changeable as …
as cheerful as …
as clear as …
as cold as …
as cool as …
as cunning as …
as dark as …
as dead as…
as deaf as…
as different as…

as drunk as …
as dry as …
as dumb as …
as easy as …
as fair as …
as fat as …
as fierce as …
as firm as …
as fit as …
as flat as …
as free as …
as fresh as …
as gay as …
as gaudy as…
as gentle as …
as good as …
as graceful as …
as grave as …
as greedy as …
as green as …
as happy as …
as hard as …
as harmless as …
as heavy as …
as hot as …
as hungry as …
as innocent as …
as keen as …
as large as …
as light as …
as like as …
as loud as …
as mad as …
as merry as …
as mute as …
as obstinate as …
as old as …
as pale as …
as patient as …
as plain as…
as playful as …
as plentiful as …
as plump as…
as poor as …
as pretty as …
as proud as …

as quick as …
as quiet as …
as red as …
as regular as …
as rich as …
as ripe as…
as round as …
as salty as …
as sharp as …
as silent as …
as silly as…
as slender as …
as slippery as …
as smooth as …
as sober as …
as soft as…
as sound as …
as sour as …
as steady as…
as timid as …
as tough as …
as tricky as …
as true as …
as ugly as …
as vain as …
as warm as …
as watchful as …
as weak as …
as wet as …
as white as …
as wise as …
as yielding as …

VI. Lexical difficulties

Use the BNC to study the differences between the following pairs of words:

Adverse, averse
Acute, chronic
Among, amid
Amoral, immoral
Between, among
Biannual, biennial,
Bimonthly, biweekly
Broach, brooch
Cement, concrete
Cession, session
Compare to, compare with

Complement, compliment
Continual, continuous
Convince, persuade
Creole, pidgin
Definite, definitive
Different from, to, than
Disinterested, uninterested
Disposal, disposition
Distinct, distinctive
Each other, one another
Economic, economical
Elicit, illicit
Fewer, less
Flammable, inflammable
Ingenious, ingenuous
Lay, lie
Plethora

False friends

Use the BNC and dictionaries to study the following:

Actual (topical, current)
Alley (avenue)
Costume (suit)
Fabric (factory)
Faction (fraction)
Fantasy (imagination)
Formula (form)
Fraction (decimal fraction)
Human (humane)
Isolate (insulate)
Manager (CEO)
Marmalade (jam)
Massive (solid)
Novel (novella)
Pathetic (emotional)
Patron (cartridge)
Physician (physicist)
Preservative (condom)
Programme (TV) channel
Public (audience)
Receipt
Receipt (recipe)
Smoking (tuxedo, dinner jacket)
Theme (topic, subject)
to conserve (to preserve)
to control (to check, monitor)

to realise (to implement)
to send (to broadcast)
to dislocate

Vocabulary enhancement exercises:

ADVERB + ADJECTIVE COLLOCATIONS

In good English it is untypical and disappointing to describe something or somebody with a simple, single adjective and answer a question like:
„What did you think of the match last night?“ with
„It was good“.

In such cases either an absolute adjective is used e.g. **fantastic,** to which nothing can be added or as a minimum an adverb like „**really“** or „**extremely**“ (or even **terribly, awfully, dreadfully** depending upon the social milieu) to decribe the degree of the adjective.

Make a list of some of these adverbs under the categories of degree and feeling.

Incredibly, astonishingly, magnificently, infuriatingly, irritatingly, impossibly, desperately, passionately, disappointingly, irresisitably, disarmingly.

Try to think of different adverbs which might be used with the following adjectives: (check in the corpus)
Clever, cunning, kind, adept at, meticulous, loyal, viciuos, careful etc.

- Replace the adjective **important** by other adjectives: **critical, crucial, major, serious, significant, vital.**

Synonyms

Use the BNC to study the following:
Ambivalent, ambiguous
Abdicate, abrogate, abjure, adjure, arrogate, derogate
Allay, alleviate, assuage, relieve
Arbitrate, mediate
Assume, presume
Avenge, revenge
Barbaric, barbarous
Between, among
Born, borne
Contrary, converse, opposite, reverse

Empathy, sympathy, compassion, pity, commiseration
Fickle, flexible
Fractious, factitious, fractious
Healthy, healthful, salutary
Imply, infer, insinuate
Sparing, frugal, thrifty, economical
Concise, terse, succinct, laconic, pithy
Conclusive, decisive, determinative, definitive
Dominant, predominant, paramount, preponderant
Doubtful, dubious, problematic, questionable
Effective, effectual, efficient, efficacious
Apparent, illusionary, seeming, ostensible

GLOSSARY

Alignment – tekstų paralelinimas, išlygiavimas

Alignment is the practice of defining explicit links between texts in a parallel corpus; the matching or linking of a text and its translation(s), usually paragraph by paragraph and/or sentence by sentence.

Annotation – anotavimas

Annotation is the practice of adding explicit additional information to machine-readable text.

ASCII – The American Standard Code for Information Interchange

A standard character set that maps character codes 0 through 27 (low ASCII) onto control functions, punctuation marks, digits, upper case letters, and other symbols.

CALL – kompiuterinis kalbų mokymas

Computer-aided (or assisted) language learning.

Character – ženklas, raidė, skaitmuo.

This is a term used to mean generally a letter of an alphabet, but a set of characters includes punctuation marks and other symbols on computer keyboards.

COBUILD

An acronym for Collins Birmingham University International Language Database.

Colligation - koligacija

The likelihood that a grammatical pattern or feature will occur near another grammatical feature or lexical item.

Collocation – žodžių junginys

Collocation is the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening.

Comparable corpora – palyginamasis tekstynas

Comparable corpora are comparable original texts in two or more languages; they are monolingual corpora designed using the same sampling techniques.

Concordance – konkordansas

A concordance is an index to the words in a text. Concordance is a comprehensive listing of a given item in a corpus, also showing its immediate context.

Corpus – tekstynas

A corpus is a collection of naturally occurring language texts, chosen to characterize a state or variety of a language.

Corpus balance – tekstyno balansas

The range of different types of language that a corpus claims to cover.

Corpus-based analysis- tekstynais paremta analizė

Corpus-based analysis starts with a pre-existing theory which is validated using corpus data.

Corpus-driven analysis – tekstynų inspiruota analizė

Corpus-driven analysis builds up the theory step by step from the analysis of corpus data.

Context – kontekstas

The linguistic environment of any expression under scrutiny.

Co-text – kontekstas, (tiriamojo žodžio artimoji apsuptis)

A text occurring around a NODE, as can be seen in a CONCORDANCE. This is a more precise term than context.

Error-tagging – klaidų žymėjimas (mokinio tekstyne)

Assigning codes indicating the types of errors occurring in a learner corpus.

Expected frequency – tikėtinasis dažnis

The frequencies one would expect if no factor other than chance were affecting the frequencies.

Frequency – dažnumas

The actual count of a linguistic feature in a corpus, also called raw frequency.

General corpus – bendrasis tekstynas, bendrojo pobūdžio tekstynas

Idiom principle – fraziškumo principas

One of the main principles of the organization of language – the choice of one word affects the choice of others in the vicinity.

Interlanguage – tarpukalbė, tarpkalbė

The learner's knowledge of L2 which is independent of both the L1 and the actual L2.

Keywords– prasminiai žodžiai, raktažodžiai, deskriptoriai

Words in a corpus whose frequency is usually high (positive keywords) or low (negative keywords) in comparison with a reference corpus.

KWIC

This acronym stands for Key Word In Context.

Lemma – antraštinė žodžio forma

A lemma is the headword form that one would look for if consulting a dictionary.

Lemmatisers – lemuokliai

Tools that group together all of the different inflected forms of the same word.

Monitor corpus – tęstinis tekstynas

A growing, non-finite collection of texts.

Mutual information score – abipusės informacijos įvertis

A statistical score that relates one word to another by comparing the probability that the two words occur together because they belong together with the probability that their occurrence together is just by chance. The score can be used to measure the strength of COLLOCATIONS. The higher the mutual score, the stronger the connection between the two words.

Node – tiriamasis žodis

The node word in a collocation is the one whose lexical behaviour is under examination.

Observed frequency – nustatyti dažniai

The actual frequencies extracted from corpora.

Open-end principle – laisvojo žodžių pasirinkimo principas, laisvųjų žodžių junginių principas

Words are treated as independent items of meaning. Each of them represents a separate choice.

Parallel corpus – paralelus tekstynas

A corpus which contains the same texts in more than one language.

Parsing – sintaksinis tekstyno anotatavimas, sintaksinė tekstyno analizė

A process that analyses the sentences in a corpus into their constituents.

POS- part-of-speech annotation – morfologinis, kalbos dalių anotavimas

Part-of-speech annotation assigns parts of speech to each word (and other token) such as noun, verb, adjective, etc.

Representativeness – reprezentatyvumas

A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its content can be generalized to the said language variety.

Sample – imtis

Elements that are selected intentionally as a representation of the population being studied.

Sample corpus – baigtinis tekstynas

A corpus of finite size consisting of text segments selected to provide a static snapshot of language.

Semantic preference – semantinis laukas

Semantic preference is the semantic field a word's collocates predominantly belong to.

Semantic prosody – semantinė prozodija.

A consistent aura of meaning with which a form is imbued by its collocates.

A discourse function of a unit of meaning.

SEU – the Survey of English Usage

SGML – Standard General Markup Language – ženklinimo priemonė

Span – intervalas

This is the measurement, in words, of the co-text of a word selected for study. A span of -4, +4 means that four words on either side of the node will be taken to be its relevant verbal environment.

Specialised corpus – specialusis tekstynas

A corpus that is domain or genre specific and is designed to represent a sublanguage.

Subcorpus – patekstynis

A component of a corpus, usually defined using certain criteria such as text types and domains.

Tag - žymeklis

A tag is a label attached to a word with some interpretative linguistic information.

Tagging – tekstyno anotavimas, žymėjimas

An alternative term for annotation, especially word-level annotation such as POS tagging and semantic tagging.

Translationese – vertalas

A version of L1 language that has been influenced by the translation process.

References

Aijmer, K., B. Altenberg. 1991. *English Corpus Linguistics.* Longman: London and New York.

Aijmer, K., Altenberg, B. and Johansson, M (eds). 1996. *Language in Contrast: Papers from a symposium on Text-based Cross-linguistic Studies.* Lund: Lund University Press.

Aijmer, K. 2009. *Corpora and Language Teaching.* Amsterdam/Philadelphia. J. Benjamins.

Altenberg, B. 1998. On the phraseology of spoken English: the evidence of recurrent word combinations. In Cowie, A.P. (ed) *Phraseology: Theory Analysis and Applications.* Oxford: Oxford University Press.

Altenberg, B. and Granger ,S. 2001. Grammatical and lexical patterning of *make* in student writing. *Applied Linguistics* 22(2): 173-194.

Aprijaskytė, R. and E. Pareigytė, 1982. *Some Lexical Difficulties for the Lithuanian Learner of English.* Vilnius: Vilnius University Press.

Aston, G. and Burnard, L. 1998. *The BNC Handbook.* Edinburgh: Edinburgh University Press.

Barnbrook, G. 1996. *Language and Computers: A Practical introduction to the Computer Analysis of language*. Edinburgh: Edinburgh University Press.

Biber, D. 1990. Methodological issues regarding corpus-based analyses of of linguistic variation. *Literary and Linguistic Computing* 5: 257-269.

Butler, C. 1992. *Computers and Written Texts.* Oxford: Blackwell.

Chomsky, N. 1957. *Syntactic Structure.* The Hague: Mouton.

Chomsky, N. 1958. Paper given at Third Texas conference on problems of linguistic analysis in English. Austin: University of Texas.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, Mass: MIT Press.

Crowdy, S. 1993. Spoken Corpus Design. *Literary and Linguistic Computing.* Vol. 8, no. 4, 259-265.

Dagneaux E., Denness S., Granger S. and F. Meunier. 1996. *Error Tagging ManualVersion 1.1.* Centre for English Corpus Linguistics. Université Catholique de Louvain, Louvain-la-Neuve.

De Cock, S., Granger, S., Leech, G. and McEnery, T. 1998. An automated approach to the phrasicon of EFL learners. In Granger, S. (ed) *Learner English on Computer.* London: Longman, 67-79.

Fillmore, Ch. 1992. Corpus linguistics or Computer-aided armchair linguistics. In *Directions in Corpus Linguistics***.** Proceedings of Nobel Symposium 82. 35-60.

Firth, J. 1957. *Papers in Linguistics.* Oxford: Oxford University Press.

Firth, J. 1968. A synopsis of linguistic theory. In F. Palmer (ed.) *Selected Papers of J. R. Firth 1952-59.* London: Longman. 168-205.

Francis W.N. and H. Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin.

Gass S.M. and L. Selinker. 2001. *Second Language Acquisition. An Introductory Course.* Mahwah NJ: Lawrence Erlbaum.

Granger, S. 2002. A bird's-eye view of learner corpus research. In S. Granger, J. Hung and S.Petch-Tyson (eds) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Philadelphia: John Benjamins, 3-33.

Granger, S. (ed). 1998. *Learner English on Computer.* London: Longman.

Granger, S. 2003. International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37 (3), 538-546.

Granger, S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. A critical evaluation. In K. Aijmer (ed.) *Corpora and Language Teaching.* Amsterdam/Philadelphia: J. Benjamins. 13 – 32.

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunston, S. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12:2, 249-268.

Hunston, S. 2010. Using a corpus to explore patterns. In *The Routledge Handbook of Corpus Linguistics.* London, New York: Routledge. 152-166.

Jespersen, O. 1995. *A Linguist's Life: An English Translation of Otto Jespersen's Autobiography. e*d. by A. Juul, et al. Odense: University Press of Southern Denmark.

Johansson, S. 2007. *Seeing through Multilingual Corpora. On the use of Corpora in Contrastive Studies.* Amsterdam/Philadelphia: Benjamins.

Johns, T. 1991. "Should you be persuaded": two samples of data-driven learning materials. In T. Johns and P. King (eds). *Classroom concordancing ELR Journal* 4. University of Birmingham.

Kaszubski, P. 1998. Enhancing a writing textbook: a national perspective. In S. Granger (ed) *Learner English on Computer*. London: Longman, 72-185.

O'Keefe, A., McCarthy M. and Carter R. 2007. *From Corpus to Classroom:language use and language teaching.* Cambridge: Cambridge University Press.
Kennedy, G. 1998. *An Introduction to Corpus Linguistics.* London and New York: Longman.
Knowles, G. 1990. The use of spoken and written corpora in the teaching of language and linguistics. *Literary and Linguistic Computing* 5-1:45-48.
Krishnamurthy, R. 2000. Collocation: from *silly ass* to lexical sets. In C. Heffer, H. Sauntson and G. Fox (eds) *Words in Context: A Tribute to John Sinclair on his Retirement.* Birmingham: University of Birmingham.
Kučera, H. 1991. The odd couple: The linguist and the software engineer. The struggle for high quality computerised language aids. In J. Svartvik (ed.). *Directions in Corpus Linguistics* – Proceedings of Nobel symposium 82. Stockholm, 4-8 August 1991. The Hague: Mouton de Gruyter, 401-419.
Lauridsen, K.M. 1996. Text corpora and contrastive linguistics: Which type of corpus for which type of analysis? In Aijmer, K., Altenberg B. And Johansson M. (eds.) *Languages in Contrast. Papers from Symposium on Text-based Cross-linguistic Studies.* Lund Studies in English 88. Lund: Lund University Press. 63-71.
LDELC- *Longman Dictionary of English Language and Culture.* Ed. By D. Summers. 1992.Longman.
Leech, G. 1991. The state of art in corpus linguistics. In K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics*. London: Longman, 8-29.
Leech, G. 1992. Corpora and theories of linguistic performance. *Proceedings of Nobel Symposium 82.* Monton de Gruyter: Berlin, New York.
Leech, G. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research.* Vol. 28, 1- 13.
Leech, G. 1997. Teaching and language corpora: a convergence. In A. Wichmann et al. (eds) *Teaching and Language Corpora.* London: Longman.
Leech, G. 1998. Learner corpora: what they are and what can be done with them. In *Learner English on Computer.* London: Longman, xiv-xx.
Leech, G. and Fligelstone, S. 1992. Computers and corpus analysis. In Butler (ed.) *Computers and Written Texts*. Oxford: Blackwell, 115-140.

Louw, B. 1993. Irony in the text or insincerity in the writer? The Diagnostic potential of semantic prosodies In M. Baker, G. Francis & E. Tognini-Bonelli (eds) *Text and Technology. In Honour of John Sinclair.* Amsterdam: Benjamins, 157-176.

Mackin, R. 1978. On collocations: 'words shall be known by the company they keep'. P. Stevens (ed.) *Studies in Honour of A.S.Hornby.* Oxford: Oxford University Press. 149-164.

Marcinkevičienė, R. 2000. Tekstynų lingvistika. Teorija ir praktika. *Darbai ir dienos,* vol. 24, 7-64.

McCarthy, M. 1998. *Spoken English and Applied Linguistics.* Cambridge: Cambridge University Press.

McEnery, A. and T.Wilson (eds.) 1997. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McEnery T., Xiao R. And Yukio Tono. 2006. *Corpus-Based Language Studies. An Advanced Resource Book.* London and New York: Routledge.

Meyer, C. 2002. *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.

Mukherjee, J and Rohrbach, J. 2006. Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In Kettemann, B. and G. Marko (eds). *Planning, Gluing and painting Corpora: Inside the Applied Corpus Linguist's Workshop.* Frankfurt: Lang, 205-232.

Nesselhauf, N. 2004. Learner corpora and their potential for language teaching. In J. M. Sinclair (ed). *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: J. Benjamins.

O'Keefe, A. and M. McCarthy (eds). 2010. *The Routledge Handbook of Corpus Linguistics.* London: Routledge.

Partington, A. 1998. *Patterns and Meanings.* Amsterdam: Benjamins.

Partington, A. 2004. 'Utterly content in each other's company': semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1), 131-156.

Philip, G. 2011. *Colouring Meaning. Collocation and Connotation in Figurative Language.*Amsterdam: Benjamins.

Pinker, S. 1994. *The Language Instinct.*NewYork: HarperCollins.

Sinclair, J. 2000. Lexical Grammar. *Darbai ir Dienos*, t. 24, 191-203.

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

Sinclair, J. 1996a. The search for units of meaning. *Textus* (IX) vol. IX. No.1, 75-106.

Sinclair, J. (ed) 1996. *Looking Up. An Account of the COBUILD Project.* HarperCollinsPublishers.

Sinclair, J. 2003. *Reading Concordances.* Harlow: Pearson Longman.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Stubbs, M. 1996. *Text and Corpus Analysis.* . Oxford: Blackwell.

Stubbs, M. 2001a. *Words and Phrases: Corpus Studies of Lexical Semantics.* New York: Blackwell.

Stubbs, M. 2001. Texts, corpora, and problems of interpretation: a response to Widdowson. *Applied Linguistics.* 22/2:149-172.

Summers, D. 1991. *Longman/ Lancaster English Language Corpus: Criteria and Design.* Harlow: Longman.

Tognini-Bonelli, E. 2000. Corpus Classroom Currency. *Darbai ir Dienos* t. 24, 205-243.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work.* Amsterdam/Philadelphia: J. Benjamins.

Widdowson, H. 2000. The limitations of linguistics applied. *Applied Linguistics* 21/1:3-25.

Williams, M. 1988. Language taught for meetings and language used in meetings: Is there anything in common? *Applied Linguistics* 9 (1): 45-58.

Zipf, G.K. 1935. *The Psychobiology of Language*. Boston:Houghton Mifflin.

APPENDICES

Table 1. Rank order of 50 most frequent word types in the BNC, the Birmingham Corpus, Brown Corpus and LOB Corpus.

	BNC	Birmingham Corpus	Brown Corpus	LOB Corpus
the	1	1	1	1
of	2	2	2	2
and	3	3	3	3
a	4	5	5	5
in	5	6	6	6
To (inf)*.	6	4	4	4
it	7	9	12	10
is	8	11	8	8
To (prep.)	9			
was	10	10	9	9
I	11	8	20	17
for	12	13	11	11
That (conj.)*	13	7	7	7
you	14	14	33	32
he	15	12	10	12
be	16	18	17	15
with	17	16	13	14
on	18	15	16	16
by	19	29	19	20
at	20	22	18	19
have	21	24	28	26
are	22	27	24	27
not	23	25	23	23
this	24	26	21	22
‘s (Gen)	25			
but	26	20	25	24
had	27	19	22	21
they	28	21	30	33
his	29	23	15	18
from	30	32	26	25
she	31	31	37	30
That (DetP)	32			
which	33	38	31	28
or	34	28	27	31
we	35	30	41	40
‘s (Verb)	36			
an	37	39	29	34
~n’t	38			
were	39	37	34	35
as	40	17	14	13

do	41			
been	42	50	43	37
their	43	42	40	41
has	44			
would	45	44	39	43
there	46	35	38	36
what	47	41	54	58
will	48			
all	49	34	36	39
if	50	43	50	45

Useful references:

British National Corpus (BNC) home: http://info.ox.ac.uk/bnc

BNC is also available at: http://corpus.byu.edu/bnc

BNCWeb: http://bncweb.lancs.ac.uk

British Academic Spoken English (BASE) corpus - http://www.rdg.ac.uk/AcaDepts/II/base_corpus/**)**

British Academic Written English (BAWE) corpus - http://www.coventry.ac.uk/bawe)

CorALit- Corpus of Academic Lithuanian

http://coralit.lt/

Corpus of Contemporary American English (COCA): http://www.americancorpus.org

Corpus of Spoken Professional American English (CSPAE)

http://www.athel.com/cspa.html

Corpus of the Contemporary Lithuanian Language and the **Parallel Corpus** (Czech-Lithuanian, Lithuanian-Czech, English-Lithuanian, Lithuanian-English) compiled at the Centre of Computational Linguistics at Vytautas Magnus University (Kaunas) (http://donelaitis.vdu.lt/)

Corpus of Spoken Lithuanian compiled at the Regional Studies Department, Vytautas Magnus University, Kaunas
http://www.vdu.lt/LTcourses/?pg=41&menu_id=112

CELL: the Corpus of Estonian Literary Language http://www.cl.ut.ee/korpused/baaskorpus/

Corpus of Spoken Estonian http://www.cl.ut.ee/suuline/Korpus.php

International Corpus of Learner English – ICLE .

http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.html

Michigan Corpus of Academic Spoken English (MICASE) - http://www.lsa.umich.edu/eli/micase/index.htm)

Corpus of English as a Lingua Franca in Academic Settings
http://www.uta.fi/laitokset/kielet/engf/research/elfa/project.htm

CADIS - Corpus of Academic English - http://dinamico.unibg.it/cerlis/page.aspx?p=196

David Lee's Corpus-based Linguistic Links
http://tiny.cc/corpora

The Longman Learners' Corpus

http://www.pearsonlongman.com

The Macmillan World English corpus

http://www.macmillandictionary.com/essential/about/corpus.htm

International Corpus of Learner English – ICLE .

http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.html

OPUS corpus of parallel texts:
http://opus.lingfil.uu.se/index.php

The Sketch Engine:
http://sketchengine.co.uk

The Translation English Corpus (TEC)
http://www.llc.manchester.ac.uk/ctis/research/english-corpus/

The World Wide Web corpus (WebCorp)
http://www.webcorp.org.uk

ANSWER KEYS:

Conventional collective noun phrases:

a brood of chickens
a covey of partridges
a colony of ants
a drove of cattle
a flight of birds
a flock of sheep
a gaggle of geese
a herd of deer
a herd of cattle
a litter of puppies
a nest of ants
a pack of hounds
a pack of wolves
a swarm of bees
a swarm of locusts
a shoal of fish
a school of porpoises
a school of whales
a bunch of grapes
a bunch of keys
a bunch of flowers
a bouquet of flowers
a bundle of sticks
a bundle of hay
a chain of mountains
a clump of trees
a cluster of stars
a flight of stairs/steps
a gang of thieves/robbers
a group of islands
a group of people
a heap of stones
a heap of sand
a pair of shoes
a pile of books
a range of hills
a series of events
a suit of clothes

Countable v. Uncountable nouns
Definite v. Zero Article

Language

1. This course has similar entrance qualifications to 1A1/1A2. For students taking two modern European languages in their second year, there is also the opportunity to study Russian nineteenth- and twentieth-century history or the history of the **Russian language** .
2. She knew she would never be able to master the **Greek language** .
3. It is this richness of **scientific language** which I feel has been lost.
4. Some features of Richards's theory may now seem rather out of date: his notion that **poetic language** is purely emotive, his materialistic conception of literary value, his view of the author -- text -- reader relationship.
5. As it happens, in the early days of computers it was thought that it would be only a few years before computers would be able to understand **natural language.**
6. While every effort has been made to express the ideas in **everyday language** , it has been impossible to dispense with some scientific terms.
7. A popular arrangement involved four areas, devoted to reading, art, maths and **language** .

Society

1. In spite of his years, Sir Richard remains very active and will deliver one of the principal addresses at today's conference at the offices of the **Royal Society** in London.
2. The Gordon Riots (1780, described by Dickens in Barnaby Rudge), in which the London mob sacked the houses of Catholics and released the inhabitants of Bedlam, frightened all sections of **English society**.
3. There are intransigent problems about the place of the very old in **modern society**.
4. Norman Tebbit's Disraeli lecture in 1985 spelt out his distaste for the " valueless values of the **permissive society** ", of the 1960s and 1970s -- represented by legalized abortion and homosexuality, fewer constraints on what is portrayed in the media and theatre, and growing disrespect for authority.
5. Overall, the Census data suggest not only that the standard of living for those already on state benefits in 1971 has fallen further behind, " but that many more of the residents have become dependent upon benefits and have so little disposable income as to be unable to participate in the **consumer society** at all.
6. This power derives, not from any superior individual or institutional competence, but from the strategically important role which these interests have been able to mark out for themselves in **American society**.

Literature

1. This lasting antipathy coexisted in his mind with a rare mastery of philosophical debate and **classical literature** .

2. for the upper class the belief in the educative values of classical **English literature** was still strong.
3. Their failure to make the most of the **scientific literature** seems to begin early in their academic careers.
4. We could make a similar point about the **psychological literature** explaining subculture as resistance to parental norms.
5. The one fact which does stand out is this: in the creation of the Victorian town, just as in the life of the ancient universities, in the spread of learning and in the writing and publishing of **Victorian literature** , Nonconformists were a vital element in English life wielding an importance far beyond their numbers.
6. They rarely publish their arguments in the **technical literature** ; when they do, the arguments usually fare poorly.
7. All her writings are characterized by an outstanding clarity and vigour of presentation, qualities which were a reflection of her keen interest in **English literature** .
8. Report has it that they have now virtually committed racial suicide, declining to accept the deplorable standards of **modern literature** and paper.

Idioms of comparison

as black as the Ace of Spades, soot, coal, pitch, midnight, ink...
as blind as a bat, a beetle, a mole…
as bold as brass, a lion…
as brave as a lion…
as bright as silver, noonday, day…
as brittle as glass….
as brown as a berry….
as busy as a bee…
as changeable as the weather, the moon…
as cheerful as a lark…
as clear as a bell, crystal, the nose on your face…
as cold as charity, a frog, a stone, ice…
as cool as a cucumber
as cunning as a fox…
as dark as pitch**…**
as dead as a doornail, mutton**…**
as deaf as a post**…**
as different as chalk from cheese **…**
as drunk as a lord **…**
as dry as a bone, dust, a stick **…**
as dumb as a fish, a statue **…**
as easy as ABC, pie, anything **…**
as fair as a rose **…**
as fat as butter, a pig **…**
as fierce as a tiger **…**
as firm as a rock **…**
as fit as a fiddle **…**

as flat as a board, a pancake **…**
as free as the air, a bird **…**
as fresh as a daisy **…**
as gay as a lark **…**
as gaudy as a peacock, a butterfly **…**
as gentle as a lamb **…**
as good as gold, a play **…**
as graceful as a swan **…**
as grave as a judge **…**
as greedy as a wolf, a pig, a dog **…**
as green as grass**…**
as happy as a king, a lark, the day is long **…**
as hard as a stone, nails **…**
as harmless as a dove, a kitten **…**
as heavy as lead **…**
as hot as fire, pepper **…**
as hungry as a hunter **…**
as innocent as a dove **…**
as keen as mustard **…**
as large as life **…**
as light as a feather, a cork, a butterfly, air, thistledown **…**
as like as two peas in a pod, two beans **…**
as loud as thunder **…**
as mad as a hatter, a March hare **…**
as merry as a cricket **…**
as mute as a fish **…**
as obstinate as a mule **…**
as old as the hills **…**
as pale as a ghost, death **…**
as patient as Job, an ox**…**
as plain as a pikestaff, the nose on your face **…**
as playful as a kitten**…**
as plentiful as blackberries **…**
as plump as a partridge **…**
as poor as a church-mouse, Lazarus **…**
as pretty as a picture **…**
as proud as a peacock, Lucifer **…**
as quick as lightning **…**
as quiet as a lamb **…**
as red as beetroot, fire, blood, a cherry, a rose **…**
as regular as a clockwork **…**
as rich as Croesus **…**
as ripe as a cherry **…**
as round as a barrel, a ball, a globe, an apple **…**
as salty as a herring **…**
as sharp as a razor, a needle **…**

as silent as the grave, the dead, the stars **…**
as silly as a sheep, a goose **…**
as slender as gossamer **…**
as slippery as an eel **…**
as smooth as velvet, butter, oil **…**
as sober as a judge **…**
as soft as butter, down, wax **…**
as sound as a bell **…**
as sour as vinegar **…**
as steady as a rock **…**
as timid as a rabbit **…**
as tough as leather, nails **…**
as tricky as a monkey **…**
as true as steel **…**
as ugly as a scarecrow **…**
as vain as a peacock **…**
as warm as toast **…**
as watchful as a hawk **…**
as weak as a kitten, a baby **…**
as wet as a drowned rat **…**
as white as snow, a sheet **…**
as wise as an owl, Solomon **…**
as yielding as wax **…**

Idioms test
One word required, except in number 16.

1. She's got him eating out of her hand.
2. I don't understand him. He's talking over my head.
3. It's high time he started working seriously.
4. A rag and bone man.
5. For old times' sake.
6. It goes against the grain.
7. She gave him the cold shoulder.
8. He pulled a fast one on me.
9. She went off the deep end.
10. They hated each other like poison/anything.
11. That should bring him to his senses.
12. He ploughs a lone furrow.
13. It's time he learnt the facts of life.
14. It's all in the melting pot.
15. He prided himself on his ability to make people laugh.
16. Lucky in love, unlucky at cards.
17. He was very quick at putting two and two together.
18. He can't write for toffee/nuts.

19. You must keep a cool head.
20. That story's as old as the hills.
21. It's time he mended his ways.
22. The new plane is second to none in the world.
23. He sticks to his guns.
24. It isn't all beer and skittles.
25. We ought to count our blessings.
26. He made no bones about it.
27. He led his mother a dance.
28. It stands to reason.
29. The top of the morning to you!
30. I did it on the spur of the moment.
31. The wrong end of the stick.
32. He soon changed his tune.
33. He made an honest woman of her.
34. They'll smell a rat.
35. Who spilt the beans?
36. Cast your net wide.
37. He vented his spite on her.
38. He's got the gift of the gab.
39. If we stand fast/united/together, we shall be all right.
40. He couldn't do justice to his food.
41. I've been in some tight spots/corners in my time.
42. They're trying to keep up with the Joneses/times.
43. He's a crashing/terrible/deadly/crushing bore.
44. The wrong side of the blanket.
45. I haven't the foggiest/faintest/least idea.
46. He has a mind of his own.
47. Well/good and truly.
48. He's a pompous ass.
49. The goat and the sheep.
50. To his heart's content.

Two or more words required, except in no.68, one word for each dash.

51. Let them stew in their own juice.
52. It was all he could do to bring himself to say 'Thank you'.
53. Penny wise, pound foolish.
54. He told her off in no uncertain terms.
55. You can't have your cake and eat it.
56. His knowledge of Greek stood him in good stead.
57. That's right! You've hit the nail on the head.
58. Don't take it at its face value.
59. You must sometimes be cruel to be kind.
60. Every time she opens her mouth, she puts her foot in it.
61. He hasn't got the courage of his convictions.
62. I've got a bone to pick with you.
63. We could do it at the drop of a hat.

64. It’s only fair, when all is said and done.
65. Don’t make a mountain out of a molehill.
66. It made my blood run cold.
67. We’re all in the same boat.
68. Six of the best.
69. Do me a favour.
70. All dressed up, and nowhere to go.
71. Don’t throw my poverty in my face.
72. It’s enough to make him turn in his grave.
73. He’s like a fish out of water.
74. He plows a lonely furrow.
75. He was revealed in his true colours.
76. Any man who’s worth his salt would have done the same.
77. His idol has feet of clay.
78. Things have come to a pretty pass.
79. She spent a small fortune on cigarettes.
80. I shall wash my hands of the whole business.
81. He knows which side his bread is buttered.
82. It’s like trying to pull yourself up by your own boot straps.
83. You must learn to take the rough with the smooth.
84. It was a bitter pill for him to swallow.
85. She always has the last word.
86. He’s a square peg in a round hole.
87. He went in fear and trembling of his teacher.
88. He got off on the wrong foot.
89. She screamed blue murder.
90. They are all up in arms about it.
91. He couldn’t for the life of him understand her.
92. The whole bag of tricks.
93. More power to his elbow.
94. Her money’s gone to her head.
95. He can’t see beyond the end of his nose.
96. He doesn’t let the grass grow under his feet.
97. Familiarity breeds contempt.
98. Don’t wash your dirty linen in public.
99. He’s always got his nose in a book.
100. Honesty is the best policy.