CORPORA IN
LANGUAGE
STUDIES
JONĖ GRIGALIŪNIENĖ

CORPORA IN LANGUAGE STUDIES

VILNIUS UNIVERSITY
2013
Knyga leidžiama įgyvendinant projektą „Aukštos kvalifikacijos vertėjų rengimo kokybės gerinimas ir plėtra“. Projektą vykdo Vilniaus universiteto Filologijos fakulteto Vertimo studijų katedra.
ACKNOWLEDGEMENTS

In designing and developing this course of lectures, I have received inspiration from many people and countless sources. First of all, Hilary Nesi, who was the first to introduce me to corpora back in 1992 during my research visit to Warwick University. I am also extremely grateful to Christopher Tribble and Susan Maingay, the first British Council representatives in the Baltic States and who, supporting the idea of corpus use in the classroom, funded a series of lectures by Michael Rundell and prof. Geoffrey Leech. Susan Maingay and Christopher Tribble, together with Regina Siniūtė and Michael Ayers, contributed a great deal and made it possible to launch a project of “Language Corpora Centre”, the aim of which was to promote the use of corpora in foreign language learning and teaching.

I would also like to thank Laima Erika Katkuvienė, who understood the importance of corpus studies for the students of English Philology and invited me to teach a course at the Department of English Philology in 1996.

Developments in Corpus Linguistics have inspired my views on language, on how I teach and learn language as well as my research. I especially acknowledge the influence of Geoffrey Leech, whose work, and lectures at the University of Vilnius in 1996 had a major impact on the course I taught over the years. I also acknowledge the work of John Sinclair, who has influenced almost everything I taught. The work of Sylviane Granger had an enormous impact on my views and the projects I participated in as well as coordinated. The LICLE project, the idea of which originated during my first meeting with Sylviane Granger at the University of Warwick in 1998, has taken more than ten years to materialise and finally came to a successful conclusion last year. The LINDSEI-LITH, the Lithuanian component of the Louvain International Database of Spoken English Interlanguage, the compilation of which started in 2012, was inspired by the success of the ICLE project.

I acknowledge the part my colleagues played in making this book happen: Laima Katkuvienė, Rita Juknevičienė, Lina Bikelienė, Nijolė Braženienė, Inesa Šeškauskienė and Nijolė Maskaliūnienė.
I owe a great debt and deep gratitude to my students, who participated in the projects (LICLE, LINDSEI-LITH), commented on the materials used during lectures and who were the ultimate source of my inspiration.
PREFACE

The lecture materials and practical assignments included in *Corpora in Language Studies* were developed for the Corpus Linguistics course taught at the Faculty of Philology, Vilnius University. The materials are intended for undergraduate students, but could probably be useful to more advanced students for revision purposes. The design of the materials aims to ensure that the students are given opportunities to think, discuss, engage in tasks, reflect, research and read critically the texts recommended. Chapter 1 is a short introduction into corpus linguistics with a brief history of corpus studies. Chapter 2 focuses on Chomsky’s criticisms of corpus linguistics while Chapter 3 deals with the advantages of using corpora in language studies. Chapter 4 describes some key issues in corpus compilation, and Chapter 5 introduces some corpus tools of data analysis. Chapter 6 discusses learner corpora, and Chapter 7 considers the use of corpora in translation studies. At the end of the book there is a glossary of some key terms, concepts, their definitions and their equivalents in Lithuanian. The book also provides some references to the best-known corpora. The practical assignments are based, on the one hand, on some typical and common mistakes of Lithuanian learners of English and, on the other hand, they aim to reach students with information about corpora and their application in language learning, teaching and research. The data-driven application of corpora in the classroom encourages students to discover things about language without any preconceptions.
CONTENTS

1. INTRODUCTION TO CORPUS LINGUISTICS. A BRIEF HISTORY OF CORPUS LINGUISTICS. ................................................................. 9

2. N. CHOMSKY AND CORPUS LINGUISTICS. ........................................... 11

3. WHY USE CORPORÆ?......................................................................................... 16

4. CORPUS CREATION. DIFFERENT TYPES OF CORPORÆ. CORPUS DESIGN CRITERIA. ........................................................................................................ 27

5. CORPUS TOOLS AND DATA ANALYSIS. .......................................................... 37

6. LEARNER CORPORÆ. CORPORÆ AND LEARNER LANGUAGE. .............. 56

7. CORPORÆ IN TRANSLATION STUDIES. .............................................................. 68

8. GLOSSARY. ........................................................................................................ 81

9. REFERENCES. ........................................................................................................ 86

10. APPENDIX 1 ...................................................................................................... 94

11. APPENDIX 2 ...................................................................................................... 98
Corpus linguistics can be described as the study of language based on text corpora. *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. Although there are many ways to define a corpus, most scholars agree that the term *corpus* in modern linguistics is used to refer to a collection of machine-readable, authentic texts, chosen to characterize or represent a state or variety of a language. 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 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 (Aijmer et al. 1992). 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 – only a million words. Leech (1992:10) referred to them as the first generation corpora. The second generation corpora, according to 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 large 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 (or even billions) of words, many
of them being in commercial hands and using the technologies of computer text processing (for more information on text corpora see: Appendix 1; O’Keefe, et al 2007: 284-296).

The importance of corpora has not always been as widely accepted as it is nowadays. Nelson Francis, the compiler of the first computerised corpus, recalls that in the early 1960s he was asked about what he was up to at the time, and when he 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. The person who asked him the question 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 at that time was 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 and 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. <...> There were many members of the humanistic world in various academic institutions 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).
2. CHOMSKY AND 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 Oxford English Dictionary) were compiled on the basis of large collections of words; the grammars were also constructed using authentic language data (Poutsma 1914 and Kruisinga 1911 provided copious illustrative examples in their grammars), and other language documenters working 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, that reason has precedence over the senses in the acquisition of knowledge, and that much of this knowledge is innate. In language, a rationalist theory is a theory based on artificial behavioural data and conscious introspective judgements. 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. (see: McEnery and Wilson 1997).
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. According to Chomsky (1965), competence is the ‘ideal’ language system, our tacit, internalised knowledge of a language that makes it possible for us to produce and understand an infinite number of sentences and to distinguish grammatical sentences from ungrammatical sentences. 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 and Wilson 1997: 5)

Another of Chomsky’s criticisms was connected with the fact that a corpus is finite while language is infinite. The assumption that, if a linguist is patient and industrious enough, the sentences of a natural language can be collected and enumerated, just like blades of grass on a lawn, 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 linguistics to be set up alongside other empirical sciences and made language description more objective. Unfortunately, this assumption was false and, as 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 were inadequate for language study, because they would always be ‘skewed’. Some sentences would be in the corpus because they are frequent constructions, some by sheer chance. According to Chomsky (1959:159):

“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 the frequency of a feature in the language is a significant determiner of inclusion. As Chomsky himself noted “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” (McEnery and Wilson 1997:8).

One more criticism made by Chomsky is connected with 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?”(McEnery and Wilson 1997: 9). Corpus research is slow and 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. 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 similarly 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, 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 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 that 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 big enough corpus (such as the 100 million word British National Corpus) will provide a large number of utterances that one is likely to encounter in a language.

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

In the fifty years since 1961, CL 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. The issue of the status of corpus linguistics is still highly contentious. Some scholars argue that corpus linguistics is more than a methodology (Tognini-Bonelli 2001, Leech 1992b) and maintain that it is a new ‘research enterprise and a new philosophical approach to linguistic enquiry’, others claim that it is ‘a methodology rather than an independent branch of linguistics’ (McEnery, et al. 2006:7-8, but a methodology that has ‘a theoretical status’, ‘a methodology with a wide range of applications across many areas and theories of 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:


3. WHY USE CORPORA?

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

- 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, to examine authentic texts when discussing linguistic problems. 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 a corpus is examined, a more objective picture emerges, since there is no prior selection of data. Paper slips could provide useful information on features that struck the excerpter as interesting or 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

---

\(^1\) 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.
in streets and on roads, I am constantly noticing oddities of pronunciation, forms and sentence constructions”.

Most reference books, grammars and dictionaries are also only secondary sources: they present somebody’s selection or 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 taught by many current textbooks 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. For example, learners of English were taught to disagree with sb. by 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 is possible to study real language with corpora” (McCarthy 1998: 19).

Verifiability. Verifiability is a normal requirement in scientific research, therefore, 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. This 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. By studying concordances, learners can discover new meanings of the words they already know, e.g. sample concordances for the word *dimly* from the British National Corpus can expose learners to different contexts of the use of the word *dimly* and provide additional unexpected information on the meaning and use of the word *dimly*.

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, for example, 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:
consciousness. And, contrary to Hume's contention, being very
of tools for the job in hand and others that he can
such as grabbing woods, estates, and livestock. They realized
The best of deciding to be ordained is the knowledge that you
makes sense but, as both Mr Gandhi and Mr Singh began
old, and in those lands our crypts and dark foundations are
its familiar bi-temporal headache. There it is at last, a
dimly perceived in the pages of Scripture. Much as we love our
round his shoulders and his mother was ruffling his black hair.
Doyle. I thought you’d want to hear this.’

‘She said, ‘Lefs do it, Nate!’

The lamp outside in the yard suddenly went out, plunging the
‘ she said. ‘Lefs do it, Nate!’

receive such close scrutiny in a public place the two of them
-cracked paving led to a front door that had a crescent of
dimly noticed by few people at the time. Because of the happy
world of cultivated Roman paganism had receded into a past now only
with instructions to be there upon their return. Wilson felt only
earth. Earth-digested, come to dust. Someone, I thought
ringing. It brought me back too precipitately into the small,
a sanctum than a room in a museum. The room is
er of another: ‘the solution posed the question’.
and there is something of the anonymity of the confessional in this
were no switches or dangling cords. She could still see,
cover, ‘there’s ... ’ she pointed to a figure
dimly glimpsed beside a pillar in the background. ‘Naw it’s

Figure 1. Sample concordances for dimly from the BNC.
‘But...’ Bramble began, but she silenced him with a glare from the astonishing blue eyes. Strange, thought Bramble, that of warm light as excruciatingly awkward as it is refreshingly flattering —

The Jerusalem Post noted, ‘Israel suddenly finds itself in take note of your views, sir.’ Cotton gave him a glare from mother, who had shunted him to transport, and he, the other pretended nothing was going on. Jasper knew she a glare of contempt. ‘What got into you? Were you chicken

But...’ Bramble began, but she silenced him with a glare, glared angrily at the world outside the Cages, all! Kraal was inclined to say, fixing Woil with a glare and flexing his talons back and forth to make it clear that his wings, glared angrily at the world outside the Cages, quite.’ ‘All right,’ said Miller, with a glare at his colleague. ‘For the ladies’ benefit, Peter

set out A kayak stitched with pain Your tree — your oak group. Stockier than Morpurgo, with a less untroubled brow and a glare that took in the Zoo, the Men, the trees, all! Kraal was inclined to say, fixing Woil with a glare and flexing his talons back and forth to make it clear that

‘Giant hangover?’ Mitch offered, and got himself a glare at Luke Denner that would have turned a lesser man to stone, let alone ever designing one.’ She went, with a glare at Luke Denner that would have turned a lesser man to stone. ‘For the ladies’ benefit, Peter

The keen eyes narrowed suspiciously, and something very like a glare wasn’t so far away all. He sent her sending a significant glance towards Doreen. The remark was met with a glare of warm light as excruciatingly awkward as it is refreshingly flattering —, and stalked out. When Morton came in, a few

a glare from mother, who had shunted him to transport, and he

from mother, who had shunted him to transport, and he

she snapped, ‘Sadie? I

Sadie?’ cried one enthusiastic onlooker. May fixed him with a glare without thinking of measuring the glare index. Notwithstanding problems

...Marvellous. We will see whether this spider exists or not [pause] she said

Richard Holt, the longstanding Conservative MP, was carried out in start out your lights and bouncer first. No-one wants to have a glare of national publicity. Dr. Ashok Kumar won for Labour with a lamp in their face at closing time nor to have a doorman.

one thing and another. Irene giggled, Mademoiselle fixed her with a glare. We will see whether this spider exists or not [pause] she said.

Figure 2. Concordances for a glare from the BNC.
3. WHY USE CORPORA?

chosen to prefer the banal clarity of ‘communication’ over

They screwed up their eyes, trying to get used to

a piece of paper over his eyes to see her better in

They were ostracised. The unusual closeness of the two families obscured

technocrats, high-priced consulting firms, and politicians, conducted outside

in the traditional ‘sugar cubes’ style, whitewashed to reflect

significant degree of intra-party bargaining (within coalitions), away from

garish blue has failed to ameliorate the effect. At night,

mainly in its own shadows (figure XX), thus reducing

! It’s you. Are you OK? Blinking in

and right. The enemy opened up with machine guns, but

wildlife experiences take place as they mow down interesting species caught in

Japanese transport. Fastidiously-dressed Victorian tourists silhouetted against

see well enough to steer away to safety, but ironically,

casting deceptive shadows. I could barely see ten yards because of

the roof. I think she spotted me, in spite of

of revelation and the clouded vision of horror. INVERSIONS Regressive

The sky was blue, cloudless. There were meadow browns

of public scrutiny.’ Eventually however the ARC conditionally agreed to

during of the sun. The rooms are of typical Turkish standard and

of parliamentary scrutiny: The concessions made by Cabinet ministers in

of the power station lights transform the complex into something like a

of white, and thereby allowing a less contrasting tonal interplay between

Richard followed the other boy’s gaze and looked down.

of fires made it difficult for them to aim accurately. Stirling

of their XR2’s headlights on the way home from a hill

from the volcano provided a weird and somewhat improbable lighthouse.

, and it was no use trying to look far ahead.

from the footlights for when I called at her dressing room back-stage

1990 was identical and sketching was a tedium of narrowed eyes against

faultless blue of the sky. He screwed up his eyes against

also almost wrecked their personal lives as they tried to cope with

Letters, 3 March, p 607) need not worry about

sues. Bates also wonders if anyone has given any thought to

lamps give off. Public lighting engineers give very considerable thought to

welcome the smaller point-source of light provided by this single-ended lamp. The glare

from the smaller lamps will restrict their use to well-screened fittings,

belts are also more effective windbreaks, and provide more protection from

between source and observer can produce very strong lines which account for

were all led to safety by using sign language to communicate amongst

of wolves. Celandines and violets grew there bright and innocent against

of the bones. There were also a great many red squirrels

Figure 3. Concordances for the glare from the BNC.
Similarly, the word *budge* in the LDELC (1992: 151) is defined as ‘to (cause to) move a little’. An examination of the corpus data, however, shows that the language does not talk about moving. In fact, all the occurrences are either grammatically or lexically negative: (more on that below: Extended unit of meaning).

![Figure 4. Sample concordances for *budge* from the BNC.](image-url)
Enhance learner motivation

Another argument in favour of using corpora 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”.

Using corpora in the classroom can also impact the way languages are learned and taught. Thus, McCarthy (1998: 67-68) argues that the traditional 3 Ps approach methodology – Presentation – Practice – Production should be supplemented by the 3 Is method: Illustration – Interaction – Induction. Illustration means looking at real data where possible; Interaction means talking among learners and teachers about language, sharing and forming views about the language observed and patterns discovered. /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 very strong motivating effect on the learner. What is more, the 3 Ps 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 original observations and discoveries that have probably never been identified 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 two reasons: the first is the Hawthorne effect (see: Mayo 1933) – 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. At present, there is evidence enough in support of the thesis that memory is conditioned by an active cognition of the past. In other words, we may assume that recognizing 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 the circumstances of that input. Using corpora and reading concordancing lines is the learner’s willful mental activity carried out while trying to cope with the complex textual environment of a linguistic item.

It is also 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 some matters 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 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 (When do they use the word X?), second-person elicitations (When do you use the word X?), and first-person introspection (When do I use the word X?). Corpora provide us with third person facts about what people do, but not the facts about what people know, or what they think they do; the facts come from the perspective of the observer, not from the introspection of the insider. Corpora therefore deal with the textually attested, but not with the possible, nor with the contextually appropriate. The data provided are partial, so problems may arise when this partial description is directly applied in the classroom. Widdowson argues that the linguistics of the attested is just as partial as the linguistics of the possible, but admits that “all enquiry is partial” (2000:23). Stubbs (2001) responding to Widdowson's criticism of corpus linguistics admitted that Widdowson's observation about the three types of data was an elegant one; however, he regretted that Widdowson did not discuss how the three levels of reality relate to each other or how such relations can be investigated. He then added that corpora provide us with repetitions of patterns (collocations, colligations, etc.) that can now be discovered and studied. The discussion presented clearly shows that there are some reservations concerning the use of corpora data in the classroom,
however, corpora have already been accepted as valuable data resources in linguistic analysis, language teaching and learning.

**Discussion and research points**

What are corpora unable to tell us?

**Further reading**


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) argued 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 that can be accessed free or for a symbolic fee (when used for research purposes), as well as 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 a corpus, its size, and then choose the texts for inclusion. The whole process of corpus building is unfortunately much more complicated.

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

**Getting permission**

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 puts additional strain on corpora builders. The issue of getting permission has often been addressed by corpora linguists, but unfortunately there is as 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 permission is further aggravated by the fact that different countries have different copyright 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).

**Design**

The design of a corpus is dependent upon the type of 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 the 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, parallel, comparable, 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, the 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) of 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 own limitations and are inappropriate for the study of discourse, infrequent words and text cohesion aspects.

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.

Monitor corpora have a large and up-to-date selection representing the current available state of a language, e.g. the Bank of English forms part of the Collins Corpus, which is an extremely large 2.5 billion word database.

General corpora

General corpora are assembled to serve as a reference base for unspecified linguistic research (Kennedy 1998:19). 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, a general corpus is needed, a collection of texts of as many different types as possible to make the corpus representative. However, 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? Nonetheless, as Kennedy observes (1998:62):

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.

The representativeness of most corpora, as 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 supposed to be representative of the language or the variety under consideration (McEnery 2006:16). However, one should bear in mind that, since at the present time there is no reliable scientific measure of corpus balance, 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 the 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'. Sinclair (2004) warns against perfectionism in corpus building. He argues that corpus building “is an inexact science, and no-one knows what an ideal corpus would be like” (ibid). Therefore, his advice for corpus builders would be to keep a careful record of what is in the corpus, so that material can be added or amended later, and the users would know what the corpus contained. The principle of careful recording of the contents of a corpus was followed in the compilation of the big corpora, such as the Cobuild and the BNC.
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 very considerable imbalance in the frequency of words. Nowadays, with 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. 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 only once 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 it is to accurately represent 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 texts does not constitute a corpus (cf. the difference between a corpus and an archive). The third generation corpora have very often been collected according to what sources of material were made available and are therefore haphazard collections of texts. Initiatives are under way to assemble archives comparable in scope to those 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 in particular to the legal systems and copyright issues discussed above. Copyright holders are unlikely to grant permission 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 that 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 a corpus is likely to be limited by technical constraints. An in-memory concordancer imposes an absolute restriction on the quantity of text that can be analyzed at one time. With other types of software there may be no fixed maximum, but if a 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, and 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. In addition, spoken language is not very well researched and most 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 is quite difficult to obtain. Interviews, debates and discussions on 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, and recording people without their permission is an unjustifiable invasion of privacy.

Film scripts, drama texts, etc., are of little value in a general corpus. They are ‘considered’ language, written to simulate speech in artificial settings, and therefore 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. The presence of only a small amount of spoken language in an otherwise written corpus might yield very peculiar results.

**Discussion and research points**

Discuss the approach to text selection for the BNC. Look at the composition of the BNC:
Table 1. The composition of the British National Corpus.

<table>
<thead>
<tr>
<th>A1 WRITTEN LANGUAGE COMPONENT: INFORMATIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRIMARY SUBJECT FIELD (or DOMAINS)</td>
</tr>
<tr>
<td>Natural and pure science</td>
</tr>
<tr>
<td>Social science</td>
</tr>
<tr>
<td>Commerce and finance</td>
</tr>
<tr>
<td>Belief and thought</td>
</tr>
<tr>
<td>Biography</td>
</tr>
<tr>
<td>Applied science</td>
</tr>
<tr>
<td>World affairs</td>
</tr>
<tr>
<td>Arts</td>
</tr>
<tr>
<td>Leisure</td>
</tr>
<tr>
<td>GENRE</td>
</tr>
<tr>
<td>Books</td>
</tr>
<tr>
<td>Periodicals</td>
</tr>
<tr>
<td>Written to be spoken</td>
</tr>
<tr>
<td>Miscellaneous (published)</td>
</tr>
<tr>
<td>Miscellaneous (unpublished)</td>
</tr>
<tr>
<td>LEVEL</td>
</tr>
<tr>
<td>Specialist</td>
</tr>
<tr>
<td>Lay</td>
</tr>
<tr>
<td>Popular</td>
</tr>
<tr>
<td>DATE: 1975-PRESENT</td>
</tr>
</tbody>
</table>

| 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

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

Further reading
Discussion and research points

Study and report on the composition of the Corpus of the Contemporary Lithuanian Language.
Study and report on copyright laws in Lithuania.

Further reading

Kennedy 1998; Hunston 2002; Meyer 2002, McEnery 2006 provide further information and discussion of corpus design issues in general.
Once a corpus has been compiled, the user has to decide what kind of data is required. The most common tool for extracting data from a corpus is a concordancer – a kind of search engine designed for language study. It looks through a corpus and lists every single occurrence of a word or phrase. To use a concordance, a word or a phrase is entered and the software finds all instances of that word or phrase. The search word is sometimes referred to as a node or a keyword and each instance is displayed together with its immediate co-text. The search word is centred (and very often highlighted). This way of display is called KWIC (Key Word In Context). Figure 5 shows a sample of concordance lines for the word *ambivalent* from the BNC.

Most concordancing tools also make it possible to sort the concordances. They allow patterns to be made visible and easily discoverable. Concordances can be sorted alphabetically to the left or to right of the key word. Figure 6 presents the concordances sorted to the right of the key word *ambivalent*.

Concordances can also be sorted to the left. Figure 7 shows the concordances of the word *ambivalent* sorted to the left.

With parallel corpora it is possible to get bilingual concordances. Figure 8 and Figure 9 present bilingual aligned concordances extracted from the English-Lithuanian parallel corpus:
words, it is worth remarking that 'revolution' is an
narrative plot. From JOHN LICHFIELD in Washington A MUDDLED response and the Toshack era takes off. Thus far his press has been
towards it is not positive, partly as a result of their
by a majority of 25 per cent. Telegraph/Times readers were more
the window-panes' is a passage with an ambiguous, or better
that conjunction of hatred, paranoia, and desire (repressed,
there surfaces the play's underlying knowledge, at once exhilarating,
a straightforwardly 'correct' political attitude. Consider two kinds of
dimension to camp, but one inseparable from a more problematic and
' gives rise to a complex history, and the complex,
to the Conservatives, but their' relationship to the party was often
civil servants in charge of delivering on policy conditions may indeed feel
when they emerge. The Icelandic attitude to the new usurpers is
brain damage in the ring. Public reaction to boxing is now
fields of daisies.' Since Christmas his feelings about Keelan were
central character. an engineer, is portrayed as a more deeply
on one of the central criticisms made of Blackeyes: its dangerously
the complete collapse of the economy. But the public remains
' Cold wind that stirs up bitter memories Peter Hetherington reports on
, the Admiralty was soon fighting back. The Army was more
— is the bleakness of the alternative facing Mr Hussein: an
. It is familiar to us from our own experience. The
on the Spanish Civil War as being as 'torpid and as

ambivalent trope. On the one hand, it denotes an absolute,
attitudes in the Bush administration contributed to the failure of Tuesday's
ambivalent attitudes toward sex crimes, but mostly because it is policed by
ambivalent, however, and our small sample of Guardian readers preferred the
ambivalent, reference, since it holds together in one term both the
ambivalent, or overt) which characterizes some of those same heterosexually identified
ambivalent, appalling, and violent; a knowledge which incites yet also
ambivalent transgressive reinscription within gay culture, camp and machismo. As styles
ambivalent relation to both femininity and women: the gay male parody of
ambivalent, conflicted identity of the subordinated (Black Skin, 83)
ambivalent . 'Not sure that I want the Americans, fairly certain
ambivalent . As a display of complete athleticism a great boxing match is
ambivalent , He was flattered by her interest. In the desert in
ambivalent figure, driven by his ambition to build a plane that goes
ambivalent attitude to its central character. There was never much chance that
ambivalent about the Prime Minister's free-market policies and there are rumblings of
ambivalent feelings created by a memorial service in the town scarred by last
ambivalent : while welcoming the end of National Service and the return to
ambivalent slab of northern territory run for an indefinite period by foreign armies
ambivalent world of need, high hopes for the future, conflict
ambivalent as that of the National Government. The Spanish Civil War

Figure 5. Concordance display of the word *ambivalent* from the BNC.
did the rural foresters in West Germany and Sweden. It was
children who disapproved with the smoking behaviour of those who were
— had the sensuality but was losing it — might be more
think you really believe in ghosts!’ ‘I’m fairly
who rule the roost at Westerburgh Higli. But she’s secretly
, I think it is. So it adds up to something
him for what he was having and they wanted. They were
possesses a degree of candour unusual in political life, is still
recommending the path of virtue. I understood that my mother was
thinks so, too.’ From the beginning, Constance felt
sectarianism and schism. Unlike the ILP, Marasts tended to be
state are obvious — but the women’s liberation movement is also
arise for the individual and those around him/ her. Society is
work and they take a lot of changing, so I am
work in Hamburg is proceeding apace. German shipyard workers have been
civil servants in charge of delivering on policy conditions may indeed feel
EAN was not an American- or Western-controlled body, although they remained
have noted that many were either pressured to present for help or
centre and places for people to sit and enjoy and I’m
practitioners were to refrain from referring patients in cases where they were
— though without compromising printer performance. The company is at best
a pocket Cotswolds for your Manchester of the East. You were
the complete collapse of the economy.’ But the public remains
certain extent ‘blind-a-Hey’ situations. Tho’g Beveridge was rather
some sausage rolls?’ asked Betty. ‘You are very
mullahs in Morocco. Moreover, King Hassan was reploled to be
twenty-one-year-old Iranian student of architecture. John Kennedy was even more
she intend to call on him for support. Nicholson remained fairly
about acid rain because of its experiences with lakes. It managed
about advertising falling between that of the two groups ... Children
about being raped. Sam said he agreed and that I should
about ghosts,’ he admitted with a smile. This
about her friends and an open breach looms when she’s drawn
about him, so much so that I have to wonder.
about his handling of economic policy. ‘If I’d taken
about marriage, bitter about her own experience of it while seeing
about Nicky Scott Wilson. She found him physically attractive, but
about parliamentary socialism although most of them accepted the cc
about paying housewives. On the one hand the Wages for Houseworld
about recognising that elderly people have a legitimate wish to contr
about that one. Well, thank you very much. Carol
about the contract. Even if they had wanted to black the
about the desirability of an individual structural adjustment loan. SW
about the developing emphasis on regional political cooperation with
about the helping regimes. And, finally, according to their
about the leaf stem. I understand it’s been offered several
about the necessity for specialist intervention. Another reason for exp
about the Network Printing Alliance however — yet another industry
about the new Poland; detested Russia, but grew tearful singing
about the Prime Minister’s free-market policies and there are rumbling
about the role of the ‘blind alley’ his recognition
about the sausage,’ Lydia accused her. ‘Do you
about the Shah. It was widely said that he had long
about the Shah than Eisenhower had been. When Kennedy became pi
about the thought of having a son. mainly, he explained

Figure 6. Concordances of the word *ambivalent* sorted one word to the right.
...the Spanish Civil War as being as ‘torpid and ambivalent’ as that of the National Government. The Spanish Civil War

However, the moral status of the victim is often portrayed ambivalent in murder trials. There are a range of behaviours and relationships

Gunners, a Carnegie Medal winner, I would at best be ambivalent, and at worst, reluctant to invite pupils to share its

part of the person initiating the divorce. Some people can be ambivalent in this way for years. A frequent cause of complaint was

tasks is more difficult to arrange and criterion performance may thereby be ambivalent. But how precisely can test items be specified by a criterion

mullahs in Morocco. Moreover, King Hassan was reported to be ambivalent about the Shah. It was widely said that he had long

is needed to maintain a favourable position; BCG seems to be ambivalent here about the need to invest to increase market share. In

sectarianism and schism. Unlike the ILP, Marxists tended to be ambivalent about parliamentary socialism although most of them accepted the combin

Hopkins’s attitude to the publication of his poetry had always been ambivalent, but his habit of sending copies to Bridges indicates that he

turntable. The company attribute to matters digital however has always been ambivalent. In fact that’s putting it mildly indeed. The launch

to portraits of himself and his bride. There always had been ambivalent feelings between father and oldest son. At the wedding ceremony,

the first of her articles reproduced here, the TUC had been ambivalent. The general opinion is that Real are tighter in defence.

the Toshack era takes off. Thus far his press has been ambivalent. The feminist response to abortion as a moral problem has been ambivalent. Feminists and other progressives have always been uncomfortable with it:

West. The climbing world’s response over the years has been ambivalent. The BMC’s current view is a wanly pragmatic notion that

because of being in such deep sleep. Research findings have been ambivalent in this area and no firm results have been shown. Children

have expressed an opinion about the idea. Mostly they have been ambivalent, admitting on the one hand that both creativity and psychosis probably

work in Hamburg is proceeding apace. German shipyard workers have been ambivalent about the contract. Even if they had wanted to black the

is critical of attitudes stressing the paramountcy of biological parenthood as being ambivalent and contradictory. In her view society overvalues children’s ties with

— though without compromising printer performance. The company is at best ambivalent about the Network Printing Alliance however — yet another industry body

the window-panes is a passage with an ambiguous, or better ambivalent, reference, since it holds together in one term both the

of their own power? Well I think I feel a bit ambivalent on this point because [pause] and it’s quite possible I was smiling

a popular and psychological shorthand for her conclusions, has important but ambivalent significances for psychology and feminism. Gilligan’s assertion of a differ

conflict once the sons began to internalize the primal father through chronically ambivalent, guilt-ridden mourning for him. Thereafter, the entire conflict became

Figure 7. Concordances of the word ambivalent sorted one position to the left.
Word **FACILITY** is found 11 times.

1. Where use is made of the *facility* in paragraph 1, the party concerned shall undertake in writing to pay a sum equal to that which he would have been required to pay had he given a security and that security has subsequently been forfeited in part or entirely.

2. I met her in January 2000 on a school tour of the *facility*.

3. The world’s largest scientific research facility - Switzerland’s Conseil Européen pour la Recherche Nucléaire (CERN) - recently succeeded in producing the first particles of antimatter.

4. But CERN has now broken ground on its new Antiproton Decelerator - an advanced antimatter production *facility* that promises to create antimatter in much larger quantities.

5. I run a research *facility*.

6. But as soon as a camera left the *facility* and was out of range, the signal was lost, and the screen went blank.

7. You crossed a line when you made that specimen, and you have put this entire facility at risk.

8. “Does this object belong to your *facility*?”

9. The director of your *facility* refused to share any details about this object with me over the phone except to say that I needed to find it immediately.

10. “Despite my attire, which I realize you find troublesome, I am a senior level physicist at the world’s most advanced subatomic research *facility*.

11. He assumed somewhere in this colossal *facility* was a body...
3. THE FRAMEWORK FOR COHESION POLICY 2007 - 2013

3. 2007 - 2013 M. SANGLAUDOS POLITIKOS STRUKTŪRA

In this section, the new framework for cohesion policy is discussed in relation to the need for thematic and geographical concentration. Siame skyrui aptariamas naujosios sanglaudos politikos struktūros ryšys su būtinybe sukurti tematine ir geografinę santaką.

Account also needs to be taken of other Community policy priorities, first because of their potential to contribute to growth and jobs within the framework of cohesion policy programmes and, second, in order to foster synergies with other EU policies on the ground. Taip pat reikia atsižvelgti į kitų Bendrojo politikos kryptių prioritetų, primiausia dėl jų potencialo prisidėti prie ekonomikos augimo ir darbo vietų kūrimo įgyvendinant sanglaudos politikos programas, ir antra, siekiant teigiamos sąveikos su kitomis vietos lygmenių įgyvendinamomis ES politikos kryptimis.

Basic infrastructures and services will need to be created, upgraded and expanded in order to open up regional and local economies, set up a proper business support framework and exploit the opportunities afforded by the Single Market. Siekiant atverti regionų ir vietos ūkio sistemą, sukurti tinkamą verslo paramos struktūrą ir pasinaudoti visomis bendriosios rinkos teikiamomis galimybėmis, reikės sukurti, atnaujinti ir išplėsti pagrindinę infrastruktūrą ir paslaugas.

Capacity building for public administrations at national, regional and local level, to improve the administrative framework for economic activity, good policy design and implementation, including better lawmaking, evaluation and impact analysis of policy proposals, and regular screening of delivery mechanisms, should therefore be supported by the Structural Funds. Nacionaliniu reginio ar vietos lygmenių struktūriniu fondai turėtų remti viešojo valdymo administracinių gebėjimų stiprinimą siekiant pagerinti ūkinių veiklos administravimo struktūrą, geros politikos formavimą ir įgyvendinimą, išskaitant geresnes teisės aktų kūrimo sistemą, politinių pasiūlymų vertinimą ir poveikio analizę bei reguliarius vykdyto mechanizmų patikrinimą.

In considering whether to undertake a particular project as a PPP, it is important that certain conditions are met before contracting out the provision of goods and services, in particular the putting in place of an appropriate legal framework. Svarstant, ar taikyti konkrečiam projektui PPP modelį, prieš sudarant santykis sutartis dėl prekių ir paslaugų teikimo, svarbu atlikti tam tikras sąlygas, visų pirma įdėti vietovę atitinkamą tvarką.

The guidelines rather represent a single framework which Member States and regions are invited to use when developing national and regional programmes, in particular with a view to assessing their contribution to the objectives of the Union in terms of cohesion, growth and jobs. Tiesą sakant, gairės atspindinti bendrą struktūrą, kuria naudojotis kviestamos valstybės narės ir regionai rengiant nacionalines ir regionines programas, tam, kad galėtų vertinti savo indėlį siekiant Sąjungos tikslų sanglaudos, ekonomikos augimo ir darbo vietų kūrimo srityse.

Third, particular attention should be paid in the Convergence regions to modernising the railway system by carefully selecting the priority sections, ensuring their interoperability within the framework of the European Rail Transport Management System (ERTMS). Trečia, ypatingas dèmesys turėtų būti skiriamas konvergencijos regionams siekiant modernizuoti geležinkelių sistemą atidžiai pereinant prioritetines srities, užtikrinant jų sąveikomą su Europos geležinkelių transporto valdymo sistema (angl. ERTMS).
Another way of extracting data from a corpus is to have a corpus annotated. Annotation means adding “interpretative” linguistic information to a raw corpus (Leech 1993: 275). Corpus annotation can be undertaken at different levels: prosodic, morphosyntactic, syntactic, semantic, and pragmatic/discoursal. The most popular and widely used is a grammatical or part-of-speech annotation, when labels or tags are attached to a word to indicate its grammatical class. Other types of annotation are not so well developed. Corpus annotation is a controversial issue and some linguists are against any type of ‘human intervention’ in a raw corpus; they claim that annotation imposes a certain type of linguistic analysis upon a corpus user - a user is led by annotations rather than by real corpus data, and annotations always raise the issue of accuracy and consistency. On the other hand, however, an annotated corpus has added value (McEnery 2006: 29-32) and can both expand the range of research questions that can be asked and also facilitate the process of data analysis. It is much easier to extract information from annotated corpora. Thus, for example, without part-of-speech annotation, it would be difficult to extract right as an adjective from a raw corpus, as its various forms and meanings cannot be identified on the basis of the word form itself.

**Frequency lists and other statistics**

Frequency information plays a very significant role in language research and language learning. Nation (1990) claimed that the 4,000-5,000 most frequent words account for nearly 95 percent of written text and the 1,000 most frequent words account for 85 percent of speech. Although frequency data cannot be the only source of information for a learner when choosing which words to learn, it can be a very good starting point when prioritizing which words to learn first. Corpus-based frequency lists - that is a list of all the words that occur in a corpus, are a great help for a learner. Frequency lists can be ordered alphabetically or by frequency and can help to identify the most and the least common items used in a corpus. They can help researchers in choosing the most frequent, or the most unusual, creative lexical items for study. Frequency data however should be approached with caution. *Raw frequencies* can be used only when no comparison between different corpora is involved, otherwise frequencies should be normalized. Besides, counting words, even with the help of computers, is not as straightforward as it might appear. Gardner (2007) pinpoints the problems that are connected with multiple word meanings, grouping words into families, lemmas, etc. Nevertheless, most linguists
agree that “human intuition about language is highly specific and not at all a good guide to what actually happens when people use the language” (Sinclair 1991: 4, Alderson 2007), therefore, frequency lists and frequency dictionaries are the best current estimate of a word’s importance.

Table 2. Rank order of 50 most frequent word types in the BNC, the Birmingham Corpus, Brown Corpus and LOB Corpus.

<table>
<thead>
<tr>
<th></th>
<th>BNC</th>
<th>Birmingham Corpus</th>
<th>Brown Corpus</th>
<th>LOB Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>of</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>and</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>a</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>in</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>to (Inf.)</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>it</td>
<td>7</td>
<td>9</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>is</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>to (Prep.)</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>was</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>I</td>
<td>11</td>
<td>8</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>for</td>
<td>12</td>
<td>13</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>that (Conj.)</td>
<td>13</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>you</td>
<td>14</td>
<td>14</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>he</td>
<td>15</td>
<td>12</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>be</td>
<td>16</td>
<td>18</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>with</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>on</td>
<td>18</td>
<td>15</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>by</td>
<td>19</td>
<td>29</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>at</td>
<td>20</td>
<td>22</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>have</td>
<td>21</td>
<td>24</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>are</td>
<td>22</td>
<td>27</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>not</td>
<td>23</td>
<td>25</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>this</td>
<td>24</td>
<td>26</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>'s (Gen)</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>but</td>
<td>26</td>
<td>20</td>
<td>25</td>
<td>24</td>
</tr>
<tr>
<td>had</td>
<td>27</td>
<td>19</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>they</td>
<td>28</td>
<td>21</td>
<td>30</td>
<td>33</td>
</tr>
<tr>
<td>his</td>
<td>29</td>
<td>23</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>from</td>
<td>30</td>
<td>32</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>she</td>
<td>31</td>
<td>31</td>
<td>37</td>
<td>30</td>
</tr>
</tbody>
</table>
The information provided in Table 2 shows that the most frequent items have a relatively stable distribution – the top 20 most frequent words are nearly the same in the four corpora, regardless of the differences in size. There is however a very considerable imbalance in the frequency of the words – the most common words are the articles, particles, prepositions, pronouns and other “small words” while content words do not occur among the 50 most frequent words in the four corpora. This fact is significant for corpora compilers and means that corpora have to be as large as possible, especially if the aim is lexical studies. The patterns of word occurrence in texts were first noted by Zipf (1935).

Many computer programs will have some other tools that provide lists of statistics for a particular corpus. The most useful among them are word type, token and type-token ratio measures. The number of tokens refers to the total number of running words in the corpus, types mean the number of different words in the corpus, while the token-type ratio refers to the relationship between the number of running words in a corpus and the types of different words used in the corpus. This tool makes it possible to compare different corpora in order to determine the variety of the word

<table>
<thead>
<tr>
<th></th>
<th>BNC</th>
<th>Birmingham Corpus</th>
<th>Brown Corpus</th>
<th>LOB Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>that (DetP)</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>which</td>
<td>33</td>
<td>38</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>or</td>
<td>34</td>
<td>28</td>
<td>27</td>
<td>31</td>
</tr>
<tr>
<td>we</td>
<td>35</td>
<td>30</td>
<td>41</td>
<td>40</td>
</tr>
<tr>
<td>‘s (Verb)</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>an</td>
<td>37</td>
<td>39</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>~’n’t</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>were</td>
<td>39</td>
<td>37</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td>as</td>
<td>40</td>
<td>17</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>do</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>been</td>
<td>42</td>
<td>50</td>
<td>43</td>
<td>37</td>
</tr>
<tr>
<td>their</td>
<td>43</td>
<td>42</td>
<td>40</td>
<td>41</td>
</tr>
<tr>
<td>has</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>would</td>
<td>45</td>
<td>44</td>
<td>39</td>
<td>43</td>
</tr>
<tr>
<td>there</td>
<td>46</td>
<td>35</td>
<td>38</td>
<td>36</td>
</tr>
<tr>
<td>what</td>
<td>47</td>
<td>41</td>
<td>54</td>
<td>58</td>
</tr>
<tr>
<td>will</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>49</td>
<td>34</td>
<td>36</td>
<td>39</td>
</tr>
<tr>
<td>if</td>
<td>50</td>
<td>43</td>
<td>50</td>
<td>45</td>
</tr>
</tbody>
</table>
forms used. Some software, such as *WordSmith Tools*, have built-in standardized type-token ratio calculations, which allow for the calculation of the type-token ratio for every 1,000 words.

Corpus linguistics has a wide range of statistical tools and measures (Oakes 1998), but many of them are used by linguists only as built-in features of the software they use to process data. Thus for example, *WordSmith Tools* provides researchers with the KeyWords function, Wordlists function, etc. Linguists also use some other statistical measures, such as tests of significance: the Pearson chi-square test, the log-likelihood, the mutual information test for significant collocations, etc. (more on the use of statistics in corpus linguistics in: Barnbrook 1996, Oakes 1998, McEnery and Wilson 1997). Many linguists claim that a rough indication of frequencies, an observation of broad tendencies in the data would be enough for linguists to suggest some interesting areas of research.

Software tools can also be used to study collocations, i.e. 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 (Sinclair 1991: 170). 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. Determining the 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”.

Computerized corpora allow us to study collocations by observing concordancing lines to see which words occur to the right or to the left of the node or search word as in Figure 10.
calls his immediate and unconditional release. He was denied practices, including the recruitment of ethnic minorities and women, and At present almost no horticultural information is stored within the system, on matters which might be thought to be contentious. In providing of minority shareholders. Accounting, taxation and legal due diligence, sorted by Amnesty International's concern that Vic Williams was not guaranteed Lord Fraser contends that the reforms will 'ensure the continuation of credible motive as well as reasonable opportunity of committing the crime and base value. Arrangements will be made for shortlisted bidders to have Lawyers] [contact]. In order to allow all parties but shall remain our property. We shall have the right of other similar accommodations for individuals with disabilities. The concept of, 'meaning such an individual: who, with or without to Whitehall as a biased anti-nationaliser who was unwilling to accept a lower income groups and the unemployed that are suffering very much [unclear] move to a guest house or hotel but we would only pay to think that you're getting good quality of training, and Despite this, the hiring of a personal assistant might be a or isolated branch or site rather than incur investment costs in making the job. If they can do the job with or without the general principles established by US law. A failure to make evident that neither of Lord Greene M.R.'s constructions of the term give them some reassurance that the British heavyweight champion could give a Thiercelin managed to piece together the events of the previous night with There are various ways in which popular opinion can be represented with support could, in consultation with the client, be identified with show and explain the company’s transactions, to disclose, with reference strip and will sometimes trigger spurious diagnostics. It measures with

Figure 10. Sample concordances for reasonable from the BNC.
Although it is possible to study collocations by scrutinising concordancing lines, computer software can also compute and present collocations in tabular, list or graphic form (see Figures 11 – 14 below).

<table>
<thead>
<tr>
<th>No.</th>
<th>Word</th>
<th>Total No. in whole BNC</th>
<th>Expected collocate frequency</th>
<th>Observed collocate frequency</th>
<th>In No. of texts</th>
<th>Mutual information value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>foreseeability</td>
<td>21</td>
<td>0.005</td>
<td>6</td>
<td>3</td>
<td>10.1371</td>
</tr>
<tr>
<td>2</td>
<td>firmness</td>
<td>151</td>
<td>0.038</td>
<td>8</td>
<td>3</td>
<td>7.7061</td>
</tr>
<tr>
<td>3</td>
<td>precautions</td>
<td>577</td>
<td>0.146</td>
<td>30</td>
<td>20</td>
<td>7.6789</td>
</tr>
<tr>
<td>4</td>
<td>suspecting</td>
<td>144</td>
<td>0.037</td>
<td>7</td>
<td>5</td>
<td>7.5819</td>
</tr>
<tr>
<td>5</td>
<td>endeavours</td>
<td>338</td>
<td>0.086</td>
<td>13</td>
<td>6</td>
<td>7.244</td>
</tr>
<tr>
<td>6</td>
<td>eminentiv</td>
<td>221</td>
<td>0.056</td>
<td>8</td>
<td>8</td>
<td>7.1566</td>
</tr>
<tr>
<td>7</td>
<td>infer</td>
<td>318</td>
<td>0.081</td>
<td>9</td>
<td>8</td>
<td>6.8015</td>
</tr>
<tr>
<td>8</td>
<td>timescale</td>
<td>215</td>
<td>0.055</td>
<td>6</td>
<td>6</td>
<td>6.7812</td>
</tr>
<tr>
<td>9</td>
<td>diligence</td>
<td>312</td>
<td>0.079</td>
<td>8</td>
<td>7</td>
<td>6.6591</td>
</tr>
<tr>
<td>10</td>
<td>assume</td>
<td>4,052</td>
<td>1.028</td>
<td>90</td>
<td>73</td>
<td>6.4519</td>
</tr>
<tr>
<td>11</td>
<td>believing</td>
<td>1,360</td>
<td>0.345</td>
<td>30</td>
<td>15</td>
<td>6.442</td>
</tr>
<tr>
<td>12</td>
<td>approximation</td>
<td>333</td>
<td>0.084</td>
<td>7</td>
<td>7</td>
<td>6.3724</td>
</tr>
<tr>
<td>13</td>
<td>sized</td>
<td>438</td>
<td>0.111</td>
<td>9</td>
<td>9</td>
<td>6.3396</td>
</tr>
<tr>
<td>14</td>
<td>grounds</td>
<td>6,062</td>
<td>1.538</td>
<td>109</td>
<td>49</td>
<td>6.1471</td>
</tr>
<tr>
<td>15</td>
<td>suspicion</td>
<td>1,562</td>
<td>0.396</td>
<td>27</td>
<td>13</td>
<td>6.0902</td>
</tr>
<tr>
<td>16</td>
<td>expectation</td>
<td>1,285</td>
<td>0.326</td>
<td>22</td>
<td>20</td>
<td>6.0763</td>
</tr>
<tr>
<td>17</td>
<td>occupier</td>
<td>371</td>
<td>0.094</td>
<td>6</td>
<td>2</td>
<td>5.9941</td>
</tr>
<tr>
<td>18</td>
<td>unreasonable</td>
<td>978</td>
<td>0.248</td>
<td>15</td>
<td>8</td>
<td>5.9177</td>
</tr>
<tr>
<td>19</td>
<td>excuse</td>
<td>2,934</td>
<td>0.744</td>
<td>44</td>
<td>16</td>
<td>5.8852</td>
</tr>
</tbody>
</table>

Figure 11. MI scores for the word *reasonable*. 
Computer software can also help in carrying out more sophisticated collocational analyses. Thus, the BNCWeb can provide the collocational tree of the word reasonable ordered by MI (mutual information), t-score, or z-score, which are the measures of significance for selecting and evaluating significant collocations. The MI score is the

<table>
<thead>
<tr>
<th>No.</th>
<th>Word</th>
<th>Total No. in whole BNC</th>
<th>Expected collocate frequency</th>
<th>Observed collocate frequency</th>
<th>In No. of texts</th>
<th>T-score value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>2,164,238</td>
<td>549.118</td>
<td>1738</td>
<td>928</td>
<td>28.5177</td>
</tr>
<tr>
<td>2</td>
<td>to</td>
<td>2,593,729</td>
<td>658.090</td>
<td>1409</td>
<td>677</td>
<td>20.0047</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>990,281</td>
<td>251.258</td>
<td>716</td>
<td>452</td>
<td>17.3682</td>
</tr>
<tr>
<td>4</td>
<td>it</td>
<td>1,054,279</td>
<td>267.495</td>
<td>737</td>
<td>477</td>
<td>17.2944</td>
</tr>
<tr>
<td>5</td>
<td>that</td>
<td>1,118,985</td>
<td>283.913</td>
<td>654</td>
<td>431</td>
<td>14.4716</td>
</tr>
<tr>
<td>6</td>
<td>be</td>
<td>650,082</td>
<td>164.941</td>
<td>481</td>
<td>294</td>
<td>14.411</td>
</tr>
<tr>
<td>7</td>
<td>seems</td>
<td>20,738</td>
<td>5.262</td>
<td>129</td>
<td>109</td>
<td>10.8945</td>
</tr>
<tr>
<td>8</td>
<td>care</td>
<td>24,369</td>
<td>6.183</td>
<td>129</td>
<td>49</td>
<td>10.8134</td>
</tr>
<tr>
<td>9</td>
<td>time</td>
<td>152,502</td>
<td>38.693</td>
<td>179</td>
<td>113</td>
<td>10.487</td>
</tr>
<tr>
<td>10</td>
<td>would</td>
<td>245,366</td>
<td>62.255</td>
<td>215</td>
<td>152</td>
<td>10.4171</td>
</tr>
<tr>
<td>11</td>
<td>grounds</td>
<td>6,062</td>
<td>1.538</td>
<td>109</td>
<td>49</td>
<td>10.293</td>
</tr>
<tr>
<td>12</td>
<td>for</td>
<td>878,727</td>
<td>222.954</td>
<td>436</td>
<td>286</td>
<td>10.2031</td>
</tr>
<tr>
<td>13</td>
<td>within</td>
<td>45,546</td>
<td>11.556</td>
<td>122</td>
<td>86</td>
<td>9.9991</td>
</tr>
<tr>
<td>14</td>
<td>assume</td>
<td>4,052</td>
<td>1.028</td>
<td>90</td>
<td>73</td>
<td>9.3785</td>
</tr>
<tr>
<td>15</td>
<td>price</td>
<td>18,315</td>
<td>4.647</td>
<td>84</td>
<td>72</td>
<td>8.6581</td>
</tr>
<tr>
<td>16</td>
<td>take</td>
<td>69,531</td>
<td>17.642</td>
<td>107</td>
<td>60</td>
<td>8.6386</td>
</tr>
<tr>
<td>17</td>
<td>fair</td>
<td>8,963</td>
<td>2.274</td>
<td>77</td>
<td>48</td>
<td>8.5158</td>
</tr>
<tr>
<td>18</td>
<td>steps</td>
<td>6,634</td>
<td>1.683</td>
<td>71</td>
<td>39</td>
<td>8.2264</td>
</tr>
<tr>
<td>19</td>
<td>beyond</td>
<td>11,346</td>
<td>2.879</td>
<td>73</td>
<td>60</td>
<td>8.2071</td>
</tr>
</tbody>
</table>

Figure 12. T-scores for the word reasonable.
measure of the *strength* of collocation. The higher the measure, the stronger the link between the two items. Hunston (2002:71) claims that an MI score of 3 or higher could be taken as proof that the two items are collocates. The t-score measures the strength of association between the two items, and in particular the confidence with which we can claim that there is some association (Church and Hanks 1990). The t-score takes the size of the corpus in which the items occur into consideration. A t-score of 2 or higher is normally considered to be statistically significant (McEnery et al. 2006: 56).

The z-score is the number of standard deviations from the mean frequency. A higher z-score indicates a greater degree of collocability of an item with the node word. The z-score is used relatively less frequently than the MI test in corpus linguistics (McEnery et al. 2006: 57).

Log-likelihood is a statistical test to compare the fit of two models; in informal contexts, “likelihood” is used as a synonym for “probability”. The log-likelihood test is based on the likelihood ratio, which expresses how many times the data under one model is more likely than under the other (see also McEnery 2006: 55).

Corpora and computer software can also be useful in many other ways. Semantic prosody is another aspect of meaning that can be studied using corpora. The term semantic prosody was first introduced by Louw (1993), developed by Sinclair (2004) and used by many other scholars (Stubbs 1996, 2001, Tognini-Bonelli 2001, Partington 1998, 2004, Hunston 2007). Louw defines semantic prosody as “a consistent aura of meaning with which a form is imbued by its collocates”. Semantic prosodies, as Louw argues (1993: 156), are inaccessible to human intuition and cannot be retrieved through introspection. Only big corpora and the study of concordance lines can reveal the fact that many words occur frequently in recurring sequences which have in common a particular attitudinal meaning. Well-known examples include Sinclair’s analysis of *set in* (1987: 155-156):

The most striking feature of this phrasal verb is the nature of the subjects. In general they refer to unpleasant states of affairs. The main vocabulary is *rot* (3), *decay, malaise, despair, ill-will, decadence, impoverishment, infection, prejudice, vicious circle, rigor mortis, numbness, bitterness, mannerism, anticlimax, anarchy, disillusion, disillusionment, slump*. Not one of them is desirable or attractive.
## 5. CORPUS TOOLS AND DATA ANALYSIS

---

<table>
<thead>
<tr>
<th>No.</th>
<th>Word</th>
<th>Total No. in whole BNC</th>
<th>Expected collocate frequency</th>
<th>Observed collocate frequency</th>
<th>In No. of texts</th>
<th>Z-score value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>assume</td>
<td>4,052</td>
<td>1.028</td>
<td>90</td>
<td>73</td>
<td>87.255</td>
</tr>
<tr>
<td>2</td>
<td>grounds</td>
<td>6,062</td>
<td>1.538</td>
<td>109</td>
<td>49</td>
<td>86.2464</td>
</tr>
<tr>
<td>3</td>
<td>precautions</td>
<td>577</td>
<td>0.146</td>
<td>30</td>
<td>20</td>
<td>76.7173</td>
</tr>
<tr>
<td>4</td>
<td>foreseeability</td>
<td>21</td>
<td>0.005</td>
<td>6</td>
<td>3</td>
<td>75.2752</td>
</tr>
<tr>
<td>5</td>
<td>perfectly</td>
<td>4,310</td>
<td>1.094</td>
<td>64</td>
<td>59</td>
<td>59.6775</td>
</tr>
<tr>
<td>6</td>
<td>seems</td>
<td>20,738</td>
<td>5.262</td>
<td>129</td>
<td>109</td>
<td>53.7257</td>
</tr>
<tr>
<td>7</td>
<td>steps</td>
<td>6,634</td>
<td>1.683</td>
<td>71</td>
<td>39</td>
<td>53.0428</td>
</tr>
<tr>
<td>8</td>
<td>a</td>
<td>2,164,238</td>
<td>549.118</td>
<td>1738</td>
<td>928</td>
<td>50.7134</td>
</tr>
<tr>
<td>9</td>
<td>believing</td>
<td>1,360</td>
<td>0.345</td>
<td>30</td>
<td>15</td>
<td>49.6321</td>
</tr>
<tr>
<td>10</td>
<td>excuse</td>
<td>2,934</td>
<td>0.744</td>
<td>44</td>
<td>16</td>
<td>49.5544</td>
</tr>
<tr>
<td>11</td>
<td>fail*</td>
<td>8,963</td>
<td>2.274</td>
<td>77</td>
<td>48</td>
<td>49.2208</td>
</tr>
<tr>
<td>12</td>
<td>care</td>
<td>24,369</td>
<td>6.183</td>
<td>129</td>
<td>49</td>
<td>49.1912</td>
</tr>
<tr>
<td>13</td>
<td>endeavours</td>
<td>338</td>
<td>0.086</td>
<td>13</td>
<td>6</td>
<td>42.3918</td>
</tr>
<tr>
<td>14</td>
<td>suspicion</td>
<td>1,562</td>
<td>0.396</td>
<td>27</td>
<td>13</td>
<td>41.4649</td>
</tr>
<tr>
<td>15</td>
<td>beyond</td>
<td>11,346</td>
<td>2.879</td>
<td>73</td>
<td>60</td>
<td>41.0336</td>
</tr>
<tr>
<td>16</td>
<td>expect</td>
<td>10,269</td>
<td>2.605</td>
<td>68</td>
<td>56</td>
<td>40.2035</td>
</tr>
<tr>
<td>17</td>
<td>doubt</td>
<td>11,883</td>
<td>3.015</td>
<td>71</td>
<td>56</td>
<td>38.8655</td>
</tr>
<tr>
<td>18</td>
<td>firmness</td>
<td>151</td>
<td>0.038</td>
<td>8</td>
<td>3</td>
<td>38.1213</td>
</tr>
<tr>
<td>19</td>
<td>expectation</td>
<td>1,285</td>
<td>0.326</td>
<td>22</td>
<td>20</td>
<td>37.0826</td>
</tr>
</tbody>
</table>

---

Figure 13. Z-scores for the word `reasonable`. 
<table>
<thead>
<tr>
<th>No.</th>
<th>Word</th>
<th>Total No. in whole BNC</th>
<th>Expected collocate frequency</th>
<th>Observed collocate frequency</th>
<th>In No. of texts</th>
<th>Log-likelihood value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>2,164,238</td>
<td>549.118</td>
<td>1738</td>
<td>928</td>
<td>1679.2764</td>
</tr>
<tr>
<td>2</td>
<td>grounds</td>
<td>6,062</td>
<td>1.538</td>
<td>109</td>
<td>49</td>
<td>716.258</td>
</tr>
<tr>
<td>3</td>
<td>to</td>
<td>2,593,729</td>
<td>658.090</td>
<td>1409</td>
<td>677</td>
<td>664.2076</td>
</tr>
<tr>
<td>4</td>
<td>assume</td>
<td>4,052</td>
<td>1.028</td>
<td>90</td>
<td>73</td>
<td>629.2831</td>
</tr>
<tr>
<td>5</td>
<td>seems</td>
<td>20,738</td>
<td>5.262</td>
<td>129</td>
<td>109</td>
<td>579.2362</td>
</tr>
<tr>
<td>6</td>
<td>is</td>
<td>990,281</td>
<td>251.258</td>
<td>716</td>
<td>452</td>
<td>578.0315</td>
</tr>
<tr>
<td>7</td>
<td>it</td>
<td>1,054,279</td>
<td>267.495</td>
<td>737</td>
<td>477</td>
<td>562.947</td>
</tr>
<tr>
<td>8</td>
<td>care</td>
<td>24,369</td>
<td>6.183</td>
<td>129</td>
<td>49</td>
<td>539.3242</td>
</tr>
<tr>
<td>9</td>
<td>be</td>
<td>650,082</td>
<td>164.941</td>
<td>481</td>
<td>294</td>
<td>401.1908</td>
</tr>
<tr>
<td>10</td>
<td>perfectly</td>
<td>4,310</td>
<td>1.094</td>
<td>64</td>
<td>59</td>
<td>396.1395</td>
</tr>
<tr>
<td>11</td>
<td>fair</td>
<td>8,963</td>
<td>2.274</td>
<td>77</td>
<td>48</td>
<td>393.7901</td>
</tr>
<tr>
<td>12</td>
<td>steps</td>
<td>6,634</td>
<td>1.683</td>
<td>21</td>
<td>39</td>
<td>393.6239</td>
</tr>
<tr>
<td>13</td>
<td>that</td>
<td>1,118,985</td>
<td>283.913</td>
<td>654</td>
<td>431</td>
<td>356.2823</td>
</tr>
<tr>
<td>14</td>
<td>within</td>
<td>45,546</td>
<td>11.556</td>
<td>122</td>
<td>86</td>
<td>354.8712</td>
</tr>
<tr>
<td>15</td>
<td>beyond</td>
<td>11,346</td>
<td>2.879</td>
<td>73</td>
<td>60</td>
<td>332.3982</td>
</tr>
<tr>
<td>16</td>
<td>price</td>
<td>18,315</td>
<td>4.647</td>
<td>84</td>
<td>72</td>
<td>328.1538</td>
</tr>
<tr>
<td>17</td>
<td>expect</td>
<td>10,269</td>
<td>2.605</td>
<td>68</td>
<td>56</td>
<td>313.3957</td>
</tr>
<tr>
<td>18</td>
<td>doubt</td>
<td>11,883</td>
<td>3.015</td>
<td>21</td>
<td>56</td>
<td>313.172</td>
</tr>
<tr>
<td>19</td>
<td>excuse</td>
<td>2,934</td>
<td>0.744</td>
<td>44</td>
<td>16</td>
<td>273.177</td>
</tr>
</tbody>
</table>

Figure 14. Log-likelihood scores for the word *reasonable*. 
Other known examples include Stubbs’ *cause* (1996: 173-4), where he claims that *cause* has ‘a negative (or unfavourable) semantic prosody’, suggesting that an association with evaluatively negative things is a property of the verb. Thus, *cause* is used with *anger, amazement, damage, dizziness, misery, outrage, problems, pain, concern, etc.* On the other hand, the verb *provide* tends to collocate with positive or neutral words such as *insights, impetus, assistance, clues, opportunities, stimulus, facilities, shelter, information, details, etc.*

Although these are very interesting observations as such, the concept of negative or positive semantic prosody seems to be a somewhat simplistic view of attitudinal meaning. Such meaning is often not reducible to a simple ‘positive or negative’ (Hunston 2007: 256). Ascribing semantic prosody to a word is over-simplistic. The attitudinal meaning of a word may be altered by its immediate phraseology. Sinclair uses the term ‘semantic prosody’ in a more subtle way, referring not to simple co-occurrence but to the consistent discourse function of the unit formed by a series of co-occurrences: ‘the unit of meaning’. Sinclair (1996a) in identifying and defining the meaning of words takes into account their collocational and colligational patterns as well as their pragmatic implications. Thus, to define word meaning, or what Sinclair calls “the extended unit of meaning”, one needs to take the following methodological steps:

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

Looking at the word *budge*, Sinclair (2004: 142-147) notes that it usually occurs in an extended sequence that includes:

1. Negative lexical and lexico-grammatical patterns: *she could not budge, the snake will not budge, it will not budge, horses could not budge, the doors did not budge, the British would not budge, they won't budge, he would not budge.*
2. Determination, unwillingness not to move: *he'll refuse to budge, none of us will be able to budge tomorrow, he refuses to budge, he knew he couldn't budge.*

This inability or determination not to budge causes frustration (or similar emotion), despite pressure being applied. This sense of frustration is a more complex concept than simple negative evaluation, and it is not only achieved through the meaning of the word *budge*, but also belongs to the whole unit of meaning. The semantic prosody
of an item is the reason why the word is chosen – we choose *budge* rather than *move* when we want to show that something does not move despite attempts to move it.

**Research points:**


**A mini-research 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 the extended unit 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)

Use the BNC and study the following groups of words:

ambivalent, ambiguous
allay, alleviate, assuage, relieve
apparent, illusionary, seeming, ostensible
arbitrate, mediate
assume, presume
avenge, revenge
barbaric, barbarous
between, among
born, borne
concise, terse, succinct, laconic, pithy
conclusive, decisive, determinative, definitive
contrary, converse, opposite, reverse
dominant, predominant, paramount, preponderant
doubtful, dubious, problematic, questionable
effective, effectual, efficient, efficacious
empathy, sympathy, compassion, pity, commiseration
fickle, flexible
fractious, factitious, fractious
healthy, healthful, salutary
imply, infer, insinuate
sparing, frugal, thrifty, economical

Using the BNC, study the language specific, culturally determined collocations of superordinates, e.g. colours (black, blue, red, white, green), animals (ass, bee, butterflies, cat, donkey, horse, lark).
LEARNER CORPORA. 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 the 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 covering only one L1 background. The HKUST Corpus of Learner English is a ten million-word corpus that contains written
essays and examination scripts of Chinese learners of English at the University of Hong Kong. The JEFL (Japanese EFL Learner) corpus is a one million-word 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 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 and 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 (Granger et al. 2009). 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 ICLE 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 revealed through the overuse or underuse of words or structures with respect to the target language norm.

One of the aims of the ICLE project was to compile localized corpora of learner English, the Lithuanian Learner of English Corpus (LICLE) among them. The compilation of the Lithuanian component of ICLE (LICLE), which will also be included in the ICLE database, was completed in 2011. It contains advanced student essays from Vilnius University and Kaunas Vytautas Magnus University. The size of LICLE is 240,332 words (420 essays): 175,990 words (305 essays from Vilnius University) and 65,342 (115 essays) from Kaunas Vytautas Magnus University. The corpus contains argumentative essays on a variety of topics (for more see: Grigaliūnienė et al. 2008).
The success of the ICLE project prompted the launch of its spoken counterpart, the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin et al. 2010) as well as its Lithuanian component (LINDSEI-LITH). The data for the LINDSEI-LITH are collected using the LINDSEI format: the interviews are of approximately the same length – 2000 words each and follow the same pattern: they start with an informal discussion of university life, hobbies, travel or future plans. Then the interviewees are asked to choose one of three topics: an experience that taught them a lesson; a country which impressed them, a film or a play which they particularly liked or disliked. The students are asked not to make any notes and to speak without preparation – this is done for the sake of spontaneity. Each interview ends with a short picture-based story telling. The interviews are then transcribed using an orthographic transcription scheme.

The compilation of ICLE and LINDSEI and their localized corpora has given rise to a plethora of studies in the field of learner language research. 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 describing 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) such as the 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 permit the analysis of 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 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.

**Corpora and language learning and teaching**

Although 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 the 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 ways to reach students and teachers with information about corpora and what they can do. On the one hand, this 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). Learners who use corpora in language learning are encouraged to become more autonomous and to take more responsibility for their own learning, which in turn is conducive to increased motivation and learning effectiveness (Bernardini 2004: 27).

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, since they do not automatically provide straightforward understandable answers. Students should also be familiarized with inductive methods, otherwise they will find corpora boring and difficult. Teachers should introduce students to corpus analysis by specially prepared exercises that 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. The students should start using corpora for solving very practical language issues, such as the use of articles, countable/uncountable pairs of words, etc. The students can be asked to determine the difference in meaning between the noun as countable and as uncountable and then try to complete the ‘gapped’ citations to see if they have formed the correct hypothesis.
### Language

```
“Guidelines tell us we cannot 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.

A Flocks, elected chair, explained that he had for sometime been anxious as to where the society was drifting.

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.

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.

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.

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.

Much of the technical literature on the subject seems to confuse the two sets of questions distinguished in this section.

Forsyth is a top-selling and stylish author — decidedly not a purveyor of great literature, but a man writing for men, with thrills guaranteed.

---

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.

Using learner corpora in the classroom is a very new practice and before it becomes standard it still 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). For example, the compilation of Lithuanian learner of English corpora (LICLE and LINDSEI-LITH) made it possible not only to research various aspects of the English writing and speech of Lithuanian learners (Grigaliūnienė and Juknevičienė 2011, 2012, 2013), but also to use empirical data from LICLE and LINDSEI-LITH in language classes. Thus, for example, discussing the issue of what makes learner language sound unnatural and different from native speaker English, learners are exposed to comparative data from learner corpora (both LICLE and LINDSEI-LITH) and native speaker corpora (LOCNESS and LOCNEC), which can provide information on specific problem areas for English learners in Lithuania.

**Corpora and materials design**

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 comparing 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 books 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. Moreover, 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**


Study and discuss the data on the use (overuse and underuse) of connectors in native and non-native speaker writing.

**Table 3.** Non-native speaker overuse of connectors: raw frequencies

<table>
<thead>
<tr>
<th>Connector</th>
<th>NNS*</th>
<th>NS**</th>
</tr>
</thead>
<tbody>
<tr>
<td>actually</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>indeed</td>
<td>65</td>
<td>16</td>
</tr>
<tr>
<td>of course</td>
<td>53</td>
<td>20</td>
</tr>
<tr>
<td>moreover</td>
<td>52</td>
<td>1</td>
</tr>
<tr>
<td>e.g</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>for instance</td>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>namely</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>on the contrary</td>
<td>34</td>
<td>3</td>
</tr>
</tbody>
</table>

(Taken from Granger 1998)

*NNS – non-native speaker
**NS – native speaker

**Table 4.** Non-native speaker underuse of connectors: raw frequencies

<table>
<thead>
<tr>
<th>Connector</th>
<th>NNS</th>
<th>NS</th>
</tr>
</thead>
<tbody>
<tr>
<td>however</td>
<td>47</td>
<td>197</td>
</tr>
<tr>
<td>instead</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>though</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>yet</td>
<td>17</td>
<td>46</td>
</tr>
<tr>
<td>hence</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>then</td>
<td>27</td>
<td>65</td>
</tr>
<tr>
<td>therefore</td>
<td>33</td>
<td>123</td>
</tr>
<tr>
<td>thus</td>
<td>36</td>
<td>56</td>
</tr>
</tbody>
</table>

(Taken from Granger 1998)
Study the lists of formulaic sequences in LICLE and LINDSEI-LITH and make detailed comparisons of learner writing and speech.

**List 1.** Formulaic sequences in LICLE and their normalized frequencies per 100,000 words.
For example (80), in order to (72), according to (70), able to (68), because of (61), of course (47), as well (45), as well as (44), first of all (41), on the other hand (41), the fact that (40), more and more (37), in conclusion (34), in addition (32), what is more (31), at all (30), for instance (30), this way (29), used to (27), at least (26), even though (26), more than (26), deal with (24), to sum up (24), in my opinion (24), at the same time (21), instead of (21), due to (20), even if (20), going to (18), in other words (17), all over (16), no longer (15), in fact (14), out of (14), I believe (14), so much (13), up to (13), all in all (12), I think (12), in mind (12), in public (12), in the future (12), as if (12), on the contrary (12), in front of (11), take place (11), at first (10), in general (10), no need (10), find out (9), thing is that (9), no matter (9), in favour (8), start with (8), by law (7), rather than (7), take into consideration (7), after all (7), and so on (7), apart from (7), as much as (7), on the one hand (7), as soon as (6), once again (6), that way (6), at the moment (5), carried out (5), in reality (5), in terms of (5), take care of (5), go through (5), in case (5), in comparison (5), so far (5), supposed to (5), as such (5), away from (5), go on (5), grow up (5), make sense (5), on their own (5), all around (4), all things considered (4), far from (4), for the sake of (4), no wonder (4), what about (4), as far as (3), get to know (3), goes without saying (3), in future (3), in spite of (3), in turn (3), no doubt (3), put of (3), step by step (3), take into account (3), take part in (3), take up (3), what if (3), at once (3), come out (3), come up with (3), in the course of time (3), in the end (3), later on (3), look through (3), needless to say (3), a matter of fact (2), as for (2), at night (2), bring up (2), each and every (2), ever since (2), for granted (2), get rid of (2), give up (2), I guess (2), in common (2), in love (2), in search of (2), in the field of (2), last but not least (2), like that (2), little by little (2), on purpose (2), on the way (2), on time (2), put down (2), take over (2), I know (2), I mean (2), in contrast (2), in practice (2), in the public eye (2), looked upon (2), make up (2), more like (2), pass away (2), stand up (2), to a certain extent (2), to point out (2), turn on (2), when it comes to (2).

**List 2.** Formulaic sequences in LINDSEI-LITH and their normalized frequencies per 100,000 words.
I think (356), I don't know (264), you know (136), of course (103), for example (85), I guess (69), very much (60), I mean (56), going to (54), thank you (54), as well (49),
because of (49), used to (40), kind of (38), something like that (38), in general (36), I suppose (34), at first (31), first of all (29), in reality (29), like that (27), so much (27), and so on (25), in the end (25), at the same time (20), in particular (20), more than (20), at least (18), bit more (18), even though (18), away from (16), out of (16), thing is that (16), able to (13), in the future (13), the fact that (13), this way (13), at all (11), more like (11), sort of (11), and stuff like that (9), and that’s it (9), as well as (9), I know (9), in love (9), later on (9), one day (9), so far (9), take care of (9), that way (9), that’s all (9), what is more (9), after all (7), as far as (7), at night (7), changed my mind (7), even if (7), get to know (7), good question (7), once again (7), supposed to (7), take place (7), a difficult question (4), again and again (4), all around (4), all in all (4), all the way (4), as if (4), as much as (4), at last (4), but most of all I (4), but then (4), far from (4), find out (4), get out (4), go through (4), I believe (4), make up (4), up to (4).

(taken from Grigaliūnienė and Juknevičienė 2013: 222)

Practical assignments:

Using the BNC, carry out a contrastive study of words frequently confused by non-native speakers:

- adverse  averse
- biannual  biennial
- broach  brooch
- cession  session
- compliment  complement
- definite  definitive
- disinterested  uninterested
- disposal  disposition
- distinct  distinctive
- economic  economical
- elicit  illicit
- ingenious  ingenuous
- lay  lie
- lightening  lightning
7. CORPORA AND TRANSLATION STUDIES

The use of corpora in translation studies is relatively new, having been first advocated by Mona Baker in 1993 (Baker 1993, Olohan 2006). The perception of translation has not been very favourable in linguistics, corpus linguistics included, translations were not considered as representing language use and were excluded from general language corpora (e.g. BNC). Linguists viewed translations with suspicion, assuming them to be ontologically different from non-translated texts and referred to them as ‘interlanguage’ (Selinker 1972), ‘third language’ (Duff 1981), ‘third code’ (Frawley 1984), or ‘translationese’ (e.g. Gellerstam 1986, Doherty 1998, Mauranen 1999, Tirkkonen-Condit 2002). Nowadays, however, the situation is changing and the use of corpora is growing rapidly in translation studies research, translator training and translation practice. The compilation of parallel corpora proved to be an extremely valuable source of information in translation studies.

Parallel and comparable corpora

A parallel corpus is a corpus composed of source texts and their translations in one or more different languages (e.g. the works of Shakespeare and their translations into Lithuanian); parallel corpora can be aligned at a word, phrase or sentence level thus establishing correspondences between units of bilingual or multilingual texts.

There is some variation and some confusion in the definition of parallel corpora. Some linguists, for example Aijmer et al 1996, Granger 1996, argue for the term ‘translation corpus’, but this term is sometimes used to describe corpora consisting of only original texts in two or more languages, and referred to by other authors (Johansson & Hasselgärd 1999) as ‘comparable corpora’, corpora consisting of original texts in two or more languages, matched by criteria such as the time of composition, text category or intended audience. The term ‘parallel corpus’ is used in this book since this usage is consistent with work done by other translation scholars.

A parallel corpus is a resource that instantly provides a translation for a linguistic item. Unfortunately, there are only a few publicly available parallel corpora since it is
difficult to obtain permission to use both a text and its translation. Parallel corpora are compiled mainly in business communication, in multilingual societies, such as the United Nations, NATO, the EU and in bilingual countries such as Canada. One of the publicly available parallel corpora is the OPUS project, which is a growing collection of translated texts from the web. It consists of several sub-corpora such as the European Central Bank subcorpus, the European Medicines Agency documents, the European Parliament proceedings, etc. The following concordances of the words *galioja* and *sustainability* come from the European Parliament Proceedings subcorpus (Figures 15–16.).

Such concordances give information about how a particular issue has been dealt with by other translators and they are therefore important resources for translation studies. As Aijmer and Altenberg (1996:12) noted, they can provide new insights into the languages compared, insights that cannot be obtained in studies of monolingual corpora, they can also be used for different comparative purposes and enhance our understanding of language-specific, typological and cultural differences as well as universal features, they can highlight differences between source texts and translations, and they can also be used for a number of practical applications in translation teaching.

Aligned parallel corpora can provide illuminating insights into the nature of translation, they can help to devise tools to aid translation, and they can also be used in training contexts. Parallel corpora can be unidirectional (e.g. from English into Lithuanian or from Lithuanian into English), or bidirectional (e.g. containing both English source texts with their Lithuanian translations as well as Lithuanian source texts with their English translations), or multidirectional (e.g. the same text with its English, German, French, Russian, Spanish, Italian, etc. versions).

Translators can also build their own parallel corpora, using source and target texts of their previous translations and aligning them on a sentence-by-sentence basis.

**Comparable corpora**

Comparable corpora are comparable original texts in two or more languages. They are monolingual corpora designed using the same sampling techniques, e.g. the Aarhus Corpus of Contract Law (McEnery 2006: 47), the International Corpus of
Tai galioja ne tik dirbamai žemei, bet ir viešosioms erdvėms, tokioms kaip parkai, žaidimų aikštėlės, sporto aikštės, takelai ir geležinkelio keliai.

Tas pats galioja ir teritoriniams susitarimams, kuriuos J. Kacin gana teisingai išdėstysavo pranešime, nes niekas nesuprasi, ko mes norime, jei to aškiai neišakysime.

Tas pats galioja kovai su terorizmu.

Dabar vėl mėginama pratęsti pereinamą į laikotarpį iki aštuonerių metų šalims, kuriose jau galioja aukšti apsaugos lygmenys.

Tačiau policijos darbe galioja tam tikros taisykletės.

Rezultatas buvo vertas pastangų, ir įrodo, kad afrikinčiai nėra pasirengę priimti visko, kas jiems brukama; tai galioja ir diplomatijos bei politikos srityse.

Norėčiau pasakyti, kad stojimo deryboms pradėti ir Komisijos rekomendacijai priimti galioja visų valstybių narių Taryboje vienbalsiškumo reikalavimas, todėl nors šį klausimą laikome dvišaliu, taip pat reikia atsižvelgti ir į politinią tikrovę.

Jei Europa yra erdvė, kuriuje galioja bendrosios vertybės, mes negalime nekreipti dėmesio į nerimą keliančias Turkijos valdžios institucijų tendencijas, susijusias su teisinės valstybės principais, saviraiškos laive ir mažumų teisių pagarba.

Gairių sistema šiuo metu jau galioja Komisijai ir Tarybai skirtimi žodiniams klausimams.

Parlamentas balsavo už REACH sistemą, kuri galioja šiuo metu ir saugo žmonių sveikatą bei aplinką nuo cheminių medžiagų keliamo pavojaus.

Aš pats maniau, kad pramongalutinį rezultatą susipinino ir stengėsi suardyti, bet bent jau šie plačiausios apimties pasaulyje teisės aktai dėl cheminių medžiagų galioja.

Taip pat gaila, kad pvz., nepateikus reikalingos informacijos, sutartis toliau galioja vienus metus.

Šalia Komisijos pasiūlymo geros kvalifikacijos imigrantams galioja privalomi nacionaliniai reikalavimai, taip pat gali būti taikomi ir papildomi reikalavimai.

Bus ir kitų, kuriems bus netgi dar sunkiau pradėti dirbt arba gauti paaukštinimą, leisiantį užsidirbti pensiją; visi šie dabartinės situacijos aspektai galioja ne vienam laikotarpiui.

Antra, iki šiol visada vadovavomės nuostata - nors pastaruoju metu daugiau dvejų metų laikotarpiu buvo kitaip - kad klaidų toleravimas yra nulinis ir tai galioja visų rūšių kliaudoms.

Šis apsisprendimas dar galioja.

Mūsų manymą, dar kyla abejonių, ar iš tiesų galioja teispateikti apeliacinį skundą, jei vairuotojas nesutinka su jam paskirta nuobauda.

Pranešime yra bent du lygmenys: lygmuo, kuris yra visuotinis, nuolatinis ir galioja visose situacijose, ir lygmuo, kuris yra skubus atsakas i į dabartine krize.

Figure 15. Concordances for the word galioja taken from the Opus corpus.
That is true not only on farmland but also in public spaces such as parks, playgrounds, sports fields, pathways and railway tracks.

The same goes for territorial arrangements, which Mr Kacin has quite rightly brought into his report from the Commission, because no one will understand what we want if we pointlessly water it down.

The same applies to combating terrorism.

Now comes a new attempt to extend the grace period to eight years for countries where high levels of protection are currently in force.

It must be equally clear, however, that there are certain ground rules in police work.

This was all worthwhile, and it proves that Africans are not prepared to accept just anything when it comes to diplomacy, or politics for that matter.

I would like to point out that in order to start accession negotiations, in order for the Commission recommendation to be accepted, in the Council there is a requirement of unanimity of all EU Member States and therefore, even though we see this as a bilateral issue, we have to take this into account as a political reality.

If Europe is a space of shared values, we cannot close our eyes to the disturbing tendencies of the Turkish authorities in relation to the principles of the rule of law, freedom of expression and thought, or respect for minority rights.

A system of guidelines is currently in place for oral questions addressed to the Commission and to the Council and the Corbett report suggests a method which is already implemented by Parliament and well known by the Council and Commission.

The Parliament voted in favour of the REACH system which is now in place and which protects people’s health and the environment from the danger of chemicals.

I myself thought that the final result was watered down and lobbied to pieces by industry, but at least it is there, the world’s most wide-ranging chemicals legislation.

It is also regrettable, for example, that, if crucial information is not provided, a contract will still exist one year on.

Alongside the Commission proposal, national regulations remain in force for migrants with sound qualifications, and additional requirements are possible.

There will be others who will find it even more difficult to start their working lives or to find promotion to build up their pensions: all those aspects of what happens if you are not working for a period of time.

Secondly, we have until now - although the past two years or so have been different - always assumed that there is zero tolerance for any error.

That vote is, in fact, still valid.

In our opinion, there are also doubts about the effective right of appeal when the driver does not agree with the penalty applied.

The report involves at least two levels: that which is general, stable and valid in all situations, and that which is an emergency response to the current crisis.
The **sustainability** criteria must have a clearly defined place in this directive.

Reforms have been introduced to improve the **sustainability** of social protection systems and public accounts in many European countries and there has of course been a heightening of awareness of the need for us to take care of the environment in our model for growth.

We will not make compromises in connection with **sustainability**, either in social or environmental terms, and we therefore insist that fulfilment of the 10% target remains an obligation, and that this can be done in a sustainable way.

The **sustainability** scheme would therefore not be able to address trading practices such as the one referred to in the question from the honourable Member.

I would like to reflect a little on what Mr Hokmark said about developing the best possible system to ensure security of supply, **sustainability** and competitiveness.

The European Council of 19 and 20 June highlighted the importance of ensuring the **sustainability** of biofuel policies, for example by improving the production of first-generation biofuels and by encouraging the development of second-generation biofuels made from by-products.

Biodiesel can be produced in accordance with **sustainability** principles without endangering food supply - which you have referred to - and biodiesel is the only renewable energy where Europe holds a strong, wide leadership.

The **sustainability** of the pension system, aligned to both demographic and economic developments and based on risk-avoiding investment strategies in the long term, should be accorded high priority.

They contribute towards environmental **sustainability**, improved efficiency, enhanced safety and non-discriminatory competitiveness in road passenger and freight traffic.

Since issues as important as the regulation and supervision of financial markets and the **sustainability** of public finances were discussed, I would like to give special emphasis to the issue of sustainable economic growth and employment.

The ailing economy and investment in **sustainability** have been placed high on the agenda, and rightly so.

The report on the Green Paper insists on reforms that I believe to be necessary for better implementation of this policy on a regional level, particularly decentralisation of fisheries management to take account of the specific features of each region and to link in the **sustainability** of the sector through the preservation of species.

Figure 16. Concordances for the word *sustainability* taken from the Opus corpus.
It

Šioje direktyvoje turi būti aiškiai apibrėžti tvarios plėtros kriterijai.

Reformos skatina socialinės apsaugos sistemų nuoseklumą, o daugelyje Europos šalių rimčiau atsižvelgiama į poreikį atidžiau saugoti gamtą pagal mūsų ekonomikos augimo modelį.

Mes neisime į kompromisą dėl tvarumo, socialinio ar aplinkos, todėl reikalaujame, kad 10% tiksls vykdymas būtų privalomas ir to galima pasiekti tvariu būdu.

Todėl tvarumo schema nebus tinkamas atsakas kovoje su gerbiamojo Parlamento nario klausime nurodyta praktika.

Norėčiau Šiek tiek panagrinėti G. Hokmarko teiginį dėl tinkamsiausios sistemos vystymo tiekimo saugumui, tvariai plėtrai ir konkurencingumui užtikrinti.

Birželio 19 ir 20 dienomis Europos Vadovų Taryba nurodė, kaip svarbu užtikrinti biologinio kuro strategijos nuoseklumą, pavyzdžiui, siekiant efektyvesnės pirmosios kartos biologinio kuro gamybos ir skatinant plėtoti antrosios kartos biologinio kuro gamybą iš šalutinių produktų.

Biodyzelinas gali būti gaminamas laikantis tvarumo principų ir nekeliant grėsmės aprūpinimui maisto produktais, kurį jūs paminėjote, be to, biodyzelinas yra vienintelis iš atsinaujinančių energetikos Šaltinių gaminamas biokuras, kurio gamybos srityje Europa išlaiko stiprą ir plačią lyderystę.

Pensijų sistemų tvarumui, kartu su demografiniais ir ekonominiais postūmiais parentam rizikos vengiančiomis ilgalaikėmis investavimo strategijomis, turėtų būti skiriamas svarbus prioritetas.

Jos padeda palaikyti aplinkos tvarumą, gerinti keleivinio ir krovininio kelių transporto veiksmingumą, didinti jo saugumą ir nediskriminojantį konkurencingumą.

Kadangi buvo aptartų tokie svarbūs klausimai kaip finansų rinkų reguliavimas ir priežiūra bei viešųjų finansų tvarumas, norėčiau daugiau dėmesio skirti ekonominio augimo ir užimtumo klausimams.

Darbotvarkėje didelšvarba teikiama sergančiai ekonomikai ir investicijoms į tvarumą - ir teisingai daroma.

Pranešime dėl žaliosios knygos primygtinai raginama vykdyti reformas, kurios, mano manymu, reikalingos norint šią politiką geriau įgyvendinti regioniniu lygmeniu, pirmiausia decentralizuoti žvejybos valdymą siekiant atsižvelgti į konkrečius kiekvieno regiono ypatus ir tai susieti su sektoriaus tvarumui išsaugant rūšis.
English (ICE), which consists of individual regional corpora built to a common design or the International Corpus of Learner English, which comprises texts written by advanced learners of English as a foreign language from diverse mother tongue backgrounds. This kind of monolingual comparable corpus is particularly useful in studying intrinsic features of translations, improving the translator’s understanding of the subject domain, terminology and idiomatic expressions in the specific field.

In reality, there are also multilingual corpora, which are a mixture of parallel, and comparable corpora (e.g. the *English-Norwegian Parallel Corpus*, in spite of its name, can be seen as a combination of a parallel and comparable corpus (Xiao, 2012:241)).

Comparable corpora can provide a translator with information about the domain and genre as well as terminology in both the source and target languages.

Comparable and parallel corpora are related, but still different, types of corpora. Kenning (2010:487) argues that “what distinguishes parallel from comparable corpora is that parallel corpora imply a common source text”.

Comparable and parallel corpora are of great interest to translation studies, translator training, translation practitioners, language teachers and language learners, and in addition are often claimed to be complementary. Parallel corpora which deal with translated texts are sometimes viewed with suspicion because, as Teubert (1996:247) claims “working with translations means working with distorted mirror images of the source language in the medium of the target language, when the objective is to analyse and describe the language in its own right, not just a target language”. Because of that, parallel corpora should be just starting points that need to be supplemented. Comparable corpora that are collections of authentic, naturally occurring language can easily serve this function.

**Corpus-based translation research**

Corpus-based translation studies emerged as a new translation research paradigm in the 1990s and contributed a great deal to translation studies. Corpus translation studies provide access to very large quantities of data – more data than any single human being could ever manage to gather and investigate in a lifetime without electronic assistance. Corpus translation studies changed the content and methods
of the discipline of translation not only in quantitative, but also in qualitative ways. Corpus translation studies focus on both the process and the product of translation and contribute to the debates going on in the discipline. One of the most important debates in intellectual domains is connected with research on the universals of translation. The notions of ‘human universals’ and ‘linguistic universals’ are not new and have been notoriously contentious (Crystal 1987: 84-5). Baker (1993, 1996) is interested in patterns that are not the result of interference from either source or target language (Olohan 2006:92). She argues that these features are characteristic of any translated text and they do not vary across cultures, unlike norms of translation, which are considered to be social, cultural and historical (for more detailed discussion of the norms of translation see: Toury 1995, Hermans 1999). The translation universals are represented by explicitation, i.e. translations tend to be more explicit on different levels than the originals, simplification – when the content or form is simplified compared with non-translated texts and normalization, i.e. the language used in translations is more conventional and normalized than that of the originals (Olohan 2006:37). Corpus linguists try to approach translations objectively and rely on the properties that can be measured: average word length, average sentence length, lexical density, type-token ratio, etc.

Explicitation

Baker refers to explicitation as “an overall tendency to spell things out rather than leave them implicit in translation” (Baker 1996:180) and discusses the ways to research this feature. She argues that translations tend to be longer than their source texts, and this can be attributed to explicitation. This can be tested by using parallel corpora, comparing lengths of texts and text segments and analyzing the differences. Explicitation can be found at various linguistic levels ranging from lexis to syntax and textual organization. Syntactic and lexical explicitation can be investigated by using comparable corpora and looking into the frequency of explanatory vocabulary and conjunctions, e.g. cause, reason, due to, lead to, because, therefore, consequently (Baker 1996:181), in order to find out whether they were more frequently used in translations to make the relations between propositions more explicit.

The research of explicitation did not start with the emergence of corpus-based translation studies. The tendency to make translations more reader-friendly, more explicit and more conforming to target language norms was noted by many researchers
before the arrival of corpus-based translation studies. Thus, Vanderauwera (1985) studied the translations of some Dutch novels into English and commented on some of the ways in which cultural information is made explicit to target language readers: by using repetitions to avoid ambiguity, specifying references for disambiguation purposes, expanding condensed passages and filling out elliptical units (see Olohan 2006:93).

Other authors (e.g. Overas 1998) put forward a hypothesis that explicitation is norm dependent rather than a universal and concludes that only findings from different comparable corpus-based studies can lead to a better understanding of this phenomenon.

**Normalization and conventionalization**

Baker (1996:176-7) defines normalization as a “tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them”. The discussion of normalization focuses on typical collocational patterns, clichés, grammatical structures and punctuation. The issue of what is typical in a language can be best answered if based on corpora. Therefore, corpus-based research can provide a great deal of information on the most typical grammatical and collocational patterns in both source and target languages.

**Simplification**

There is a clear relation between normalization and simplification, as simplification is reflected in various strategies such as the breaking up of long sentences, omissions of redundant or repeated information, shortening of complex collocations, etc., which are aimed at adhering to target language norms and conventions. The simplification hypothesis, however, is controversial – evidence produced to support this hypothesis is still insufficient. Therefore, more studies should be carried out to throw new light on the universals of the translation process.

The computerized corpora can also enable us to explore a number of other theoretical issues that are difficult to deal with on the basis of small-scale studies.
Corpora in translator training

Recently there has been a growing interest in using corpora in translator training. Corpora may be integrated into translator training and may meet various needs of translator trainers.

Parallel corpora are especially useful as they can be used to retrieve terminology, explore collocations, phrasal patterns, lexical polysemy, translation of collocations and idioms, etc. (Botley et al. 2000). Students can also be encouraged to compile their own specific corpora that can be very useful for content information, terminology, and phraseology for some particular domains or topics. Such corpora are often referred to as DIY corpora or disposal corpora, i.e. they are compiled for a specific, one-off purpose and can be easily compiled from texts on the Internet. There are some dangers involved in using texts from the Internet as the quality and reliability of texts to be included needs to be carefully assessed (more on that in Zanettin 2003). A corpus compilation experiment can be carried out as a real-life translation assignment (Varantola 2003). Prior to corpus compilation, the students can be asked to speculate on the usefulness of their exercise, the difficulties that they might face and the decisions that they may have to make concerning text selection for their corpora. According to Varantola (2003: 266), “the main benefit that the students gained from corpus use was that they felt reassured in their strategic decisions and lexical choices, particularly when the corpus data provide justification for decisions that they perceive as more radical”.

Comparable corpora can also be helpful in translator training as they can be used to check terminology and collocates, identify text-type-specific formulations, validate intuitions and provide explanations for the appropriateness of certain solutions to problems (Pearson 2003).

A learner translation corpus is another type of corpus which may be relevant to translator training and could provide useful information for teaching and research. Learner corpora, consisting of student translations, can be easily compiled in the training process. Students submit their translations electronically and they can be put into a Student Translation Corpus, which can be used for the analysis of different aspects of student performance, help to identify common problems in student translations and to improve teaching contents and materials; longitudinal studies could track the students’ progress and look at the efficiency of different training methods and techniques.
Corpora in translation practice

Corpora have been around for quite a long time now, but their use by professional translators is still a relatively rare occurrence. Olohan (2006:176) noted that corpora can be potentially useful in the translation profession: specialized corpora can be used to familiarize translators with concepts and terms from a specific domain, translators can study corpora output to understand text-type conventions, or look up how a particular translation issue has been dealt with by other translators. Corpora can complement traditional reference sources such as dictionaries, which are often insufficient for professional translators. Corpora are not a universal panacea for the translator, but as translators spend a great deal of time looking for information of various kinds, they can provide a significant resource: readily consultable collections of texts, a basis for acquiring conceptual and terminological information, reassurances for checking their hunches, help in finding equivalents, reassurance in checking the readability, comprehensibility, coherence, grammaticality and terminological consistency of their translations (Kübeler 2010:501). Although professional translators understand the potential usefulness of the use of corpora, they do not have the time and the skills required to use corpora as translation resources. Therefore, learning how to use corpora as translation resources should be part of the curriculum of future translators and become part of their professional competence.

Discussion and research points

Study Olohan, M. 2006. *Introducing Corpora in Translation Studies.* London and New York: Routledge, 90-144 (‘Features of translation’) and discuss the features of translations, focusing on the findings of the studies and examples given. Provide your own examples.

Using the BNC, the Corpus of Contemporary Lithuanian and parallel (EN-LT) corpus, study the following words and their translation equivalents in Lithuanian:

- actual (topical, current) — patron (cartridge)
- alley (avenue) — physician (physicist)
- costume (suit) — preservative (condom)
- fabric (factory) — programme (TV) (channel)
- faction (fraction) — public (audience)
Using the BNC, complete the following idioms of comparison, then try to find the Lithuanian equivalents using the Corpus of Contemporary Lithuanian:

### Idioms of comparison

Complete the following:

<table>
<thead>
<tr>
<th>English</th>
<th>English</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>as black as ...</td>
<td>as gentle as ...</td>
<td>as quick as ...</td>
</tr>
<tr>
<td>as blind as ...</td>
<td>as good as ...</td>
<td>as quiet as ...</td>
</tr>
<tr>
<td>as bold as ...</td>
<td>as graceful as ...</td>
<td>as red as ...</td>
</tr>
<tr>
<td>as brave as ...</td>
<td>as grave as ...</td>
<td>as regular as ...</td>
</tr>
<tr>
<td>as bright as ...</td>
<td>as greedy as ...</td>
<td>as rich as ...</td>
</tr>
<tr>
<td>as brittle as ...</td>
<td>as green as ...</td>
<td>as ripe as ...</td>
</tr>
<tr>
<td>as brown as ...</td>
<td>as happy as ...</td>
<td>as round as ...</td>
</tr>
<tr>
<td>as busy as ...</td>
<td>as hard as ...</td>
<td>as salty as ...</td>
</tr>
<tr>
<td>as changeable as ...</td>
<td>as harmless as ...</td>
<td>as sharp as ...</td>
</tr>
<tr>
<td>as cheerful as ...</td>
<td>as heavy as ...</td>
<td>as silent as ...</td>
</tr>
<tr>
<td>as clear as ...</td>
<td>as hot as ...</td>
<td>as silly as ...</td>
</tr>
<tr>
<td>as cold as ...</td>
<td>as hungry as ...</td>
<td>as slender as ...</td>
</tr>
<tr>
<td>as cool as ...</td>
<td>as innocent as ...</td>
<td>as slippery as ...</td>
</tr>
<tr>
<td>as cunning as ...</td>
<td>as keen as ...</td>
<td>as smooth as ...</td>
</tr>
<tr>
<td>as dark as ...</td>
<td>as large as ...</td>
<td>as sober as ...</td>
</tr>
<tr>
<td>as dead as...</td>
<td>as light as ...</td>
<td>as soft as ...</td>
</tr>
<tr>
<td>as deaf as...</td>
<td>as like as ...</td>
<td>as sound as ...</td>
</tr>
<tr>
<td>as different as...</td>
<td>as loud as ...</td>
<td>as sour as ...</td>
</tr>
<tr>
<td>as drunk as ...</td>
<td>as mad as ...</td>
<td>as steady as...</td>
</tr>
<tr>
<td>as dry as ...</td>
<td>as merry as ...</td>
<td>as timid as ...</td>
</tr>
<tr>
<td>as dumb as ...</td>
<td>as mute as ...</td>
<td>as tough as ...</td>
</tr>
<tr>
<td>as easy as ...</td>
<td>as obstinate as ...</td>
<td>as tricky as ...</td>
</tr>
<tr>
<td>as fair as ...</td>
<td>as old as ...</td>
<td>as true as ...</td>
</tr>
<tr>
<td>as fat as ...</td>
<td>as pale as ...</td>
<td>as ugly as ...</td>
</tr>
<tr>
<td>as fierce as ...</td>
<td>as patient as ...</td>
<td>as vain as ...</td>
</tr>
<tr>
<td>as firm as ...</td>
<td>as plain as ...</td>
<td>as warm as ...</td>
</tr>
<tr>
<td>as fit as ...</td>
<td>as playful as ...</td>
<td>as watchful as ...</td>
</tr>
<tr>
<td>as flat as ...</td>
<td>as plentiful as ...</td>
<td>as weak as ...</td>
</tr>
<tr>
<td>as free as ...</td>
<td>as plump as ...</td>
<td>as wet as ...</td>
</tr>
<tr>
<td>as fresh as ...</td>
<td>as poor as ...</td>
<td>as white as ...</td>
</tr>
<tr>
<td>as gay as ...</td>
<td>as pretty as ...</td>
<td>as wise as ...</td>
</tr>
<tr>
<td>as gaudy as ...</td>
<td>as proud as ...</td>
<td>as yielding as ...</td>
</tr>
</tbody>
</table>
Alignment – tekstų paralelinimas, išlygiavimas
Alignment is the practice of defining explicit links between texts in a parallel corpus.

Annotation – anotavimas
Annotation is the practice of adding explicit additional information to machine-readable text; the matching or linking of a text and its translation(s), usually paragraph by paragraph and/or sentence by sentence.

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.
9. REFERENCES


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.

Birmingham’s Bank of English, currently a 650 million word corpus of present-day English.

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

The Brown Corpus

The Cambridge and Nottingham Corpus of Discourse in English (CANCODE) is part of the Cambridge International Corpus. The corpus comprises five million words of transcribed speech recorded in Britain and Ireland between 1994 and 2001, covering a wide range of mostly informal settings: casual conversation, people working together, people shopping, people finding out information, discussions and many other types of interaction.

The Child Language Data Exchange System (CHILDES) is an international database organized for the study of first and second language acquisition. The CHILDES database contains transcripts of data collected from children and adults who are learning both first and second languages. The total size of the database is approximately 20 million words, covering 25 languages.
The Corpus of Contemporary American English (COCA) – a 450 million word corpus. 1990 - present, both written and spoken. Freely searchable online.

The Corpus of Contemporary Lithuanian approximately 140 million words of contemporary Lithuanian. It is meant to represent current Lithuanian and it consists of printed material from Lithuania's independence period (since 1990). The larger part of the corpus is comprised of General Press, Popular press and Special Press; the rest of the corpus consists of fiction, non-fiction, memoirs, and various official texts.

The Corpus of Historical American English (COHA) 400 million word corpus of American English from 1810 to 2009.

The English-Norwegian Parallel Corpus (ENPC) is one of the earliest and best-known parallel corpora. The corpus is bi-directional in that it contains both original and translated texts in the two languages. The ENPC consists of 100 original texts between 10,000 and 15,000 words in length in English and Norwegian together with their corresponding translations, totalling 2.6 million words.

The English-Swedish Parallel Corpus (ESPC) follows the English-Norwegian Parallel Corpus in its design. The ESPC consists of 64 English texts samples and their translations into Swedish and 72 Swedish text samples and their translations into English, amounting to 2.8 million words.

The Freiburg-Brown Corpus of American English (FROWN) and The Freiburg-LOB Corpus of British English (FLOB) the compilation of both began in 1991. Both corpora were intended to match the BROWN and LOB corpora as closely as possible in size and composition, with the only difference that they should represent the language of the early 1990s.

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 words from British English texts from 850 – 1720; the dialectal includes 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 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.

The Michigan Corpus of Academic Spoken English (MICASE) is a collection of 1.8 million words of transcribed speech from a wide range of speech events (including lectures, classroom discussions, lab sessions, seminars, etc.). It was compiled at the University of Michigan in Ann Arbor by researchers and students.

The OPUS parallel corpus is a growing collection of translated texts from the web.
In the Opus project free online data is converted and aligned, linguistic annotations are added and a publicly available parallel corpus is provided. It consists of several subcorpora, such as the European Central Bank corpus, the European medicines Agency documents, the European Parliament proceedings, etc. It offers a range of languages.

The Survey of English Usage Corpus (SEU) is a corpus compiled at University College, London, under the direction of R. Quirk and S. Greenbaum. It contains one million words of British English (50 per cent spoken and 50 per cent written) from 1953 to 1987.
The Translation English Corpus (TEC) is a collection of contemporary translational English text. It consists of written texts translated into English from a variety of source languages, European and non-European. TEC consists of four subcorpora: fiction, biography, news and inflight magazines. The overall size of the corpus is currently around 10 million words.

WebCorp is a World Wide Web corpus, which allows access to a large collection of texts from the World Wide Web. It was developed by the Research and Development Unit for English Studies based in the School of English at Birmingham City university.
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 (CSPAЕ)**
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.

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


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.

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