LATVIJAS UNIVERSITĀTE

BAKALAURA DARBS

RĪGA 2008

UNIVERSITY OF LATVIA
FACULTY OF MODERN LANGUAGES
DEPARTMENT OF ENGLISH STUDIES

THE ROLE OF CORPUS EVIDENCE IN MODERN ENGLISH LEXICOGRAPHY

TEKSTU KORPUSU LOMA MODERNAJĀ ANGĻU LEKSIKOGRĀFIJĀ

BACHELOR THESIS

Author: **Ilona, Bulāne**

Matriculation Card # AngF 020314

Adviser: lect. Laura Karpinska

RIGA 2008

ABSTRACT

This research paper deals with the issue of corpus application in modern English lexicography. The main goal of this research paper is to analyse use of text corpora in learner's dictionaries and to investigate at what extent they help to produce real and authentic English language.

The method of analysis is to compare three learner's dictionaries which claim to describe and explain English language with the help of the evidence of corpus data and see how corpus evidence is used in practice.

The result of analysis proves that text corpora are reliable sources of linguistic data and provide true evidence of actual and real use of the English language and therefore are valued as best resources of compilation of more complete and comprehensive dictionaries for learners of the English language.

ANOTĀCIJA

Šis noslēguma darbs apraksta tekstu korpusu izmantošanu modernajā angļu leksikogrāfijā. Noslēguma darba mērķis ir analizēt tekstu korpusu piemērus vienvalodas angļu mācību vārdnīcās un izpētīt kā tekstu korpusi veicina angļu valodas apguvi.
Pētniecības darbā tiek izmantota salīdzinošā metode attiecībā uz tekstu korpusu atspoguļojumu angļu mācību vārdnīcas, lai noskaidrotu tekstu korpusu pielietojumu. Pētniecības gaitā tika noskaidrots, ka tekstu korpusi satur ticamus piemērus autentiskas angļu valodas lietošanai un tie ir piemēroti vienvalodas angļu mācību vārdnīcu sasdādīšanai.

TABLE OF CONTENTS:

List of abbreviations

ESL – English as a Second Language

ESF – English as a Foreign Language

LDOCE - Longman Dictionary of Contemporary English

COBUILD - Cobuild Advanced Learner's English Dictionary

Macmillan - Macmillan English Dictionary for Advanced Learners

INTRODUCTION

This research paper emphasizes the importance of corpus evidence in modern English lexicography and more specifically the importance of corpus use in ESL lexicography. Producing better dictionaries is the main concern of ESL lexicography and availability of large linguistic databases is essential in the creation of more comprehensive dictionaries for learners of the English language.

Nowadays contemporary learner's dictionaries are extensively based on the corpus data for they provide most reliable evidence on how English is actually used. The goal of this research paper therefore is to analyse how corpus evidence is applied in practical lexicography and how corpus evidence reflects in monolingual English learner's dictionaries.

Hypothesis:

Text corpora present an invaluable source of linguistic evidence for the creation of complete and comprehensive English learner's dictionaries for they reflect most reliable evidence of actual and real use of the English language.

To implement the analysis of corpus evidence in compiling of English learner's dictionaries the author of this research paper has undertaken the following steps:

- the literature on the subjects of Corpus linguistics and corpus application in lexicography was reviewed;
- for the analysis of corpus evidence three English learner's dictionaries were selected for they claim to describe and explain English language with the help of the evidence of corpus data. The following dictionaries were analysed:
  - Longman Dictionary of Contemporary English, fourth edition 2005 (*LDOCE*) – makes use of the Longman Corpus Network, a database of 300 million words;
  - Cobuild Advanced Learner's English Dictionary, fourth edition 2003 (*COBUILD*) – makes use of the Bank of English, corpus of 520 million words;
  - Macmillan English Dictionary for Advanced Learners, International student edition 2002 (*Macmillan*) – is compiled using the World English Corpus of 200 million words.
- the framework of analysis was outlined;
- criteria of analysis were defined.

The main literature sources used:

- S. Landau “Dictionaries, The Art and Craft of Lexicography”, who describes the history and development of text corpora;
- H. Jackson “Lexicography, An Introduction”, who specifies typical and distinctive features of general purpose dictionaries and learner’s dictionaries;
- C. Meyer “English Corpus Lexicography, An Introduction”, who describes issues in relation with corpus linguistics.

This research paper consists of 5 chapters:

Chapter 1 - deals with the issue on Corpus linguistics and text corpora.

Chapter 2 - deals with the use of text corpora in modern ESL lexicography.

Chapter 3 - deals with general purpose monolingual English dictionaries for learners of English.

Chapter 4 - describes the framework of analyses which is intended to analyse corpus application in English learner’s dictionaries.

Chapter 5 - is the analysis of corpus application in English learner’s dictionaries.

1. CORPUS LINGUISTICS – A LINGUISTIC DISCIPLINE BASED ON TEXT CORPORA

The main goal of *ESL* lexicography is to create better dictionaries for learners of English. Corpus linguistics is a discipline which enables the process of compiling dictionaries to be much more precise and much more productive. Having access to huge linguistic databases lexicographers can obtain more reliable information on lexical items, their various meanings and possible constructions they form or participate.

The overall objective of this chapter is to present some general information on the Corpus linguistics, to define a general notion of text corpora, to state their typical definitions and describe their characteristic features, as well as emphasize their various types and possible application within different linguistic fields of research. The issue of corpus use in lexicography is addressed in separate chapter which describes the ways how electronic corpus is applied, consulted and analysed in modern lexicography.

1.1. Corpus linguistics and lexicography

It is generally accepted that Corpus linguistics has emerged in 1960s, at the time when the Brown Corpus - the first computer corpus was finished. It was time when advances in computer technologies made it possible to create "machine-readable corpora" (Leech, 2002: 3) and store large amounts of linguistic data. Availability of huge amounts of data in electronic form brought about significant changes in lexicography, lexicographers had access to more "reliable linguistic evidence" and they could base their decisions on "naturally occurring" texts rather than depend on their intuition or assumptions (Landau, 2001: 286).

Although linguists claim that Corpus linguistics is not a branch of linguistics in its own right but is rather "an approach to language" (Teubert, 2007a: 50) or more "a way of doing linguistics" (Meyer, 2002: xi), advances in computer technologies with subsequent creation of text corpora have made the Corpus linguistics a "powerful linguistic discipline", which according to Granger (2002: 45) has the potential "to change perspectives on language". In Granger's opinion the suggested potential of Corpus linguistics implies the discovery of new facts and "far-reaching new hypotheses" about language and discovery of previously "unsuspected linguistic phenomena" (ibid.). What Granger suggests is that Corpus linguistics makes use of huge text corpora which allow implementation of more complex studies of

language and corpus evidence is what makes lexicographers to observe new linguistic phenomena and produce better and more-up-to date dictionaries.

Advances in computer technologies and creation of electronic text corpora turned out to be the key aspects in the changes which emerged within the field of lexicography during the past few decades. They enabled commencement of such corpus-based studies which played an immense role in ESL lexicography and brought about a number of innovations.

With the advent of large corpora one of the innovations lexicographers encountered with was in its capacity to provide evidence on typical contexts in which words were used and lexicographers could associate different meanings with different contexts (Hanks, 2002: 157). What Hanks suggests is that dictionaries for learners need to show the proportion of senses when different meanings of words are in question and shall demonstrate the pragmatic aspects of words in use. He suggests application of so called syntagmatic approach – where the pragmatic aspects of words are addressed in terms of syntax and context rather than directly in terms of semantics (2002: 159). Such syntagmatic approach in lexicographical study was not possible until very large corpora became available.

Other innovations with electronic corpora were such that everything could be counted or compared; for the first time it was possible to come up with the lists of the most frequent words and phrases. Among the many advantages of using a corpus in lexicography, in Teubert's opinion, frequency counts are the most important (2007a: 55).

In conclusion it shall be added that the text corpora are the resource of linguistic data and provide a great variety of contexts, they are invaluable in compiling contemporary dictionaries for learners of English. They denote certain notion, differ in types and scope and apart from lexicography are used in many other fields of linguistics.

1.2. Text Corpora as resources of linguistic data

To start with the issue of the general notion of text corpora it should be mentioned that the term 'corpus' was first introduce by Nelson Francis, one of the originators of the Brown Corpus, who used it when he first referred to his electronic collection of texts (Teubert, 2007a: 53).

1.2.1. The notion of text corpora

A general notion of text corpora denotes the idea of the entity or the basis on which the study of language is implemented. Text corpora and Corpus linguistics are interrelated notions, the

former being a collection of language material while the latter of them – a linguistic discipline. This statement can be supported by the following definitions:

> "Corpus linguistics can be described as the study of language on the basis of text corpora" (Aijmer, 1991: 1);
>
> "At its most general, a corpus (plural *corpora*) may be defined as a body or collection of linguistic data for use in scholarship and research" (Leech, 2002: 3);
>
> "In language study, a corpus is any body of text collected with the aim of analyzing its features" (Landau, 2001: 273).

A typical definition of text corpus denotes the notion of corpus (from Latin *body*) being: "a collection of texts or parts of texts" (Meyer: 2002), collection of "real language data" (referred to Svartvik in Aijmer's: 1991), "a collection of naturally occurring language texts" (Teubert: 2007a), "systematic collections of samples" (Leech: 1991), or "a body of electronically encoded texts" (Biber *et al*: 1998).

In some way corpora are similar to text databases or archives of texts, except that they differ in terms of scope and purpose they were created for. As Leech (1991: 10-11) suggests "[text archives] are collected more or less opportunistically, according to what sources of data can be made available", but a corpus "is designed or required for a particular 'representative' function". Baker (2006: 26) claims that "corpora tend towards having a more balanced, carefully thought-out collection of texts that are representative of a language variety or genre".

More specific definition of text corpus, as suggested by Aarts, is one of those metaphoric ones, which embodies the following idea: "Corpus data reflect the way in which language is actually used ... what one finds in a corpus is 'performance', that is, evidence of linguistic behaviour" (2002: 63). Text corpora contain real language data – providing evidence of actual instances of speech or writing rather than consisting of made up or artificially compiled data (Meyer, 2002: xiii). Availability of real language instances is an essential feature of text corpora and is particularly appreciated in ESL lexicography for they provide evidence on how words are actually used.

From the above listed definitions it can be concluded that the idea of text corpus first denotes its integrity with Corpus linguistics, then the idea of the entity which is made up of carefully selected texts and examples rather that examples chosen at random. Text corpus represents instances of actual or real language use and enables lexicographers to contextualize their analysis of language (Meyer, 2002: 6), this allows them to see, for instance, how words behave in particular contexts.

It requires mention here that text corpora differ in size and scope depending on the research goals and objectives they are created for. For fuller description of the various types of text corpora see sub-chapter 1.3. of this research paper.

1.2.2. History of text corpora

The issue on the history of text corpora is addressed in the following 2 sub-chapters - one comprising the corpus studies implemented before the advent of computer technologies and the other comprising corpus studies based on the computer corpora. Such division is aimed at highlighting and distinguishing the characteristic features of pre-electronic and electronic corpus studies.

1.2.2.1. Pre-electronic corpus studies

Text corpora were created long before computers came into existence. For instance, there were attempts to investigate "position and possible frequency counts" of words in the texts of the Bible or the works of Shakespeare (Landau, 2001: 273). The use of authentic examples from "selected texts" was a typical feature of early corpus studies (Aijmer, 1991: 1), for as early as 1755 Johnson for his *Dictionary of the English Language* used a corpus of texts from which "authentic uses" of words were collected (Biber, 1998: 21).

Examples of authentic texts in the middle of 18th century formed a kind of a corpus, from which citations were used to extract the entry words of dictionaries, to look for "confirmation of [lexicographers'] intuition", and to collect quotations. The use of words from authentic texts and use of those texts as a source of quotations is regarded by Bejoint as "a capital innovation in lexicography" (2000: 97-8). Johnson in his *Dictionary of the English Language* used quotations for almost every entry word – his intention was to demonstrate the meanings of words in context, to "establish that a word had been used by a reputable authority" and most importantly "to lend authority to the dictionary" (ibid.). The use of quotations in this respect is similar to the approach of contemporary ESL lexicography – where definitions of entry words are exemplified by illustrative examples, which are taken form electronic corpus and denote typical contexts of words, their meanings and usage.

Although the authentic texts' approach of early lexicography was highly appreciated and well received yet it posed certain limitations if opposed with the practice of contemporary lexicography. For instance, citing of authentic literary works implies that examples of only best literary authors' are used. Examples being paragraph-long or sentence-long are limited in

context and are not representative of a language. Selection of authentic examples to confirm lexicographer's intuition largely depends on lexicographer's knowledge and proficiency to locate the meanings. While the authority of contemporary dictionaries is established and dictated by their coverage and accessibility – there is demand for them to be comprehensive or user-friendly in terms of their definitions and examples, which are based on the instances of actual or real language and are justified by corpus evidence.

To continue with the early corpus studies it should be noted that the first large-scale corpus of English compiled for lexical study in Landau's opinion was Edward L. Thorndike's word count of 4.5 million words, published as the *Teacher's Word Book* in 1921. Later this corpus was enlarged to a corpus of 18 million words, based on which *The Teacher's Word Book of 30,000 Words* was produced. This book was compiled basing on texts from magazines and juvenile reading, contained word lists of relative frequencies, and was designed as a helpful tool for teachers to determine which words "are common enough" to be used at particular grade levels (Landau, 2001: 273).

Another early corpus, mentioned by Landau, was Ernest Horn's *A Basic Writing Vocabulary: 10,000 Words Most Commonly Used in Writing* (published in 1926), it was compiled mainly from letters (personal and published) and other sources of writings (2001: 273). Both Thorndike's and Horn's studies were practically aimed at investigating the frequency of most common words, but the first of them was the study specially dedicated for educational purposes.

There were two other pre-electronic corpus studies which "were destined to have the most impact on lexicography" (Landau, 2001: 274) and were related to the field of English language teaching. One of them was the corpus work of Michael West, Harold E. Palmer and A.S. Hornby of 1930s and was aimed to help in teaching of English to foreign learners. In Landau's opinion it was "seminal and has had a lasting impact on ESL dictionaries to this date". Harold Palmer, one of the initiators of this corpus, was the leading figure in the so called "vocabulary control" movement, the movement which was concerned with limiting of vocabulary for foreign learners, "which had a deeper or more lasting effect on the early history of the learner's dictionary" than any other activity of that time (Landau, 2001: 274). Whereas Jackson (2002: 129) argues that a leading contributor to the "vocabulary control" movement was Michael West, who sought ways in identifying the essential vocabulary to learners of English. Harold Palmer studied grammatical patterning of words, and together with A.S. Hornby, as observed by Jackson, performed studies of collocations and idioms, which led into the first general-purpose learner's dictionary – the *Idiomatic and Syntactic Dictionary of English* (1942).

The other study, related to the field of education, was *The Interim Report on Vocabulary Selection* prepared by H. Palmer, M. West and L. Fauccet in 1936. It "was an extremely influential study" (Landau, 2001: 274), since it reviewed grammatical functions and meanings of words and made a distinction of various parts of speech. When the part of *The Interim Report* was supplemented with information on semantic frequency counts (by M. West with an assistance of I. Lorg), it was expanded into *A General Service List of English Words*. In this study M. West introduced a new system of dividing the senses of the words and indicated relative frequencies of each sense. Unfortunately, this work was of "limited practical use in lexicography", because the number of words it contained was not large enough, but "as a model of what should be done to improve corpus research in lexicography, it is of major importance" (Landau, 2001: 275). In fact, as it is highlighted by Landau, M. West was the one to publish the first EFL dictionary, *The New Method English Dictionary*, in 1935 (ibid.: 274-5).

Teubert in his "Corpus Linguistics: A Short Introduction" refers to Randolph Quirk's *Survey of English Usage* as the first large-scale project in collecting language data 'for empirical grammatical research', later it led to what became known as the standard English grammar: *A Comprehensive Grammar of the English Language*. This project was commenced in 1950s and at the moment of implementation did not consider computerizing the data. Computerization of data was undertaken only in the mid-1980s within a subsequent project known as International Corpus of English. The significance of this project, as Teubert claims, was such that "it formed a reference point for anyone interested in empirical language studies" (2007a: 51). As Teubert explains, Quirk's Survey consisted of spoken and written components. The spoken part consisting of 500,000 words of spoken English – it was the first to be put on a computer by Jan Svartvik and in the late 1970s it became the London Lund Corpus. It was a corpus to which transcriptions of phonological and phonetic information were added, moreover it became the first spoken corpus widely available for use and remains one of the largest computerized corpora of spoken material. Although it was mostly interested in grammar and not in meaning, "it was one of the very few projects working on empirical data" (ibid.: 51-2), respectively linguistic data were collected with the aim to investigate "different grammatical patterns and features of each word" (Landau, 2001: 280). Moreover, Quirk's *Survey*, being the first systematic collection of texts and spoken material to be subjected to grammatical analysis, was used as the basis for tagging and parsing of corpora in 1980s (ibid.). The issue of tagging and parsing of text corpora is addressed in more detail in the chapter on corpus application in lexicography.

Another important early corpus study is the project known as *English Lexical Studies* commenced in 1963 by John Sinclair. The significance of this project is such that J. Sinclair was the first to use corpus specifically for lexical investigation and his study was based on the concept of the collocation introduced by H. Palmer and A.S. Hornby in their *Second Interim Report on English Collocations* (1933) and later taken up by J.R. Firth in his paper 'Modes of meaning' (1957). The aim of Sinclair's study was to investigate the meaning of collocations on the basis of rather small electronic text sample (amounting to less than 1 million words) (Teubert, 2007a: 53-4).

The review presented in this sub-chapter shows that text corpora existed long before they appeared in electronic format. It also shows that corpus data were collected for the application in learner's dictionaries as early as in 1930s by Michael West, Harold E. Palmer and A.S. Hornby and as a result of their activities the established conventions of dictionary microstructure were changed.

1.2.2.2. Electronic corpus studies

1960s was the time when advances in computer technologies facilitated creation of large corpus databases of huge text files enabling lexicographers to undertake much deeper studies of word meanings – investigating their frequencies, various senses and uses in different contexts. It was time, as Leech notes, when in 1961-64 the compilation of "a systematically organized" computer corpus (the first computer corpus ever created (Meyer, 2002: xii)), was undertaken in the USA. The Brown University Corpus of American English (known as Brown Corpus) also referred to as the *Standard Corpus of Present-Day Edited American English* consisted of 500 written text samples (each consisting of about 2000 words) and was made up of "a systematic range of publications" in the USA during 1961. From then on Leech argues that the "machine-readable corpora have gradually established themselves as resources for varied research purposes" (2002: 4). The Brown Corpus consisted of one million words, was made up of 500 texts of American English belonging to 15 text types, each text sample consisting of 2,000 words. It was "a carefully organized" corpus and was proofread for no mistakes (Teubert, 2007a: 52).

A corpus of British English, similar in scope and composition to the Brown Corpus, was the Lancaster-Oslo-Bergen Corpus (known as LOB Corpus), which was compiled in 1970s. After both corpora – the Brown Corpus and LOB Corpus - were completed, it was soon realized that a corpus of one million words represented only a small portion of the vocabulary of a certain language (Teubert, 2007a: 52). Leech values the size of the Brown Corpus as

adequate for the study of common features (e.g., common grammatical constructions, punctuation marks or some affixes) but inadequate for application in lexicography (2002: 6). Nevertheless, as Landau emphasizes (2001: 279), the Brown Corpus served as a model for how a representative language corpus could be compiled. Moreover the Brown Corpus and the LOB Corpus allowed direct comparison of both varieties of the English language – British English and American English, but the real significance of these two corpora was in "establishing a benchmark for much larger corpora" that eventually followed (ibid.: 280).

A significant project, the so called COBUILD project was initiated in 1980. This project was a joint venture of Collins, the British publishing company, and the University of Birmingham. Originally the corpus was called the Birmingham Corpus, the acronym COBUILD stands for '*Collins Birmingham University International Language Database*". In 1982 the Birmingham Corpus consisted of about 7.3 million words and was compiled primarily, as ascertained by Landau (2001: 287), for the creation of a new advanced level dictionary for foreign learners – the *Collins Cobuild English Language Dictionary* (*Cobuild ELD*). The Birmingham Corpus consisted of texts written from 1960 and later, texts were taken from a wide variety of sources. In 1987, when *Cobuild ELD* was completed, the corpus had grown to 20 million words. Later corpus was given a new name and it became the Bank of English, and by 1997 it consisted of 300 million words.

The corpus has kept growing and nowadays consists of about 524 million words, it is composed of many different types of writing and speech. Written texts come from newspapers, magazines, fiction and non-fiction books, brochures, reports, and websites. Spoken material comes from television and radio broadcasts, meetings, interviews, discussions, and conversations. The Bank of English "provides evidence about the English which people read, write, speak and hear every day of their lives" (http://www.collins.co.uk/books.aspx?group=153).

In the late 1980s the Birmingham Corpus was followed by the Longman Lancaster English Language Corpus, which was developed by the Lancaster University and, as claimed by Landau (2001: 288), was designed to be used in editing dictionaries for foreign learners. The Lancaster Corpus together with other two corpora – one of spoken English and another of learners' English – comprise the Longman Corpus Network, which was used as the primary source for compilation of the third edition of the *Longman Dictionary of Contemporary English*, a dictionary for advanced-level learners of English.

The Longman Corpus Network is an increasingly large database, which amounts to 330 million words consisting of a wide range of "real-life sources such as books, newspapers and magazines" (http://www.pearsonlongman.com/dictionaries/corpus/index.html).

The Longman Learners' Corpus is a remarkable component of the Longman Corpus Network, it consists of essays and exam scripts compiled and sent by students and teachers from all over the world. The corpus compilers claim that it represents "every language level" and provides "an unprecedented insight" into learner's English for it provides information on what type of mistakes students make. The Longman Learners' Corpus was used for instance in the compilation of the Usage Notes for the *Longman Active Study Dictionary*, the Usage Notes were written considering the typical mistakes students make (http://www.pearsonlongman.com/dictionaries/corpus/learners.html).

Another large-scale corpus to mention is the British National Corpus, which was developed starting from 1991 by major academic centers, publishers and public institutions. Oxford University Press is one of the main contributors to the British National Corpus (Landau, 2001: 288). The British National Corpus consists of 100 million words, all British English, is compiled from over 4000 texts – 90% written and 10% spoken. The written part of the corpus includes "extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, [and] many other kinds of text", while the spoken part consists of "orthographic transcriptions of unscripted informal conversations, … spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins" (http://www.natcorp.ox.ac.uk/corpus/index.xml). Oxford Advanced Learner's Dictionaries make extensive use of the British National Corpus (Landau, 2001: 289).

The Cambridge International Corpus (*CIC*) was developed by Cambridge University Press, which produces the *Cambridge International Dictionary of* English - a dictionary designed for advanced level learners of English language. The CIC consists of both British and American English and was also compiled for writing books for learners of English.

As the *CIC* creators (http://www.cambridge.org/elt/corpus/international_corpus.htm) state "the English in the CIC comes from newspapers, best-selling novels, non-fiction books on a wide range of topics, websites, magazines, junk mail, TV and radio programmes, recordings of people's everyday conversations and many other sources". The CIC includes the following corpora:

- Cambridge and Nottingham Corpus of Discourse in English (recordings of spoken English across the UK), consisting of 5 million words;
- Cambridge and Nottingham Spoken Business English (recordings of business language in commercial companies) of 1 million words;

- Cambridge Cornell Corpus of Spoken North American English (recordings of spoken English across North America), consisting of 0.5 million words;
- Cambridge Corpus of Business English (business reports and documents from the UK and US) – a collection of 100 million words;
- Cambridge Corpus of Legal English (law related books and articles from the UK and US) - 20 million words;
- Cambridge Corpus of Financial English (books and articles relating to economics and finance from the UK and US) - 55 million words;
- Cambridge Corpus of Academic English (text from academic books and journals from the UK and US) - 30 million words;
- Cambridge Learner Corpus (exam scripts written by students taking Cambridge ESOL exams) - 30 million words. The Cambridge Learner Corpus (CLC) is a large collection of exam scripts written by students taking Cambridge ESOL English exams around the world. It currently contains over 95,000 scripts and is constantly growing.

To conclude the sub-chapter on the history of text corpora the following distinctive features of pre-electronic and electronic corpus studies can be distinguished:

- **early pre-electronic corpus** studies made use of authentic examples from selected texts, such approach implied that texts and excerpts were limited in types and language genres; examples were limited in context; citations were usually paragraph-long or sentence-long and were not representative of a language (e.g. only best literary authors were cited); citation files were manually created and manually stored.
- **electronic corpus** studies had the capacity to create and store huge data-bases of written and spoken (when transcribed) language varieties; various language genres were represented: writings (academic, scientific, fiction etc.), press reportages etc.; different corpus-based linguistic studies could be implemented – more complex and reliable studies were carried out, they were based on natural language texts and huge amounts of linguistic data; electronic corpus data are representative of the real usage of language.

1.3. Types of text corpora

Text corpora differ depending on their size and purpose they were created for. Early corpora were smaller in size and scope if to compare with text corpora available today. For instance the Brown corpus amounts to 1 million words, while the Bank of English contains over 500

million words. The Brown Corpus was primarily created for the purpose of linguistic studies - as Meyer suggests (2002: xii) to implement "systematic study" of individual text types of written English and its size of 1 million words was enough, while the Bank of English was created to compile a dictionary for learners of English and its size was of major importance. There were studies to investigate how the English language changed and developed or studies to investigate particular language genre (e.g. language of newspapers or academic writings). For different studies different text corpora are created and their types can be listed as follows:

- a **'balanced corpus'** – is a corpus which contains different text types of written English or, as Meyer suggests (2002: xii), has a balanced grouping of genres and can be used for study and comparison of those individual genres.
- a '**representative corpus' –** is a corpus, which is compiled of a variety of texts, but each text type consists of "equally sized" samples (Baker, 2006: 27).
- a **'reference corpus'** – is a corpus which is made up of a wide variety of texts and is consisting of millions of words (Baker, 2006: 30).
- a **'monitor corpus'** – is large in size and is continuously updated with new words and new meanings; such corpus in Meyer's opinion (2002: 15) is neither fixed nor static.
- a **'learner corpus'** – is a corpus, which consists of the texts either writing assignments or test papers, written (e.g. by learners of the English language) (Landau, 2001: 293).
- a '**specialized corpus'** – can be used to study aspects of a particular variety or genre of language (Baker, 2006: 26).
- a '**historical corpus**' – is a corpus which is aimed at studying "linguistic developments" of the English language. It investigates how language changed in earlier periods and how it has changed from the past to the present (Meyer, 2002: 20).
- a '**diachronic corpus'** – is a corpus which is created to be representative of a language or language variety over a particular period of time and makes it possible to study linguistic changes within that period of time (Baker, 2006: 29).
- a **'synchronic corpus'** – covers language of a particular period of time rather than corresponds to its historical development.
- **'national corpora'** - text corpora that were created at a 'national level'. Within this type of corpora Landau specifies: the Kolhapur Corpus of Indian English, the Wellington Corpus of Written New Zealand English, the Australian Corpus of English and the corpus of English-Canadian writing (2001: 292).
- '**parallel corpora**' –are used in translation theory and foreign language teaching. This type of corpora is designed to implement comparative analysis of English and other languages (Meyer, 2002: 22).

It is interesting to note that the British National Corpus, which is one of the corpora used in ESL lexicography for compilation of Oxford advanced learners' dictionaries is classified into the following types (www.natcorp.ox.ac.uk):

- it is **monolingual** it deals with modern British English, not other languages used in Britain (although non-British English words do occur in the corpus).
- it is **synchronic** - it covers British English of the late twentieth century;
- it is **general** - it contains texts of many different styles and varieties, and is not limited to any particular subject field, genre or register and contains examples of both spoken and written language;
- it is **sample -** written sources are made of samples of 45,000 words, they are taken from various parts of single-author texts. Shorter texts up to a maximum of 45,000 are included in full.

The example with the British National Corpus demonstrates that the different corpus types can combine. For instance the Brown Corpus being made up of 500 texts which belong to 15 genres (e.g. fiction texts, technical writings, documents compiled by authorities etc.) is classified as a 'balanced corpus' (Teubert, 2007a: 52) But its equally sized samples of 2,000-words for each genre make it a representative corpus. Therefore it is concluded that the Brown Corpus is balanced and representative at the same time.

Another example where corpora types combine is example with national corpora – they were structured along the same lines as the Brown Corpus and therefore can be used for comparative studies of English varieties (Landau, 2001: 292), this makes them in a way parallel corpora too, moreover they can be classified as specialized corpora for they represent particular language variety.

From the above stated definitions it can be concluded that the text corpora are created with a particular purpose. For instance specialized corpora are created with the aim to study particular aspects of a language - one might be interested to study the language of academic essays or the language of newspapers then only texts which would correspond to such special criteria would be collected (Baker, 2006: 26).

The Helsinki Corpus and ARCHER (A Representative Corpus of English Historical Registers) fall into the classification of historical corpora – they were created to study linguistic developments of the English language. The Helsinki Corpus contains texts from the Old English period (beginning of the 8th century) through the early Modern English period (the first part of 18th century) and ARCHER contains texts covering the period of 1650-1990. Meyer defines the two of these corpora also as '**multi-purpose general corpora'** for they

represent different texts and their fragments of various periods of English (Meyer, 2002: 20-1).

A typical example of reference corpus is the British National Corpus consisting of about 100 million words, while the Bank of English corpus consisting of over 500 million words falls into the category of reference and monitor corpus – it is huge in size and is constantly upgraded (Meyer, 2002: 15). It is difficult to claim whether the Longman Corpus Network and the Cambridge International Corpus, both being reference corpora, can be also classified as monitor corpora for it is not clear whether they are constantly updated. It is assumed that they are kept updated since they are corpora which were created for ESL lexicography - dictionaries for learners are published in newer editions, it requires corpora being kept updated.

To sum up this sub-chapter an attempt is made to describe a corpus which would best fit the requirements of ESL lexicography. Such a corpus should be:

- representative – it can have equally sized samples of a huge variety of texts but the number of equally sized samples should be rather high,
- general - containing texts of many different styles and varieties, which is not limited to any particular subject field or genre and should contain examples of both spoken and written language,
- reference corpus – it should contain different types of texts and should consist of millions of words;
- monitor corpus - it should be kept updated and allow that new words are entered.

To sum up it can be stated that an ideal corpus for ESL lexicography denotes the entity of a corpus which has the characteristics of representative, general, reference and monitor corpus.

1.4. The use of text corpora in different fields of linguistics

Corpora differ not only according to their types but also according to their applications. They are used in a number of studies by linguists of "various persuasions" - by descriptive or theoretical linguists, computational linguists, psycholinguists, historical linguists or sociolinguists and lexicographers. Meyer suggests (2002) that the corpus is used in the following fields of linguistics:

- **psycholinguistics -** the CHILDES Corpus was studied to investigate child language acquisition, for this corpus provides transcriptions of children speaking in different communicative situations (Meyer, 2002: xii, xiv.).

- **historical linguists** - the Helsinki Corpus is extensively used by **historical linguists** interested in evolution of English for it contains various types of written texts from earlier periods of English (ibid).
- **sociolinguists** – who are interested in studying the language of particular age group of teenagers use the COLT Corpus (the Bergen Corpus of London Teenage English) because it provides the speech of London teenagers (ibid.).
- **computational linguistics -** corpora of reasonable size (where size of a corpus is more important than balance) are used for research in natural language processing – tagging, parsing, information retrieval and the development of speech recognition systems (ibid.: 24).
- **language acquisition** – to study second language acquisition a number of learner corpora have been developed. These corpora contain exam excerpts or assignments compiled by learners of English either as a second or foreign language. One of the largest corpora is the International Corpus of learner English, which amounts to 2 million words in length. Other corpus to mention is the Longman's Learner Corpus, which makes use of it when compiling usage notes in the dictionaries for learners, those are aimed at indicating typical mistakes which learners should avoid (ibid.: 26).
- **language pedagogy** – makes use of learner corpora to develop teaching strategies for learners of English as a second or foreign language. Learner corpora provide information on how students actually use the English language and such information helps to develop teaching strategies or compile lesson plans or write textbooks.

Corpus has been quite successfully used also **in writing grammar books**, too, for it provides information on conventional or more complex grammatical constructions. A good contemporary example is *the Cambridge Grammar of English* (2006, published by Cambridge University Press), prepared by R. Carter and M. McCarthy, which is based on the Cambridge International Corpus of 700 million words.

This sub-chapter was aimed at describing a number of linguistic fields where corpora have proved to be valuable resources. The corpus use in lexicography is of the greatest importance for this research paper, therefore it is discussed in separate chapter, which illustrates how corpus is applied, and how corpus is consulted and analysed in ESL lexicography because this branch of lexicography uses corpus evidence most successfully.

2. THE USE OF TEXT CORPORA IN MODERN ESL LEXICOGRAPHY

Nowadays lexicographers are no more limited to a restricted number of lexicographic resources, they can have access to previous dictionaries, apply to citation file(s) and archives maintained by the dictionary publishers and most importantly they can make use of computer corpora (Jackson, 2002: 28-9). Čermak divides lexicographic resources into primary and secondary - archives and corpus being the first type and fieldwork, other dictionaries, encyclopaedias or www sources being the second. He believes that there is hardly any alternative to corpora being the primary and main resource for lexicography, for the number of corpus-based dictionaries is steadily growing (2003: 18-20).

For instance COBUILD dictionary was one of those to use the computer corpora based on which its entry words and definitions were compiled. The *Bank of English* corpus is used for compilation of Collins dictionaries both for native speakers and for learners, while the *British National Corpus* is consulted by Oxford dictionaries, the Longman Corpus Network is used by Longman publishers and the *Cambridge Language Survey* by the Cambridge publisher (Jackson, 2002: 167).

When speaking of lexicographic practise in the past, Čermák (2003: 18) notes that "traditionally, centuries-long practise has relied on extensive and manually acquired citation files". This implied that examples and excerpts of texts were manually selected, they were written down on citation slips and citation files were created.

In some way the citation files' approach, as defined by Čermák, failed to demonstrate the meanings of examples in terms of their use in particular contexts. The author argues that different types of words require different context sizes, which were simply not considered due to practical reasons and the words in question were written down on the citation slips with a record of only surrounding sentences. Another problem arose with the choice of "what had been excerpted". A person involved in writing down the examples was mainly relying on his/her intuition on what was typical or specific use (ibid.: 18-9) and very often they chose exactly the "specific usage" not the typical one.

Nowadays on the contrary, thanks to great advances in computer technologies it has became possible to create huge corpus databases and store lots of natural language texts. In such situation contemporary lexicographers have gained several advantages over work of lexicographers in the past and lexicographic study is no more limited to sentence-length excerpts (Biber et al, 1998: 22) or on intuition in selecting proper instances of words and context examples.

However, Leech (2002: 4) is of the opinion that corpus use today is seen as a "corpus plus intuition", rather than "corpus versus intuition". In favour of this opinion Leech argues that corpus user, when analysing corpus data, is applying his/her intuition and interpretative skills using knowledge of and about the language.

2.1. The process how corpus is consulted

Traditionally, lexicographic research investigated the meanings of words and synonyms. But nowadays, as Biber *et al* (1998: 21) explain, using corpus-based techniques lexicographers can study the ways how words are used and investigate, for instance, frequency of words and their senses, "systematic associations" of words with other words and "systematic associations" of words with particular registers or dialects. These aspects in Biber's *et al* opinion are central to dictionary making (ibid.: 21). Contemporary lexicographers use corpus not only to verify frequencies and meanings, they use corpus to locate and confirm all sorts of grammatical or contextual patterns, which are essential for dictionary compiling.

To specify these aspects of lexicographic research in more detail, we shall look at the characteristic features of corpus-based studies, and define what they imply. In Biber's *et al* point of view (1998: 4) corpus-based analysis can be characterized as follows:

- corpus, being a large and principled collection of natural texts, is the basis for implementation of analysis;
- corpus-based study is empirical and investigates the actual patterns of use in natural texts;
- corpus-based studies are implemented using computers, which permit use of automatic and interactive techniques;
- corpus-based study depends on both quantitative and qualitative analytical techniques.

As Biber *et al* (1998: 23-4) continue, a typical lexicographic investigation addresses the following major research questions:

- meanings associated with a particular word – is the research question of 'the traditional focus of lexicography".

  Nowadays lexicographers identify the meanings of words by looking at their occurrences in natural contexts, as corpus provides sufficient data of all contexts in which a word occurs. Lexicographers do not have to rely on intuition about how words are used and advantage is that these conclusions are no more based on incomplete citations.
- frequency of words;

This type of investigation allows identifying of how often different words are used and define most common and rare words.

Most concordance programs can produce frequency lists of all the words in a corpus, which show the number of times that each word occurs in the corpus. This information is very useful to ESL lexicography since learners of English should know which words are most common and therefore worth learning.

- commonly co-occurring words, known as collocations;

  This research question focuses on the ways words co-occur and group together. Description of collocations, in Jackson's opinion (2002: 18), is most reliably based on the analysis of large computer corpora. Collocation refers to the restrictions on how words can be used together, for example which prepositions are used with particular verbs, or which verbs and nouns or adjectives and nouns are used together. Provision of collocations is particularly important in ESL lexicography.

- different senses of words;

  This question of enquiry focuses on the different senses of words, their frequencies and usage patterns.

- seemingly synonymous words – how they are used and distributed in different ways;

- 'non-linguistic association patterns' of words (e.g., to registers, historical periods, or dialects);

  This research question, as explained by Biber *et al* (ibid.: 32-5) concerns "the association patterns between words and non-linguistic factors". Such investigation shows the language use patterns in different varieties. For instance, a lexicographer might want to investigate relative frequency of a word across registers – how it is used and how is distributed either in scientific writings, or press reportage or fiction. Or one might investigate the development of words through time or verify how it is used by different social groups.

The corpus-driven frequency lists and more especially concordances are the basic tools of lexicographers (Leech, 2002: 12) and a concordance is a device (computer software) for retrieving information from a corpus, which allows creation of these frequency lists.

One of advantages of computer technologies is in capacity of performing quite complex and specific operations, such as automatic processing of a computer corpus data – grammatical tagging or parsing. Grammatical tagging of a corpus (done automatically by a computer program known as a tagger) is part of annotating a corpus, which means that a grammatical tag or word-class label is attached to each word of a corpus. Such a tagged corpus enables another automatic processing – production of frequency lists which are

lemmatized; where different grammatical forms of one and the same word (lemma) are listed under one entry, similarly to a standard dictionary (Leech, 2002: 10, 11).

Annotation of a corpus is particularly important in lexicography – tagged corpus provides information on word classes and provides more complete description of language use. Annotating of a corpus means supplementing corpus with linguistic or textual information, where a tagger assigns items to classes and a parser specifies what functional relations there can be between the members of these classes in context (Aarts, 2002: 67).

2.2. The process how corpus data are analysed

A corpus lexicographer examines the corpus data in the form of keyword-in-context (KWIC) lines. The usual output of a computer search is a concordance list with the keyword examples of natural language sentences. Each occurrence of the chosen word is presented on a single line, with the word in the middle and context on each side (Atkins *et al* 2003).

Furthermore, when lexicographer has performed a search for the keyword and a list of all its occurrences has been produced, a concordance program provides a possibility to rearrange the citations of the keyword (basing on various features of the keyword's context), citations might be extended into full sentences or paragraphs, and be looked up for information on relative frequencies of the keyword and its corpus collocations. Other software, which can be used similarly to concordance program, is the Word Sketch program, which as Atkins *et al* has characterized it "is more sophisticated and informative presentation of corpus data" (ibid.: 270).

"These electronic aids have speeded up traditional lexicography, and made it more efficient, more able to make sense of the mass of data on offer" is concluded by S. Atkins *et al* (ibid.: 271).

To present some more details of what concordance program may offer and produce, the following Internet site (http://www.concordancesoftware.co.uk/features.htm) was looked up and the following key features (but not limited to) were traced:

Concordance allows to:

- count words, make wordlists, word frequency lists, and indexes
- make full concordances showing every word in its context
- make fast concordances picking particular words from text
- view a full wordlist, a concordance, and original text simultaneously
- browse through the original text and click on any word to see every occurrence of that word in its context

- edit and re-arrange a wordlist by drag and drop
- lemmatise - group together any words one might choose

Concordance provides flexible, powerful textual analysis:

- User-definable reference system: identifies which section of a text each citation comes from
- User-definable contexts: words are shown in contexts which can be varied by length or sense-unit
- Select and sort words in very flexible ways
- Search for phrases, do proximity searches, sample words, and do regular expression searches
- Stop Lists one may specify words to be omitted from a concordance
- Word length chart

Comprehensive text tools:

- Built-in file viewer can display files of unlimited size
- Built-in editor allows fast editing of files up to 16MB

High usability:

- Easy user interface with good use of Windows features
- Context-sensitive help system with 150 topics
- Runs fast - can pick 15000 occurrences of a word from a 1.5MB text in under 4 seconds

The answer to the question why dictionaries for learners of English are increasingly being based on the corpora lies in the fact that the corpora with the software to analyse them first can automate the process of creating dictionaries and second it improves the quality of information which goes into the dictionary. Automation process takes seconds to perform frequency counts of words and their ranking into most frequent or less frequent 'forms'. But concordance software shows typical contexts in which words are used or what meanings they have and many other constructions it may produce – theses features all in all determine the quality of the information which goes into the dictionary (Meyer, 2002: 15)

Meanings and sub-senses, collocations and phrases in learner's dictionaries are ordered according to their frequency as defined by text corpora. In this way learners are given the guidance which meanings to learn first.

3. GENERAL PURPOSE MONOLINGUAL ENGLISH DICTIONARIES AND DICTIONARIES FOR LEARNERS OF ENGLISH

This chapter is aimed at describing two types of dictionaries – the general purpose monolingual English dictionaries and English learner's dictionaries from the following points of view:

1) general notion and aim of both dictionaries;
2) brief historical origin of the general purpose monolingual English dictionaries from which dictionaries for learners of English originated;
3) characteristic, distinctive and/or common features of both types of dictionaries.

An attempt is made to answer the question of what has changed with the advances in computer technologies which facilitated advent of corpus evidence and had an immense impact on the compilation of contemporary dictionaries for learners of the English language.

Before reviewing the issue of dictionaries it is worth emphasizing that the main object of lexicography, as stated by Bejoint (2000: 7) "is to define words and terms".
The defining of words and terms has been a challenging task to lexicographers already for centuries. What makes such a task so challenging is that a lexicographer has to "capture the 'meaning' of a word in a 'definition'" (Jackson, 2002: 15). The meaning of a word, in Jackson's opinion, is "composed of a number of features: its relation with the real world, the associations that it carries with it, its relations with other words in the vocabulary, and the regular company that it keeps with other words in sentence and text structure".

To Brumfit (in Ilson's, 1985: v) a dictionary is "the most widespread single language improvement device ever invented". For Ilson (1985: 1) "The dictionary is the most successful and significant book about language". But Atkins (1985: 23) suggests that "a good dictionary must do as much as possible to provide users not only with what they know they want, but with what they don't know they want, as well".

In Jackson's (2002: 76) opinion, dictionaries have two fundamental aims: coverage and accessibility. "Coverage includes the aim to be 'comprehensive', representing an up-to-date and wide-ranging selection of vocabulary, and the aim to be a 'faithful record' of the lexical resources of the language". Both aims are user-oriented aims – what they want has to be made available by the most straightforward means.

3.1. The general purpose monolingual English dictionaries and their historical background

When Sterkenburg (2003: 3) speaks of a dictionary, what comes to his mind is a notion of the prototypical dictionary. In his opinion the prototypical dictionary "is the alphabetical monolingual general-purpose dictionary" which employs the following characteristic features: "the use of one and the same language for both the object and the means of description, the supposed exhaustive nature of the list of described words and the more linguistic than encyclopaedic nature of the knowledge offered".

To explain what makes the monolingual general-purpose dictionary so prototypical, Sterkenburg refers to a definition formulated by Bejoint: "It is the one that every household has, that everyone thinks of first when the word *dictionary* is mentioned, it is the type that is most often bought, most often consulted, the one that plays the most important role in the society that produces it". Moreover what Bejoint (2000: 3) has mentioned is that the general-purpose monolingual dictionary "reflects ... the society in which it is produced..., it is a mythical emblem of learning and serves as a tool for improving of linguistic communication".

Definitions suggested by Sterkenburg and Bejoint imply that the general-purpose monolingual dictionary is the most typical type of a dictionary one might possess. Its characteristic feature is that one and the same language is used in the description of its macro and microstructure. The general-purpose monolingual English dictionary having a typical example here contains information only in English and is mainly compiled for the use of native English speakers.

Another issue questioned by Sterkenburg (2003: 5-8) is: "What makes a general-purpose dictionary to be called a dictionary"? Among the required criteria the author has listed - formal criteria, functional criteria and criteria regarding content.

Formal criteria would account for the form and the structure of a dictionary. Dictionary is a reference work contains linguistic information and has a double structure – macrostructure for entry words and microstructure for information given about every entry word.

The general-purpose dictionary has a social function – it is 'a reflection of social change' and performs roughly speaking the function of providing systematized information on meanings of words. Some dictionaries omit negative qualifying terms, curses or nicknames and therefore are considered to be 'guardians of moral and ideological values'.

Criteria regarding content refer to all linguistic information, which is included within the entry word.

Having evaluated the above mentioned criteria of a proper general-purpose dictionary, Sterkenburg comes up with the following definition (2003: 8):

> "The prototypical dictionary ... is a reference work and aims to record the lexicon of a language, in order to provide the user with an instrument with which he can quickly find the information he needs to produce and understand his native language. It also serves as a guardian of the purity of the language, of language standards and of moral and ideological values because it makes choices, for instance in the words that are to be described. With regard to content it mainly provides information on spelling, form, meaning, usage of words and fixed collocations".

Early monolingual dictionaries focused on the cultural and educational function, and were aiming as well at mastering user's competence in acquiring 'hard' words or loan words of Latin origin. The first monolingual English dictionaries were published in the 17th century – Cawdrey's *A Table Alphabeticall* in 1604, Bullokar's *An English Expositor* in 1616 and Cockerman's *English Dictionaire*, in 1623. Other early monolingual dictionaries, which were concerned with difficult words, were Edward Phillips' *New World of English Words* (1658), but scientific and industrial terms with more encyclopaedic information were treated by Nathaniel Bailey in his *Dictionarium Britannicum* (1730) (Sterkenburg, 2003: 11).

The first dictionary "coming close to a complete inventory of the English language" was John Kersey's *A New English Dictionary* (1702), and was based on revised Phillips' dictionary of the *New World of English Words*. Kersey's dictionary was compiled to be 'a good spelling guide'.

Great changes in the structure of monolingual dictionaries, as suggested by Sterkenburg (2003: 12-3), were introduced during the 17th and 18th centuries by the Italian and French Academies – the Academia della Crusca (established in Florence in 1582) and the Academie francaise (founded in 1635 by Cardinal Richelieu). In lexicography "an inventory of entire language" was proposed by using a corpus of literary quotations of 'purest' language texts. Archaic technical and scientific words were removed and dictionaries were given "a normative authority". At that time scholars wanted to record the language at "a certain stage in its development" – a step towards synchronic approach in lexicography.

The record of "perfectly developed" English in print was of a great importance in the second half of the 18th century. Samuel Johnson (1709-1784) when compiling his *Dictionary of the English Language* (1755) was inspired by the dictionaries of the Academia della Crusca and the Academie francaise. Johnson's dictionary was aimed at authority and at showing the best usage of words, for that a corpus of authentic literary texts was used.

Johnson was inventive in many ways – he used 114,000 citations to justify his definitions and usage of words, much attention he devoted to spelling and where required he

added his comments whenever there was doubt about usage. Although Johnson's dictionary was an innovative dictionary and greatly influenced lexicography in the 19th century, it was widely criticized for being selective in describing and exemplifying only 'the fine and good' words, while description of all the words in dictionary would be much more appreciated (Sterkenburg, 2003: 13-4). But R. Ilson remarks the significance of Samuel Johnson's dictionary of 1755, in his opinion it was intended 'to be useful to both groups' – native speakers of English and learners of English for he devoted special attention to the treatment of phrasal verbs and in explanations included synonyms of both Romance and Germanic origin (1985: 2).

The next subchapter concentrates on the analysis of general monolingual learner's dictionaries which form a sub-branch of the "prototypical" general purpose monolingual dictionaries.

3.2. The notion of English learners' dictionaries

Learners of English (as a foreign or second language) are among those users for whom, as Jackson (2002: 83) suggests, dictionaries have been specifically tailored.

Encoding needs of learners of English, when consulting a dictionary, differ from those of native English speakers. If a native English speaker, as argued by Jackson, is more concerned with the spelling of a word, learner of English is in need to verify spelling, pronunciation, inflections, and grammatical structure of a word, collocations and its usage in different contexts. This means that learner's dictionaries "need to contain more explicit, more comprehensive and more systematic information about the syntactic and lexical operation of words than a dictionary for native speakers" (2002: 84). At the same time an ideal learner's dictionary "should model [and represent] the lexical competence of the adult native speaker" (Ilson, 1985: 2).

3.3. The origin of the monolingual English learner's dictionaries

The monolingual English learner's dictionaries originated in 1930s-1940s when they derived from the activities of three teachers of English as a foreign language - H.E. Palmer, A.S. Hornby and M. West. They were involved in the studies to improve the standard of language teaching to foreign learners of English and were involved in the "vocabulary control" movement, developed the system of grammatical patterns and implemented studies on collocations and idiomatic expressions. Their activities led into the first general-purpose learner's dictionary the *Idiomatic and Syntactic dictionary of English*, 1942 (Jackson,

2002:129). From that moment on the monolingual learner's dictionaries have developed, in Rundell's opinion (1998: 316), "as a distinct lexicographic category".

What Rundell suggests (ibid.: 316-7) is that Hornby, West and Palmer developed new features to the established conventions of dictionary microstructure, which were specifically developed taking into account the needs of non-native learners of English. These features implied:

- vocabulary control – or vocabulary limitation was regarded by Hornby, West and Palmer in some respect central "to the project of creating a learner dictionary". It denotes two important aspects – the first concerns the notion of a selected subset of the lexicon to cover in a dictionary and the second concerns the notion of a restricted 'defining vocabulary' used in definitions.
- grammatical and syntactic information – was aimed at more detailed description to meet the encoding needs of the dictionary users – learners of a language;
- the role of examples – was such that they should have appeared extensively
- phraseology – Hornby's and his colleagues research revealed "the prevalence of ready made sequences in everyday speech and writing", which helped to rise interest in phraseology and describing and explaining phraseology has been one of the key features of monolingual learner's dictionaries ever since.

In Rundell's view (ibid.) the vocabulary control, pedagogically motivated examples, transparent description of syntactic behaviour and phraseological units are the defining characteristics of the monolingual learner's dictionaries.

In 1948 the *Idiomatic and Syntactic dictionary of English* was republished by Oxford University Press with the title *A Learner's Dictionary of Current English,* which in 1952 was changed into *The Advanced Learner's Dictionary of Current English*, later known as *Oxford Advanced Learner's Dictionary* (OALD) (Jackson, 2002: 129).

The first competitor to OALD was the *Longman Dictionary of Contemporary English* (LDOCE) published by Longman in 1978. The new dictionary introduced a number of improvements and innovations:

- the use of a restricted 'defining vocabulary', which was aimed at defining senses in the dictionary using only 2000 of the more common words of English;
- a new system of grammatical description was introduced, which anticipated coding system not only for verbs, but also adjectives and nouns and was aimed to be as transparent as possible;

- more attention was paid to the coverage of function words and other high-frequency items.

By the time when the second edition of LDOCE was published in 1987, a third monolingual learner's dictionary appeared – *Collins COBUILD English Dictionary (COBUILD)*, which introduced other significant innovations (Jackson, 2002: 131-2):

- it was based on a computer corpus - at the time when the use of corpus for dictionary compiling was considered revolutionary for its evidence of real language data;
- definitions of entry words were given in complete sentences and were aimed at giving some idea of typical contexts;
- the examples were based on corpus evidence of real English;
- and the grammatical information on typical patterns was provided in an 'extra column'.

The last monolingual learner's dictionary to enter the UK market, as argues Jackson (ibid.), was the *Cambridge International Dictionary of English (CIDE)* published in 1995, at the time when OALD published its fifth edition, LDOCE its third and COBUILD its second. Among innovations offered by this dictionary were:

- headwords of various senses were separately distinguished and followed by a 'guide word' to the meaning (e.g. **job** employment, **job** duty, **job** piece of work etc.);
- every grammatical pattern was illustrated by an example;
- examples indicated typical collocations;
- phraseology of words was treated in more detail by introducing an extensive 'phrase index';
- some attention was devoted to the treatment of various English varieties – American, British and even Australian.

Macmillan English Dictionary for advanced learners was first published in 2002, it did not introduce any special innovations but extensively and very successfully applied corpus evidence.

To sum up this chapter it needs to be added that learner's dictionaries have changed with the advent of large text corpora and these improvements have occurred in terms of the quality of information learner's dictionaries provide. M. Rundell (1998: 316) explains it as follows:

> "- the description of a language that a dictionary provides corresponds more closely to reliable empirical evidence regarding the way in which that language is actually used;
> - presentation of this description corresponds more closely to what we know about the reference needs and reference skills of the target user".

3.4. Common and distinctive features of general purpose monolingual English dictionaries and dictionaries for learners of English

To state briefly the common features of general purpose English dictionaries and dictionaries for learners of English, the following characteristics can be mentioned:

- both dictionaries are monolingual English dictionaries;
- are used when engaged in decoding (reading or listening) or encoding (writing or preparing to speak);
- are aimed at defining the meanings.

What makes both dictionaries distinct is to what extent this information is treated and the way it is presented in terms of its coverage and quality. The table below was compiled to emphasize some distinct features of both dictionaries (based on B. Kirkpatrick (1985: 7-13) and M. Rundell (1998: 326-334).

Type of the dictionary	**General purpose monolingual English dictionary**	**Monolingual English learner's dictionaries**
target audience	native English speakers	learners of the English language
the role	explanatory	pedagogical, explanatory
coverage	comprehensive coverage of a language	the emphasis is on covering the most common words of a language
used	to expand vocabulary, check spelling	to learn the core vocabulary
meanings, illustrative examples	are only defined and if exemplified – (short) sentence examples are used for meanings which might seem confusing	are defined using restricted vocabulary and simple language definitions; exemplified by extensive use of full sentence illustrative examples
senses	sub-senses are not differentiated in depth	senses and sub-senses are much more and clearly differentiated;
collocations	might not be indicated; when listed - might not be exemplified	- patterns of co-occurrence and combinatorial behaviour of words are distinguished; - collocations are shown in their typical contexts and exemplified by illustrative examples;
grammar and syntax	grammar codes and patterns might not be transparent	grammatical patterns tend to be transparent, are built in definitions and examples
register and field labels	might not be mentioned but in good contemporary dictionaries are used	- stylistic labels and labels demonstrating pragmatics and attitudes (formal, informal, ironic, slang etc.) as well as field labels are extensively used

The main conclusions of such comparison are such that learner's dictionaries require definitions which are provided in simpler language than the dictionaries for native speakers do.

They should define various senses and sub-senses, which require to be much more clearly differentiated than in native speaker's dictionaries, since "extended meanings and figurative usages are much more understandable to native speakers than they are to foreign learners" (Kirkpatrick, 1985: 10-2).

In relation to function or structure words – they have to be treated in much greater detail in learner's dictionaries in terms of their usage. Kirkpatrick suggests (ibid.) that this can be achieved by illustrative sentences and examples.

And final distinctive feature to mention concerns the examples of usage. It is argued that most dictionaries for native speakers do not deal in any depth with such type of examples and it is partially explained by the fact that 'great deal of language comes instinctively to one whose mother tongue it is, in a way that it does not to the foreign learner'.
The main goal of the analysis of corpus application in learner's dictionaries is to justify how their distinctive features are exemplified by corpus evidence in practice.

3.5. Learners' dictionaries chosen for analysis

The major UK learners' dictionaries have access to large and diverse corpus resources which provide a reliable description of English. The following 3 learners' dictionaries were selected for the analysis of corpus application, which is the subject matter of this research paper, for they represent the type of contemporary dictionaries which have extensively used corpus in the process of their compilation:

- Longman Dictionary of Contemporary English, fourth edition 2005 (*LDOCE*) – makes use of the Longman Corpus Network, a database of 300 million words;
- Cobuild Advanced Learner's English Dictionary, fourth edition 2003 (*COBUILD*) – makes use of the Bank of English, corpus of 520 million words;
- Macmillan English Dictionary for Advanced Learners, International student edition 2002 (*Macmillan*) – is compiled using the World English Corpus of 200 million words.

The size of corpora these dictionaries have used is viewed as an important feature in analysing corpus evidence they have applied, and this was the reason why these dictionaries were selected.

4. THE FRAMEWORK OF ANALYSIS

The quality of information and more specifically the strategies how information is provided in learner's dictionaries are the starting points for the practical analysis of the present paper. The objective of the analysis in its broader sense is not to evaluate the quality of learner's dictionaries but to analyse, to an extent it is possible, particularly the role of corpus evidence in compiling of the learner's dictionaries.

4.1. The aim of analysis

Therefore the aim of the practical analysis, briefly stating, is to investigate the ways how corpus has actually been used in practical lexicography. To implement the analysis of corpus application in learner's dictionaries, the following steps were undertaken:

- the following monolingual English learner's dictionaries which claim to describe and explain the English language accurately, with the help of the evidence of corpus data and the choice of corpus based illustrative examples, were chosen for the analysis:
  - Longman Dictionary of Contemporary English, fourth edition 2005 (*LDOCE*);
  - Cobuild Advanced Learner's English Dictionary, fourth edition 2003 (*COBUILD*);
  - Macmillan English Dictionary for Advanced Learners, International student edition 2002 (*Macmillan*).
- to decide upon the aspects to be analysed the front matter of selected dictionaries was reviewed to establish what dictionary compilers say about the application of corpus evidence in these dictionaries;
- and finally, a list of the main aspects of analysis was compiled, which shall be used as the basis for analysis of corpus application in all the three learner's dictionaries (see 4.2.).

Furthermore, the analysis aims at investigating whether the practical use of corpus evidence corresponds with what has been mentioned in theory on the corpus application in lexicography (for more information on theory see chapter 2).

4.2. The structure of analysis

To implement the analysis of corpus application in learner's dictionaries a number of entry words of varying length representing different parts of speech will be chosen, the choice of particular entry words will be given and explained in the next chapter. They will be reviewed

within all of three learner's dictionaries in terms of their morphological, syntactical and semantic treatment as justified and supported by corpus application. When describing some distinctive features of dictionaries subjected to analysis, some individual entry words will be analysed to demonstrate those features which are characteristic to one or the other dictionary.

To implement the intended analysis of how corpus evidence is reflected in these dictionaries, the following issues will be addressed:

1) the illustrative examples provided by the dictionaries;
2) collocations and phrases;
3) the arrangement of senses in polysemantic entries;
4) the frequency rating of words and their various senses.

Common and/or distinctive features of corpus application will be pointed out and exemplified.

The analysis is targeted at analysis of corpus data application in learner's dictionaries and should justify the usefulness and advantages of corpus-based approach in ESL lexicography.

5. PRACTICAL PART - ANALYSIS OF CORPUS APPLICATION IN ENGLISH LEARNER'S DICTIONARIES

All the dictionaries of intended analysis claim that they have fully relied on corpus data to provide definitions, examples and any other information for their entry words. The table below includes information on the corpora which have been used by the compilers of these dictionaries.

Table 5.1. **Corpus types used in the selected learner's dictionaries**

COBUILD	The Bank of English corpus, a database of about 520 million words
LDOCE	The Longman Corpus Network, a database of about 300 million words;
MACMILLAN	The World English Corpus, a database of about 200 million words

The number of words in the corpus of *COBUILD* is almost twice the number of words in the corpus of *LDOCE* and almost three times bigger than in the corpus of *Macmillan*. It is questioned whether dictionary based on the largest amount of data somehow differs from ones based on the data with lesser number of words.

Carefully reviewing the front matter of dictionaries and information on how the corpus evidence has been applied, some typical approaches were recognized in terms of corpus application which are more characteristic to one or another dictionary of intended analysis. For instance, what *COBUILD* emphasizes (2003: vii, x, xi) is that it has selected those corpus examples which show typical grammatical patterns, typical vocabulary, and typical contexts. Collocations provide help with set lexical and grammatical patterns but definitions use vocabulary and grammatical structures that occur naturally with the word being explained.

While *LDOCE* (2005: x) claims that most extensive use of the corpus is in the area of collocation and phraseology. Examples and collocations are illustrated by many examples drawn directly from or based on the corpus but *Macmillan* aims at giving a rich account of the collocational and syntactic behaviour of the core vocabulary of English.

Based on the information provided in the front matter of dictionaries, it can be concluded that *COBUILD* when providing corpus evidence pays more attention to exemplifying typical syntactical patterns while *LDOCE* and *Macmillan* deals in more detail with collocations and phraseology. These aspects shall be verified in the course of the analysis.

The analysis of illustrative examples, collocations, the arrangement of senses and frequency rating will start with *LDOCE*, then *Macmillan* and finish with *COBUILD*.

5.1. Illustrative examples

The aim of illustrative examples is to make dictionaries more helpful for advanced level students of English, for they help learners to remember and understand the word within a context. For the analysis of illustrative examples – mostly the entry words ***advice***, ***believe*** and ***serious*** were chosen for these are entries of varying length representing different parts of speech. In case of analysis of *COBUILD's* examples, the entry words ***accept***, ***believe*** and ***serious*** were chosen for they give a better choice of examples to mention, while examples of *Macmillan* include also the entry word ***inadequate***.

5.1.1. Treatment of illustrative examples in *LDOCE*

LDOCE claims that all the examples which are based on the corpus evidence are usually slightly edited versions of real sentences from corpus and suggests that its examples are more realistic than in other dictionaries for they prepare students for the way that words are really used (2005: x).

Examples from *LDOCE*:

1. *Take my **advice** and study something practical.*
2. *Let me give you a piece of **advice**. Wear a blue suit to the interview.*
3. *It is **believed** that the house was built in 1735.*
4. *'Have they arrived yet?' 'Yes, I **believe** so'.*
5. *the **serious** problem of unemployment*
6. *Marry Frank? You can't be **serious**!*

The main approach of *LDOCE* is to provide full sentence corpus examples, only in some cases illustrative examples are given in a shortened form – they are presented as starting with no capital letter and ending with no full stop (5) and consisting of text minimal fragments (article + adjective + prepositional phrase). Full sentences denote more natural way of expression in comparison with short forms, which sound rather artificial and probably serve only as templates on which learners can model their own sentences. Sentences of interrogative form (4, 6) denote that they have been used in a 'real' conversation and were probably taken from the corpus of spoken texts. The examples do seem slightly edited for they are presented in the shortest possible form. Irrespective of the fact that the examples are taken out of the context and they differ in length - they illustrate various meanings, denote certain situations and can be well understood. Therefore it is assumed that the examples are

indeed natural and authentic English examples, which are taken from corpus rather than made up.

As stated before, the examples in *LDOCE* come from the Longman Corpus Network – a database covering 'books, newspapers, magazines and spoken English' (2005: x). If to compare examples from the point of view which type of genre they represent, then it can be concluded that the examples 1 & 2 might have been taken from books or texts of spoken evidence in the corpus, the examples 4 & 6 are typical spoken examples, while the example 3 could have been taken from a magazine and the example 5 is a typical example taken from a newspaper. The form of expression, completeness and punctuation marks denote and characterize the genres all these sentences might belong to.

5.1.2. Illustrative examples in *Macmillan*

Macmillan does not specify whether its examples are somehow edited or not. It is assumed that the analysis of illustrative examples shall answer this question.

Examples from *Macmillan*:

1. *Tenants involved in a dispute with their landlord should seek legal **advice**.*
2. *He ignored the doctor's **advice** that he ought to lose weight.*
3. *She found it hard to **believe** that he was a real businessman.*
4. *I would never have **believed** such a place existed if I hadn't seen it for myself.*
5. *We'll have to give the situation some **serious** thought.*
6. *The police have made no **serious** attempt to address these issues.*
7. ***inadequate** provision of health care*
8. *rail tracks that are **inadequate** for the loads carried on them*

In the majority of cases, *Macmillan* provides rather long and extensive corpus examples (1-6). This seems to be the main approach adopted by *Macmillan* throughout the whole dictionary, although there are some sentence fragments given as well (7, 8). Full sentence examples seem to be explicit for they are not somehow altered. It makes them look and sound natural as if taken from the source of authentic English and therefore can be valued to represent the true confirmation of the corpus evidence.

Macmillan provides examples taken from the World English Corpus which covers 'books, magazines, newspapers, e-mails, television and radio" (www.macmillandictionary.com). The example 1 seems to belong to broadcasting style, the examples 2 & 3 probably are taken from books, while the examples 6, 7 & 8 refer more to newspapers, television or radio. Example 4

might have been taken from an interview in a magazine, and 5 refers to an interview in spoken form which probably was broadcast on TV.

5.1.3. Illustrative examples in *COBUILD*

All the examples in *COBUILD* are presented exactly as they occur in the corpus. The majority of the examples are taken word for word from the Bank of English but occasionally *COBUILD* has made very minor changes to them, so that they are 'more successful as dictionary examples' (2003: vii, x).

Examples from *COBUILD*:

1. *...a workforce generally **accepted** to have the best conditions in Europe.*
2. *Urban dwellers often **accept** noise as part of city life...*
3. *If you **believe** in yourself you can succeed.*
4. *'You've never heard of him?' – 'I don't **believe** so!'*
5. *You really are **serious** about this, aren't you?...*
6. *I hope you're not **serious**.*

The examples in *COBUILD* are presented in three ways: as sentences without a beginning (1), as unfinished or interrupted sentences with three dots at the end (2) or as full sentence examples. It looks as if the compilers of the dictionary have considered it enough to present some of sentences as unfinished for they seem to implement the task of denoting and explaining the meanings. Such approach probably is somehow related to *COBUILD's* approach in presenting its definitions. Definitions in *COBUILD* are modelled on the way people explain the meanings of words to each other, they are written in full sentence examples and contain a lot of extra information. Probably that is why the examples, following the explicit definitions, are kind of shortened. Being shortened, these examples still do denote and convey meanings, therefore it is concluded that the examples are based on corpus evidence and represent natural and 'real' English.

Similarly to the other two dictionaries, *COBUILD* takes examples from the corpus which covers 'books, newspapers, magazines, broadcasting, and conversation' (2003: x). The context of both interrupted and unfinished sentences (1 & 2) of our example denote that they are examples of broadcasting, while the example 3 might have been taken from a magazine or a book and the examples 4, 5 & 6 are typical examples of a spoken text.

To summarize it should be stated that corpus-driven examples might be presented either as slightly edited or shortened examples or on the contrary as 'word for word' full sentence examples taken directly from the corpus and representing a number of text genres.

Irrespective of their form they remain natural forms of expression and are reliable models of usage. It should be stressed that *Macmillan's* approach to use mainly full sentence examples is the most successful if to compare with the other two dictionaries for it provides examples which are in most cases not altered and do not sound unnatural.

5.2. Collocations and phrases

Collocations and phrases when looked up in the corpus can be arranged in accordance with their frequency as shown by the corpus. Such information helps lexicographers to choose which are the most common collocations and expressions to be included into the dictionary. Such approach when collocations and expressions are ordered according to their frequency is applied in all the three dictionaries subdued to this analysis.

For the analysis of collocations the entry words ***advice*** will be compared and for the comparison of typical phrases the entry word ***apart*** is chosen. These entry words were chosen from *LDOCE* on the assumption that it provides most reliable evidence of corpus application in the area of collocations and phraseology and this is what *LDOCE* has especially emphasized (2002: x.

5.2.1. Treatment of collocations and phrases in *LDOCE*

In *LDOCE* collocations are 'illustrated by many examples drawn directly from or based on the corpus'. Compilers of *LDOCE* believe that examples and collocations are both used to help the student use words, and so are important aids to producing correct natural English (2005: x, xi).

For the entry word ***advice***, *LDOCE* has given examples of the following collocations, which are listed in their order of occurrence determined by the frequency in the corpus:

1. **give advice** *Could you give me some advice about buying a home?*
2. **legal/ medical/ financial advice** *If I were you, I would get some legal advice.*
3. **professional/ expert advice** (=advice from someone who knows a lot about subject) *I want to ask your advice where to stay*.
4. **follow/ take sb's advice** (=do what they advise you) *I followed my father's advice and sold the car. Take my advice and study something practical.*
5. **a piece/ word of advice** *Let me give you a piece of advice. Wear a blue suit to the interview.*

6. **on sb's advice** *On her doctor's advice* (=because her doctor advised her) *Smith decided to take early retirement.*

Exemplifying collocations *LDOCE* lists the most frequent collocations of the noun ***advice***, which participates in the following constructions: verb + noun (1, 4), adjective + noun (2, 3), is a part of prepositional phrase (5, 6) and used in possessive case (4 & 6). This information is particularly important for productive use and if explicitly given enables one to produce more natural English. For better findability *LDOCE* highlights typical collocations in bold face and for better understanding clarifies them where required (3, 4, 6).

The entry word ***apart*** in *LDOCE* is used in a number of phrases:

1. **fall apart** *It just fell apart in my hands. He drives around in and old car that's falling apart.* 2. **be torn apart** *The play portraits a good marriage torn apart by external forces.* 3. **be worlds/ poles apart** *I realized we were still worlds apart*. 4. **grow/ drift apart** *Lewis and his father drifted apart after he moved to New York.* 5. **joking apart** *Joking apart, they did do quite a good job for us*.

6. **sb/ sth apart** except for someone or something: *The car industry apart, most industries are now seeing an improvement in their economic performance.*

7. **set sb/ sth apart** to make someone or something different from other people or things: *Her unusual lifestyle set her apart as a child.*

A phrase is a group of two or more grammatically linked words, it functions as a single unit in the sentence and has its own meaning. Phrases might be confusing, therefore need to be clearly exemplified. What *LDOCE* exemplifies is that the entry word ***apart*** in the sentence might behave like an adjective (3, 6) or like an adverb and might perform the functions: as adverbial modifier, modifying the meaning of a verb (1), as part of a verb phrase (2, 3, 4), or as part of a phrasal verb (7). Phrases when looked up in corpus are chosen among the most common ones and are listed in the order of their frequency as justified by the corpus evidence.

5.2.2. Treatment of collocations and phrases in *Macmillan*

Macmillan gives the following examples of the collocations for the entry word ***advice***:

1. **give advice** *May I give you a piece of advice?*
2. **take sb's advice** (=do what someone advises) *I took his advice and left.*
3. **legal/ medical/ expert etc advice** *Tenants involved in a dispute with their landlord should seek legal advice.*
4. **on sb's advice** *She's acting on her lawyer's advice.*

5. **on the advice of sb** (=because of someone's advice) *She applied to York University on the advice of her teacher.*

By illustrating collocations *Macmillan* provides the following typical constructions in which the noun ***advice*** participates: verb + noun (1, 2), adjective + noun (3), is a part of prepositional phrase (4, 5) and used in possessive case (2, 4). Moreover *Macmillan* mentions the following most frequently used words with the noun ***advice***: adjectives – *expert, financial, legal, medical, practical, professional* and verbs – *accept, ask for, follow, get, give, ignore, need, obtain, offer, provide, receive, seek, take, want.* Besides *Macmillan* also lists a number of sentences, which denote the idea of giving/ taking advice and not necessarily include the noun itself. For instance: *I think **you ought to** see the doctor about that lump. **If I were you** I'd stick with your job until something better comes along. **Why don't you** just tell her the truth?*
All these examples are based on the corpus evidence from which it is verified which are the most frequent words to combine with the noun advice and which are the sentences denoting the meaning of taking or giving advice.

Macmillan provides the following examples for the phrases of the entry word ***apart***: 1.**take sth apart** *If the problem is in the printer, we'll have to take the whole thing apart. 2.* **fall/ come apart** *When I picked the book up it came apart in my hands.* 3. **be poles apart/ be worlds apart** (=be very different) *Politically, Gorbachev and Thatcher were poles apart, but they became friends.*4. **set sb apart** (=make someone different from others) *He has a unique genius that sets him apart from other writers.* 5. **tear sth apart** *Yugoslavia was being torn apart by ethnic conflicts.*
Similarly to *LDOCE*, Macmillan provides corpus examples of typical patterns of the entry word ***apart*** and demonstrates when it is used as an adjective (3) or as an adverb and performs the following functions: adverbial modifier (2), is part of a verb phrase (1, 5) and is part of a phrasal verb (4).

5.2.3. Treatment of collocations and phrases in *COBUILD*

In the choice of examples, also *COBUILD* pays careful attention to collocation – so the examples provide reliable models of usage. Important collocations are highlighted in the definitions, giving help with set lexical and grammatical patterns (2003: vii). In *COBUILD* collocations and set expressions tend to be highlighted within definitions, explained by definitions and then followed by illustrative examples.
For example, it looks like this:

If you give someone **advice**, you tell them what you think they should do in a particular situation. *Don't be afraid to ask for advice about ordering the meal.... Your community officer can give you advice on how to prevent crime in your area... Take my advice and stay away from him!*

In the entry ***advice*** one can hardly recognise the presence of collocations since *COBUILD* does not highlight them but exemplifies in more implicit way. Nevertheless *COBUILD* exemplifies the following typical collocational constructions for the noun ***advice***: 1. **give advice**; 2. **ask for advice**; 3. **take advice**; 4. **on the advice of sb**; 5. **legal advice**.
Although *COBUILD* lists the typical constructions of the noun ***advice***: verb + noun (1, 2, 3), adjective + noun (5) and part of prepositional phrase (4), it provides no more information on typical adjectives or verbs, which form collocations with the noun and in this respect is much less comprehensive then the other two dictionaries.

COBUILD gives the following examples for the phrases of the entry word ***apart***: 1.**move/ pull apart** *John and Isabelle moved apart, back into the sun...* 2. **take sth apart** *When the clock stopped he took it apart to find out what was wrong... Many school buildings are unsafe, and some are falling apart.* 3. **fall/ tear apart** *Any manager knows that his company will start falling apart if his attention wanders.* 4. **set smb/ smth apart** *What really sets Mr Thaksin apart is that he comes from northern Thailand.* 5. **be long way apart** *Their concept of a performance and our concept were miles apart.* 6. **tell smth or smb apart** *I can still only tell Mark and Dave apart by the colour of their shoes!*
The entry word ***apart*** as illustrated in *COBUILD* examples performs the following functions: acts as an adjective (5), acts as an adverb and is part of a verb phrase (1, 2), acts as an adverbial modifier (3) and is part of phrasal verb (4, 6).

To summarize the analysis of collocations and phrases it should be mentioned that learner's dictionaries illustrate as many collocations as possible - to provide reliable evidence on how words combine and inform the learners about the way how native speakers use them. Phrases mainly denote figurative meanings and consist of words of different meanings and therefore are in need to be not only illustrated by the examples but fully explained by the definitions. These aspects are well resolved in all the learner's dictionaries by providing extensive illustrative examples, which are taken from corpus.

5.2.4. The arrangement and frequency of phrases and collocations

Dictionaries claim that they arrange collocations and phrases according to their frequency which is verified in corpus. The table 5.2.4.1. indicates all the collocations of the noun advice in such order as they are listed in the dictionaries.

Table 5.2.4.1. **Arrangement of collocations according to their frequency order**

LDOCE	***Macmillan***	***COBUILD***
1. give advice 2. legal/ medical/ financial advice 3. professional/ expert advice 4. follow/ take sb's advice 5. a piece/ word of advice 6. on sb's advice	1. give advice 2. take sb's advice 3. legal/ medical/ expert etc advice 4. on sb's advice 5. on the advice of sb	1. give advice 2. ask for advice 3. take advice 4. on the advice of sb 5. legal advice

The most common collocation of the entry word ***advice*** as specified by *LDOCE*, *Macmillan* and *COBUILD* is 'give advice'. The least common collocation in *LDOCE* and *Macmillan* is 'on sb's advice', while in *COBUILD* it is 'legal advice' which in *LDOCE* is among the most common ones and is viewed as fairly common in *Macmillan*. Arrangement of collocations differs to a larger extent only in one case – where frequency of 'legal advice' is somehow contradicting – from least frequent (in *COBUILD*) to almost basic (in *LDOCE*). In other respects the difference can not be regarded as major. What is common is that 'give advice', 'take advice', 'on sb's advice' and 'legal advice' are present in all the dictionaries, which seem to belong to one and the same text type e.g. broadcasting. Difference lies in the number of collocations dictionaries cover – the least number of collocations is found in *COBUILD*, while the richest account of collocations is presented in *Macmillan*.

The table nr. 5.2.4.2. is compiled to list the phrases and show their occurrence as indicated in the dictionaries. Arrangement of phrases follows the principle – to list them according to their frequency order in the corpus.

Table 5.2.4.2. **Arrangement of phrases according to their frequency order**

LDOCE	***Macmillan***	***COBUILD***
1. fall apart 2. be torn apart 3. be worlds/ poles apart 4. grow/ drift apart 5. joking apart 6. sb/ sth apart 7. set sb/ sth apart	1.take sth apart 2. fall/ come apart 3. be poles apart/ be worlds apart 4. set sb apart 5. tear sth apart	1. move/ pull apart 2. take smth apart/ come/ fall apart 3. fall / tear apart 4. set smb/ smth apart 5. be long way apart 6. tell smth or smb apart

The frequency and arrangement of the phrases of the entry word ***apart*** differ - to a large extent. First, the most frequent senses in *LDOCE*, *Macmillan* and *COBUILD* are not the same. *LDOCE* claims that 'something falls apart' is the most frequent sense, *Macmillan* lists 'to dismantle something' as the most common but *COBUILD* claims 'step aside from each other' to be the first to mention. The 2nd sense in *COBUILD* coincides with *LDOCE* and *Macmillan*

and the 3rd senses are similar in *LDOCE* and *Macmillan*. *LDOCE* and *COBUILD* contain all the senses which are included in *Macmillan*, while *Macmillan* does not mention 'drift apart', 'tell smth or smb apart' and 'joking apart'. These differences are explained by the statistical data with appear as frequency counts when corpora are consulted. But the overall conclusion could be that more or less they operate with the same linguistic data and differences are minor – it seems that their corpora provide very similar data with some minor differences.

5.3. The arrangement and frequency of senses in polysemantic entries

The meanings in the dictionaries as far as possible are ordered in accordance with their frequency in the language as shown by the corpus, which means that most frequent meanings are placed first. For the analysis of this aspect, the entry word ***handle*** was chosen for it provides a number of sub-senses.

Table 5.3.1. **The arrangement of senses for the entry word *handle***

LDOCE	Macmillan	COBUILD
1.**do work** *I handled most of the paperwork* 2. **deal with a situation** *The headmaster handled the situation very well* 3. **deal with a person** *Some customers are quite difficult to handle* 4. **not become upset** *She can't handle it when people criticize her.* 5. **hold** *He had never handled a weapon before.* 6. **control a vehicle** *I didn't know if I'd be able to handle such a large vehicle* 7. **move goods** *The Post Office handles nearly 2 billion letters...* 8. **buy/ sell goods** *Bennet was charged with handling stolen goods*	1. **deal with a situation** *The government was criticized for the way it handled the crisis* 1.a. **do particular job** *Inspector Dawkin will be handling this case.* 1.b. **deal with a large amount of work** *The newer computers can handle massive amounts of data.* 1.c. **deal with people or goods** *All the staff are trained to handle difficult customers* 1.d. **deal with a person who is likely become upset** 2. **touch or hold smth** *All chemicals must be handled with care.* 3. **control an animal or a vehicle** *She handled the pony very confidently* 4. **buy/ sell goods illegally** *He denied burglary but admitted handling stolen goods*	1.**handle a problem or situation** *I don't know if I can handle the job.* 2. **deal with situation** *I think I would handle a meeting with Mr. Siegel very badly.* 3. **control a weapon, vehicle** *I have never handled an automatic* 4. **hold or move smth** *Wear rubber gloves when handling cat litter.* 5. **understand a problem** *[informal] When you have got a handle on your anxiety you can begin to control it.*

This table shows that *LDOCE* and *Macmillan* are more explicit in defining and listing of various senses, and Macmillan provides even a more detailed subdivision of meaning. *COBUILD* provides only 5 senses, while the other 2 dictionaries provide 8. Taking into account the fact that *COBUILD* is based on the largest corpus it seems rather strange that *COBUILD* has given the least amount of senses.
Differences in the ordering of senses do not pose major problems. What counts is how well a particular sense is exemplified and whether it is mentioned at all. These aspects can well be resolved by the corpus evidence.

5.4. Frequency of words

Words appear in the dictionaries according to their frequency of appearance in the corpus. This table below indicates the frequency of the entry words which were analysed in this research paper.

Table nr. 5.4.1. **Frequency of entries analysed**

LDOCE	*Macmillan*	*COBUILD*
accept S1, W1	**accept** ∗∗∗	**accept** ♦♦♦
advice S2, W2	**advice** ∗∗∗	**advice** ♦♦
apart S3, W1	**apart** ∗∗∗	**apart** ♦♦
believe S1, W2	**believe** ∗∗∗	**believe** ♦♦♦
handle S2, W2	**handle** ∗∗∗	**handle** ♦♦
inadequate (not rated)	**inadequate** ∗∗	**inadequate** (not rated)
serious S1, W1	**serious** ∗∗∗	**serious** ♦♦♦

In *LDOCE* the top 3000 most frequent words are indicated in red. *LDOCE* is the only dictionary to distinguish between the spoken and written frequency (p xi). S1, S2 and S3 stands for 'among 1000, 2000, 3000 most frequent words in spoken English' and W1, W2, W3 stands for 'among 1000, 2000, 3000 most frequent written words in English'.
In *Macmillan* dictionary some words are printed in red with a star rating to show their frequency. A word with one star is fairly common, with two is common and with three is one of the most basic words in English.
Common words in *COBUILD* have diamond symbols - with a simple code to tell how common they are (p. vii) – similar to *Macmillan's* approach.

5.5. Some distinctive features

For the entry word ***advice*** *LDOCE* provides a warning note to avoid a common mistake when the noun ***advice*** is confused with the verb ***advise***. To exemplify this *LDOCE* gives the following corpus examples: *He gave me some useful advice. Can you advise on college courses?*

This is one of the distinctive features of *LDOCE* to provide warning notes for words which seem to be confusing for learners of English. Such evidence on common mistakes is justified by the Longman Learners' Corpus – a database of over 10 million words of English written by students from around the world (2003: xv).

Definitions provided by *COBUILD* - is the feature which requires certain appraisal. It is extremely handy to be given full sentence definitions for they provide extra help in decoding of meanings and information they denote. Although definitions do not reflect corpus evidence, they are compiled for clarity and they seem very helpful in acquiring the English language.

CONCLUSIONS

To sum up the analysis of corpus application in the dictionaries for learners' it should be pointed out that the analysis of illustrative examples concerned the aspects of whether corpus sentences presented in dictionaries seemed that real and natural sentences as dictionaries claimed them to be and how much edited or changed they might have been was the other aspect in question. The analysis of theses aspects proved that the examples given in dictionaries imply certain contexts and provide clear meanings, therefore denote a true picture of corpus evidence. The way the examples are presented - short or lengthy forms with some punctuation marks used - denotes that they have been taken from the corpus and are real rather than artificially invented examples. If one looks up the examples in several dictionaries, it becomes obvious that they are different but perform their illustrative task very successfully. The examples given in full sentences function as templates of real performance, while the examples in short form function only as templates based on which one might model his/ her own utterance.

If to compare how collocations and phrases are treated in all the dictionaries it should be mentioned that the most extensive information and explicit guidance on the use of collocations (on the example analysed) is given by *Macmillan*. All the three dictionaries exemplify them in terms of typical constructions they form, arrange them in order of their frequency as defined by corpus and illustrate them with full sentence examples taken from corpus. At certain extent collocations are idiomatic constructions/ expressions and therefore are of special importance to be exemplified to learners of English. Although highlighting of collocations and phrases is not related to corpus evidence, it helps to distinguish them in the text and memorize. *COBUILD* is far less explicit in this respect in comparison with the other two dictionaries, besides it does not treat collocations in such detail as the other dictionaries.

An interesting aspect to mention is that all three dictionaries have used their corpus data which cover linguistic evidence from 'books, newspapers, magazines and spoken English' in *LDOCE* (2005: x), 'books, newspapers, magazines, broadcasting, and conversation' in *COBUILD* (2003: x) and 'books, magazines, newspapers, e-mails, television and radio' in *Macmillan* (www.macmillandictionary.com). This leads to conclusion that in this respect all three corpora are balanced in their choice of particular text genres and they are similar corpora in respect to the choice of the genres they represent. This leads to conclusion that examples of typical collocations, phrases and typical grammar patterns more or less are derived from the linguistic resources of similar structure and therefore it might be assumed that dictionaries would arrange polysemantic entries in more or less similar order. It turned out that arrangement of senses slightly differs, arrangement of collocations and phrases in frequency

order is not the same and frequency rating of entry words differs, too. This probably might be explained by the fact that even though the origin or type of linguistic resources might be similar the extent to which one dictionary or another process their linguistic data may denote the frequency of one or the other phenomena. That is why different corpora being similar in their choice of data cannot provide completely identical results.

It requires mention here - it was assumed that the largest corpus might provide most reliable evidence on the frequency of words and distribution of their senses, the result of this analysis proved that it is not quite so. *COBUILD* fails to give richest account of senses, while *Macmillan* being the dictionary which is based on the corpus of least number of words provides a fuller account of such data. It is worth to mention that the Bank of English, being the corpus of *COBUILD* dictionary, is the corpus which falls into the category of reference and monitor corpus – being huge in size and being constantly updated.

The overall conclusion of the analysis of corpus application in learners' dictionaries is such that corpus evidence makes learners' dictionaries only better – for it provides real confirmation on how English language is actually used. Dictionaries contain an immense amount of authentic examples, which specify typical contexts and structures and corpus in all respects is indispensable tool for clarifying and exemplifying meanings.

THESES

1.

Advances in computer technologies with subsequent creation of electronic text corpora have brought about significant changes in lexicography. Lexicographers have access to more reliable linguistic evidence and they can base their decisions on naturally occurring texts rather than depend on their intuition or assumptions.

2.

The capacity of text corpora is such that they allow implementation of more complex studies of language, and corpus evidence is what makes lexicographers observe new linguistic phenomena and produce better and more-up-to date dictionaries.

3.

Text corpora contain real language data – providing evidence of actual instances of speech or writing rather than consisting of made up or artificially compiled data.

4.

Text corpora were created long before computers were invented. Johnson in his *Dictionary of the English Language(1755)* used quotations of reputable authors for almost every entry word, showing the words in context and establishing authority of the dictionary.

The use of quotations in this respect is similar to the approach of contemporary ESL lexicography – where definitions of entry words are exemplified by illustrative examples, which are taken form electronic corpus.

5.

English learner's dictionaries are increasingly being based on the corpora because the corpora with the software to analyse them can automate the process of creating dictionaries and improve the quality of information which goes into the dictionaries. Automation process takes seconds to perform frequency counts of words and their ranking into most frequent or less frequent forms.

6.

Text corpora differ depending on their size and purpose. They can be classified into the following types: *balanced, representative, reference, monitor, learner, specialized, historical, diachronic, synchronic, national, parallel* etc.

An ideal corpus which fits the requirements of ESL lexicography is a representative, general**,** reference and monitor corpus.

7.

Corpora differ not only according to their types but also according to their applications. Apart from lexicography they are used in a number of studies by psycholinguistics, historical

linguists, sociolinguists, computational linguistics, in language acquisition and language pedagogy.

8.

Computer technologies have the capacity of performing automatic processing of a computer corpus data – grammatical tagging or parsing. Annotation of a corpus is particularly important in lexicography – tagged corpus provides information on word classes and provides more complete description of language use

9.

The analysis of corpus application in three English learner's dictionaries (Longman Dictionary of Contemporary English, Collins Cobuild Advanced Learner's English Dictionary and Macmillan English Dictionary for Advanced Learners) proves that the examples given in dictionaries contain contexts and provide clear meanings, therefore denote a true picture of corpus evidence. The way the examples are presented - full sentences or sentence fragments with some punctuation marks used - denotes that they have been taken from the corpus and are real rather than artificially invented. The examples given in full sentences function as examples of real performance, while the examples in shortened form function only as templates based on which one might model his/ her own utterance.

10.

The arrangement of senses within polysemantic entries in all the three dictionaries slightly differs, arrangement of collocations and phrases in frequency order is not the same and frequency rating of entry words differs, too. This might be explained by the fact that even the corpora applied are similar, the extent to which the compilers of the dictionaries process their linguistic data may denote the frequency of one or the other phenomena. That is why different corpora being similar in their choice of data cannot provide completely identical results.

11.

The analysis of corpus application in learners' dictionaries leads to an overall conclusion that the application of corpus evidence has improved the quality of learners' dictionaries for it provides real confirmation on how English language is actually used. Dictionaries contain an immense amount of authentic examples, which specify typical contexts and structures and corpus in all respects is an indispensable tool for clarifying and exemplifying meanings.

BIBLIOGRAPY:

- Aijmer, K. and Altenberg, B. (1991). *English Corpus Linguistics, Studies in Honour of Jan Svartvik*. Longman Group UK Limited;
- Aarts, J. (2002). Does Corpus Linguistics exist? Some old and new issues. In W. Teubert, and R. Krishnamurthy (2007), *Corpus Linguistics, Critical Concepts in Linguistics. Volumes I-VI.* (pp. 58-73) Routledge
- Atkins, S., Fillmore, Ch. and Johnson, Ch. (2003). Lexicographic Relevance, Selecting information from corpus evidence. In W. Teubert, and R. Krishnamurthy (2007), *Corpus Linguistics, Critical Concepts in Linguistics. Volumes I-VI.* (pp. 269-299) Routledge
- Baker, P. (2006). *Using Corpora in Discourse Analysis.* Continuum International Publishing Group Ltd.
- Bejoint, H. (2000). *Modern Lexicography: An Introduction.* Oxford University Press
- Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press;
- Correard, M.H. (2002). *Lexicography and Natural Language Processing, A Festschrift in Honour of B.S.T. Atkins.* EURALEX 2002
- Čermák, F. (2003). Source materials for dictionaries. In P. Sterkenburg van (2003). *A Practical Guide to Lexicography.* (pp. 18-25) John Benjamins B.V.
- Granger, S. (2002). A Bird's-eye view of learner corpus research. In W. Teubert and R. Krishnamurthy. (2007). *Corpus Linguistics, Critical Concepts in Linguistics. Volumes I-VI*, (pp. 44-69) Routledge
- Hanks, P. (2002). Mapping Meaning onto Use. In M. Correard. (2002). *Lexicography and Natural Language Processing, A Festschrift in Honour of B.S.T. Atkins.* (pp. 158-198) EURALEX 2002
- Ilson, R. (1985). *Dictionaries, lexicography and language learning.* Pergamon Press in association with the British Council;
- Jackson, H. (2002). *Lexicography, An Introduction.* Routledge
- Kirkpatrick, B. (1985). A Lexicographical Dilemma: Monolingual Dictionaries for the Native Speaker and for the Learner. In Ilson, R. (1985). *Dictionaries, lexicography and language learning.* Pergamon Press in association with the British Council
- Landau, S. I. (2001). *Dictionaries, The Art and Craft of Lexicography, 2nd edition.* Cambridge University Press;

- Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer, and B. Altenberg (1991). *English Corpus Linguistics, Studies in Honour of Jan Svartvik.* Longman Group UK Limited;
- Leech, G. (2002). Corpora. In W. Teubert and R. Krishnamurthy (2007). *Corpus Linguistics, Critical Concepts in Linguistics. Volumes I-VI* (pp. 3-15) Routledge
- Meyer, Ch. F. (2002). *English Corpus Linguistics, An Introduction.* Cambridge University Press
- Rundell, M. (1998). Recent Trends in English Pedagogical Lexicography. *International Journal of Lexicography*, 11(4): 315-343. Oxford University Press
- Sterkenburg, P. van (2003). *A Practical Guide to Lexicography.* John Benjamins B.V.
- Teubert, W., and Cermakova, A. (2007a). *Corpus Linguistics: A Short Introduction.* Continuum International Publishing Group
- Teubert, W. and Krishnamurthy, R. (2007b). *Corpus Linguistics, Critical Concepts in Linguistics. V*olumes I-VI, Routledge

Dictionaries:

- Carter, R. And McCarthy, M. (2006), *Cambridge Grammar of English, A Comprehensive Guide: Spoken and Written English, Grammar and Usage*, Cambridge University Press
- *Longman Dictionary of Contemporary English*, fourth edition (2005), Person Education Limited
- *Macmillan English Dictionary for Advanced Learners,* International Student Edition (2002), Bloomsbury Publishing Plc
- *Collins COBUILD Advanced Learner's English Dictionary,* fourth edition (2003), HarperCollins Publishers

Internet sources:

1) Available from http://www.pearsonlongman.com/dictionaries/corpus/index.html [Accessed May 8, 2008]
2) Available from http://www.pearsonlongman.com/dictionaries/corpus/learners.html [Accessed May 8, 2008]
3) Available from http://www.collins.co.uk/books.aspx?group=153 [Accessed May 8, 2008]
4) Available from http://www.natcorp.ox.ac.uk/corpus/index.xml [Accessed May 8, 2008]
5) Available from http://www.macmillandictionaries.com/corpus/corpus.htm [Accessed May 20, 2008]