LATVIJAS UNIVERSITĀTE

BAKALAURA DARBS

RĪGA 2018

UNIVERSITY OF LATVIA
FACULTY OF HUMANITIES
DEPARTMENT OF ENGLISH STUDIES

APPLICATION OF CORPUS DATA IN CONTEMPORARY ENGLISH LEARNERS' DICTIONARIES

KORPUSA DATU IZMANTOŠANA MŪSDIENU MĀCĪBU VĀRDNĪCĀS ANGĻU VALODĀ

BACHELOR THESIS

Mārtiņš Jānis Možeiko
Matriculation card No. mm14118
Adviser: assist. prof. Laura Karpinska

RĪGA 2018

ANOTĀCIJA

Mūsdienu leksikogrāfijā liela nozīme tiek piešķirta dažādiem korpusa izmantošanas paņēmieniem. Korpuss atklāj raksturīgākās valodas iezīmes un sniedz informāciju par noteiktu vārdu un frāžu leksiskajām pazīmēm. Bakalaura darbā pētīta korpusa datu izmantošana *Longman, Oxford* un *Macmillan* mācību vārdnīcās angļu valodā. Salīdzinošā analīze tika pielietota, lai salīdzinātu korpusa datu izmantošanu izvēlētajās vārdnīcās. Pētījuma rezultāti norādīja, ka korpusa datu izmantošana izvēlētajās vārdnīcās atšķiras diezgan pamanāmi, papildus tam rezultāti pierādīja, ka korpuss ir neaizstājama mūsdienu leksikogrāfijas sastāvdaļa. Tā kā pētījumam tika izvēlētas tikai trīs mācību vārdnīcas, ir ieteicams veikt plašākus pētījumus par korpusa datu izmantošanu citās vārdnīcās.
Atslēgvārdi: korpuss, leksikogrāfija, dati, mācību vārdnīcas, salīdzinošā analīze

ABSTRACT

Contemporary lexicography accords great importance to various corpus application approaches. Corpus reveals the most characteristic features of language and provides information about lexical behaviour of certain words and phrases. The present bachelor thesis investigates the application of corpus data in *Longman, Oxford,* and *Macmillan* English monolingual learners' dictionaries. The comparative analysis was carried out to compare the application of corpus data in the selected dictionaries. The investigation results indicated that the application of corpus data in the selected dictionaries differs quite visibly, additionally the results proved that corpus is an indispensable part of contemporary lexicography. Since only three monolingual learners' dictionaries were selected for the analysis it is advisable to carry out further studies regarding application of corpus data in other dictionaries.

Key words: corpus, lexicography, data, monolingual learners' dictionaries, comparative analysis

Contents

LIST OF ABBREVIATIONS AND ACRONYMS

ALD - The Advanced Learner's Dictionary of Current English

BNC - British National Corpus

CALD - Cambridge Advanced Learner's Dictionary

CANCODE - Cambridge and Nottingham Corpus of Discourse in English

CIDE - Cambridge International Dictionary of English

Cobuild - Collins Cobuild English Language Dictionary

CQS – Corpus Query System

DWS - Dictionary Writing System

ICE - International Corpus of English

ICLE - The International Corpus of Learner English

KWIC - Keyword in context

LDOCE - Longman Dictionary of Contemporary English

MEDAL - The Macmillan English Dictionary for Advanced Learners

MICASE - The Michigan Corpus of Academic Spoken English

MLD - Monolingual learner's dictionary

MWALED - Merriam-Webster's Advanced Learner's Dictionary

NMED - The New Method English Dictionary

OALD - Oxford Advanced Learner's Dictionary

OED - Oxford English Dictionary

Introduction

Language is undoubtedly one of the most important values a person can have to work, study, travel and prosper freely across different continents. English is the world's most widely spoken language and knowing English ensures the possibility of getting a decent job within your home country or elsewhere. People use different approaches to acquire English: some go to university, some attend private lessons, others try to learn it by themselves. There are countless ways in which English or any other language can be mastered – one of the most popular methods is the use of English dictionaries. Language dictionaries provide several opportunities. With a good dictionary people can learn the meaning of certain words, see the possible collocations of a word, understand how to pronounce a word and most importantly see how the word functions in natural language.

To provide the possibility to see how certain words or phrases function in everyday speech, developers of these dictionaries use different corpus examples from various sources such as newspapers, literary works, blogs, television, radio, etc, additionally several corpus application approaches are employed to successfully reflect the most characteristic features of the relevant headwords. There are currently six English monolingual learners' dictionaries that provide such opportunity, yet only three dictionaries will be investigated in this study: *Longman Dictionary of Contemporary English*, *Oxford Advanced Learner's Dictionary* and *Macmillan English Dictionary for Advanced Learners.* The main reason why exactly these dictionaries were selected for the analysis is because their practice is heavily influenced by various corpus application approaches.

The current study is essentially based on the use of corpus data in dictionaries mentioned above. To narrow down the scope of the study only the following five ways how corpus data is applied in the selected dictionaries were chosen for the investigation: word frequency, lemmatization, collocations, sense ordering, and corpus examples. **The goal** of a Bachelor Thesis is to investigate application of corpus data in *Longman*, *Oxford,* and *Macmillan* English monolingual learners' dictionaries.

The research questions are the following:

1. How is the corpus data applied in each dictionary selected for the investigation?
2. Which monolingual learners' dictionary applies the corpus data most effectively?

The enabling objectives to achieve the research goal are the following:

- to study theoretical resources to establish the framework of analysis for the present research

- to identify different corpus application patterns in each dictionary selected for the investigation
- to illustrate how corpus data is obtained using the Corpus Query System and the Sketch Engine for the British National Corpus
- to prepare comparative analysis in order to determine which monolingual learners' dictionary applies the corpus data most effectively
- to draw conclusions

The main sources used to establish the theoretical background are Christopher Brumfit (1985) *Dictionaries, Lexicography and Language Learning,* Anthony Paul Cowie (2009) *The Oxford History of English Lexicography,* Philip Durkin (2016) *The Oxford Handbook of Lexicography,* Anthony Paul Cowie (1999) *English Dictionaries for Foreign Learners: A History,* Philippe Humblé (2001) *Dictionaries and Language Learners,* Susan Hunston (2002) *Corpora in Applied Linguistics,* Howard Jackson (2002) *Lexicography: An Introduction,* Hans Lindquist (2009) *Corpus Linguistics and the Description of English,* Atkins and Rundell (2008) *The Oxford Guide to Practical Lexicography,* and Laura Karpinska (2015) *English-Latvian Lexicographic Tradition: A Critical Analysis.*

Brumfit (1985), Cowie (2009), Humblé (2001), and Jackson (2002) provide a general theoretical background regarding monolingual learners' dictionaries, and explain the difference between monolingual and bilingual dictionaries. Durkin (2016) and Cowie (1999) describe the historical development of English learners' dictionaries. Hunston (2002) and Lindquist (2009) establish the theoretical information regarding corpus linguistics and introduce different types of corpora. Atkins and Rundell (2008) and Karpinska (2015) explain how exactly corpus linguistics can be used in lexicography.

The methods of research applied in the study:

The research method is the comparative analysis which is used to compare the application of corpus data in three monolingual dictionaries: *Longman Dictionary of Contemporary English*, *Oxford Advanced Learner's Dictionary* and *Macmillan English Dictionary for Advanced Learners.* Comparative analysis is a research method during which one or more different items are compared on the basis of one or more criteria (Online 1). To answer the research questions the *British National Corpus (BNC)* was employed, apart from that also *the Corpus Query Processor* and *the Sketch Engine* were used to show how the corpus data is retrieved by lexicographers.

The research paper consists of three chapters. Chapter 1 focuses on the theoretical information about English monolingual learners' dictionaries and highlights the differences between monolingual and bilingual dictionaries. Chapter 2 concentrates on the theoretical

information about corpus linguistics and defines what the concept of text corpus means. Furthermore, it introduces seven most pertinent ways of applying corpus data when compiling a monolingual learners' dictionary, paying particular attention to the five ways selected for the analysis. Chapter 3 illustrates the findings of the study regarding the application of corpus data in three monolingual learners' dictionaries.

1. ENGLISH MONOLINGUAL LEARNERS' DICTIONARIES

Since the thesis focuses on the analysis of corpus data in three English monolingual learners' dictionaries, the purpose of the first chapter is to provide a general theoretical background that would help to understand the nature and the necessity of monolingual dictionaries.

Before presenting any specific information, it is essential to provide a brief definition of what lexicography really is based on several studies that have been carried out by qualified scholars and lexicographers.

According to Jackson (2013: 1), the science of lexicography can be explored from two different perspectives: the first one refers to the compilation of dictionaries (also known as 'practical lexicography' or 'lexicography practice') and the second one refers to the research of dictionaries (also known as 'lexicography theory' or 'dictionary research'). Jackson explains that dictionary study includes variety of activities, the scholar maintains that some lexicographers study the differences between: monolingual and bilingual dictionaries, general and specialized dictionaries, alphabetical and thematic dictionaries, others focus mostly on the use of corpora in dictionary making or the design and structure of dictionaries (ibid.: 1-2).

Based on the information provided by Sterkenburg (2003: 3), 'for us, looking for a definition of "dictionary" is looking for a definition of the prototypical dictionary'. Sterkenburg asserts that the prototypical or the preliminary type of dictionary is the alphabetical monolingual general-purpose dictionary (ibid.: 3). In such dictionaries one and the same language is used for the object and the description of the object. The monolingual general-purpose dictionary provides an explanation of a standard language rather than restricted language differences and is used for a pedagogical rather than scholarly reasons (Geeraerts, 1989, discussed in Sterkenburg, 2003: 3).

Jackson states that dictionaries are either print or electronic repositories of words that are arranged in alphabetical order and are often referred to as reference books about words (2002: 1). People who compile dictionaries are lexicographers and their main task is to incorporate information that they know or anticipate people will want to find and acquire, however, the scholar also suggests that a dictionary is not only a reference book, it also serves as a collection of the vocabulary of a language (ibid.: 21-22). Every single dictionary includes words and information that only a small number of people, if any, will need to gain access to, mostly because they already recognize it, or maybe because it does not simply seem necessary to them. For illustrative purposes Jackson provides an example of the definite article *the* which is rarely searched by anyone, and yet is included in every general English dictionary on the planet only because it is the most frequently used word in English and it would not be correct to exclude it, otherwise the collection of the vocabulary would not be complete (ibid.:

22). Just as Jackson, also Siddiek confirms that in the modern sense, the dictionary is that book comprising lists of words with definitions and details about them (2013: 1745). Zgusta claims that

> A dictionary is a systematically arranged list of socialized linguistic forms compiled from the speech habits of a given speech community and commented upon in such a way that the qualified reader understands the meaning of each separate form, and is informed of the relevant facts concerning the function of that form in its community (Zgusta, 1971: 197 quoted in Siddiek, 2013: 1745).

In terms of the structure, a dictionary can be observed from two different perspectives: macro-structure and micro-structure (Jackson, 2002: 25). The macro-structure indicates that a dictionary usually consists of three parts: the front matter, the body and the appendices. The main task of the front matter is to describe the characteristics and changes of the edition as well as to explain how to use the dictionary. Some other dictionaries might offer an account of list of abbreviations, the system of transcription for pronunciation, and sometimes even some historical statement of the language. The body consists of list of headwords usually printed in bold where each headword is accompanied by several chunks of information. This combination of the headword and the information of the headword forms the entry. Dictionary editors are responsible for incorporating the latest vocabulary items in socially related areas such as fashion, the environment, computing, and so on, Jackson maintains that incorporation of these items in such areas is frequently used as a main reason for issuing a new edition (ibid.). The micro-structure concerns the order of the information within the entries which almost always vary depending on the headword. The micro-structure usually consists of the following information: spelling, pronunciation, inflections, word class, senses, definition, examples of senses and usage, possible derivatives and etymology, nonetheless some dictionaries, especially learners' dictionaries, also comprise information about collocations and examples showing how the headword functions in the sentence (ibid.: 26-27).

The general goal of Chapter 1 is to introduce the essential information about monolingual learners' dictionaries, to illustrate the differences between monolingual and bilingual dictionaries, to outline the historical development of all monolingual learners' dictionaries (MLDs), to introduce the characteristic features of MLDs, as well as to explain the meaning of decoding and encoding in the field of lexicography.

1.1. Comparison of monolingual and bilingual dictionaries

Even though the study is based solely on monolingual learners' dictionaries, it is significant to provide a comparison of both monolingual and bilingual dictionaries, since the understanding

of these differences can help to realize some essential factors, which would not be noticeable by focusing merely on monolingual ones. Sub-chapter 1.1 will give the necessary theoretical information about monolingual learners' dictionary which is the main research object of this thesis.

Atkins proposes that monolingual and bilingual dictionaries are both intended for non-native speakers who know at least one language and are different in one basic way: monolingual dictionary, being non-specific to a particular user is obliged to provide the necessary information about the word to its users of any native language, while bilingual dictionary is not responsible for it (Atkins, 1985: 15). This fundamental disparity causes several dissimilarities between both types of dictionaries in terms of design, presentation, content, accessibility and capability of the production of the target language. For example, mono- and bilingual learners' dictionaries reveal certain variations related to the structure and organization of the wordlist, owing to the fact that bilingual dictionaries are more flexible and may differ from the wordlist of few thousand most regularly used words to a full compilation of entire vocabulary, while monolingual dictionaries concentrate mainly on the most frequent words in the language (ibid.: 16). For clarity purposes the table below illustrates the main differences between mono- and bilingual dictionaries according to five criteria (based on Atkins, 1985: 21).

Table 1.1 **The main differences between mono- and bilingual dictionaries**

Criteria	**Monolingual dictionary**	**Bilingual dictionary**
wordlist	usually short	often longer in bilinguals and more flexible
explanation of senses	definitions provided only in the target language	equivalents in the first language and the target language
exemplification of usage	sometimes glossed in the target language	usually translated
treatment of fixed and semi-fixed phrases	always glossed or defined in the target language	always translated, often by equally idiomatic equivalent expressions
semantic and usage information	always in a foreign language	usually in the user's native language

According to Siddiek (2013: 1746), one of the major issues in terms of bilingual and monolingual dictionaries is to determine which dictionary is better for people who learn English as a foreign language, therefore it is essential to review reasons in favour and against both types. It is suggested that there exists obvious difference in the viewpoint of students who prefer bilingual dictionaries, and language educators who favour the use of monolingual dictionaries, Siddiek supports this idea by explaining that bilingual dictionaries are superior because there are often equivalents in the source language and numerous words are culture-

bound which makes it easier to understand the word (Landau, 1989 discussed in Siddiek, 2013: 1746). A similar idea is also provided by Atkins who approves that students predominantly favour the bilingual dictionaries because they simply require less effort in comparison to the monolingual ones, however, teachers on the other hand are against bilingual dictionaries because they weaken the learners' ability to translate and thus hinder the internalization of the target language (1985: 19). Lew's research about monolingual and bilingual dictionaries based on the frequency of their use proves that out of 932 Polish learners of English, 848 or 91% prefer bilingual over 84 or 9% who favour the use of monolingual dictionaries (Lew, 2014: 96).

Mairs (Online 2) suggests that the main benefit of using a bilingual dictionary is its conciseness and clarity, students can search a word from the target language and immediately find a direct translation for this word in their first language, without trying to understand the description of the word in the target language which may seem relatively unfamiliar. Siddiek (2013: 1747), informs that this easy accessibility to the word has its negative aspects, since bilingual dictionaries provide direct translations and the language learners persistently shift from the source language which is their mother tongue to the foreign language thereby encouraging their opinion that languages are just nomenclatures and hiding meaning discriminations.

Monolingual dictionary, on the other hand, with its dependence on target language and ability to express meanings in many ways gives a large amount of grammatical, stylistic, and semantic information that advances some further language production activities and produce more learning experience than the student had initially expected (ibid.). Mairs (Online 2) sustains the idea by claiming that the principal advantage of monolingual dictionary is its ability to provide more complete and detailed information about the language as this information very often contains extra meanings of a target linguistic unit, additional examples of how the linguistic unit functions in context, as well as helpful information about the grammatical behaviour of the linguistic unit. Cowie (1999: 195) suggests that the exact connection between the use of monolingual learners' dictionaries and bilingual dictionaries will differ based on various aspects: such as the time frame of learning process, the level of linguistic competence, and the nature of the learning activities. The scholar explains the idea by assuring that a bilingual dictionary which is used for rapid decoding purposes usually fails to provide the detailed semantic, stylistic and grammatical descriptions and the collection of illustrative examples that are usually found in monolingual learners' dictionaries (Hornby, 1981 discussed in Cowie, 1999: 196). This observation confirms that monolingual

dictionaries are mostly favoured among people who have achieved a certain level of language competence to effectively perceive the detailed information this dictionary provides.

To sum up, the most important point to remember in terms of both types of dictionaries is that monolingual dictionaries present more detailed grammatical, stylistic, and semantic information that advances some further language production activities, while bilingual ones are more flexible and benefit from their conciseness and clarity, howsoever it does not mean that one of them is better than the other.

1.2. The historical development of English Learners' Dictionaries

The origin of the learners' dictionary can be traced back to the interwar years, and people responsible for the creation of this dictionary type are Harold Edward Palmer, Albert Sidney Hornby, and Michael West.

The historical development of monolingual learners' dictionaries began in 1935, when Michael West and James Endicott compiled the first monolingual learners' dictionary *The New Method English Dictionary* also known as *NMED* which consisted of 24,000 headwords and used a restricted vocabulary of 1,490 words to provide logical definitions (Cowie, 1999: 33 discussed in Durkin, 2016: 26). For the most part the dictionary was designed for decoding purposes since the syntactic guidance and the treatment of other encoding tools was inadequate (Cowie, 2009: 393).

Palmer's dictionary *A Grammar of English Words* which was published in 1938 already offered some encoding possibilities because of its well-designed verb-pattern scheme. Although Palmer's dictionary was quite advanced it was still limited because it concentrated on simple sentence patterns only, not including any complex sentence patterns (Cowie, 1999: 28).

The first general-purpose advanced-level learners' dictionary *Idiomatic and Syntactic English Dictionary* was published in 1942 by Hornby whose work was mainly based on collocations and idioms. His work was republished after the Second World War in 1948 as *A Learner's Dictionary of Current English,* however in 1952 the title was changed once again to *The Advanced Learner's Dictionary of Current English (ALD).* Since his speciality was the study of collocations and idioms, Hornby was predominantly interested to ensure effective encoding opportunities, at the same time, making sure that the dictionary also meets learners' receptive needs by adopting *Concise Oxford Dictionary* as the fundamental source of the headwords (Cowie, 2009: 398). Afterwards, two other editions were published: ALD_2 in 1963 and *Oxford Advanced Learner's Dictionary* ($OALD_3$) in 1974. ALD_2 and $OALD_3$ provided not

only a larger scope of technical and scientific terms but also a greater quantity of examples (ibid.: 403).

In 1978 Hornby's dictionary encountered its first competitive rival: Paul Proctor's *Longman Dictionary of Contemporary English (LDOCE$_1$)* which supplied different innovative features along with highly systematic organization of grammatical categories and codes (Fontenelle, 2009 discussed in Cowie, 2009: 414). *LDOCE$_1$* was also the first substantial computerized dictionary of English.

In 1987 the third learners' dictionary entered the market: John Sinclair's *Collins Cobuild English Language Dictionary (Cobuild$_1$)* which was highly appreciated because of the wide range of corpus examples it offered (Durkin, 2016: 28). *Cobuild$_1$* marked the beginning of several corpus-based approaches in lexicography which are practiced even today. In 1987 also two other editions of *LDOCE* and *OALD* were published, however both editions presented only few changes and preserved their original approaches. (ibid.).

Durkin (2016: 28) reports that the year 1995 can be considered as one of the most productive years in lexicography because within six months, four new corpus-based MLDs were released: three new editions *OALD$_5$, LDOCE$_3$, Cobuild$_2$* and another new publication, the *Cambridge International Dictionary of English (CIDE$_1$)*. *LDOCE$_3$* and *OALD$_5$* were the first ones to show frequency data about quantitative distribution of lexical units within a particular corpus. All four aforementioned dictionaries also enhanced the accessibility of the microstructure, especially *LDOCE$_3$* and *CIDE$_1$* due to index terms and phrases they provided, also known as *signposts (LDOCE$_3$)* and *guide words (CIDE$_1$)* (ibid.). After 1995 subsequent editions kept presenting small but significant innovations, for example a more complete practice of neologisms and collocations, a general expansion of headwords, coexisting with CD-ROMs and other additional elements, nevertheless all the basic innovations were already established by 1995.

The *Macmillan English Dictionary for Advanced Learners (MEDAL$_1$)* entered the market in 2002. As stated by Bogaards (2010: 21), *MEDAL$_1$* successfully adopted the same methods that had validated themselves as being profitable in previous editions. *MEDAL$_2$* was released in 2007.

The first American learners' dictionary *Merriam-Webster's Advanced Learner's Dictionary (MWALED$_1$)* emerged in 2008 and supplied approximately 160,000 example sentences but apart from that did not come up with any new approaches (ibid.: 25).

After 1995 several other editions were introduced and provided new possibilities for the language learners. Nowadays the quality of English MLDs is high and people can select any

of the following editions for their language learning: *MEDAL*$_2$ (2007), *MWALED*$_1$ (2008), *OALD*$_9$ (2015), *CALD*$_4$ (2013), *Cobuild*$_8$ (2014) and *LDOCE*$_6$ (2014) (Durkin, 2016: 29).

1.3. Characteristic features of MLDs

Since the historical development of learners' dictionaries has already been described in the previous sub-chapter, the main purpose of sub-chapter 1.3 is to provide a description of the characteristic features of MLDs.

Based on the information provided by De Cock and Granger (2004: 72), MLDs 'are dictionaries that are specially designed to cater for the needs of foreign language learners and provide all the information in the learners' target language'. The key idea was to modify the traditional monolingual dictionary and make it acceptable to people who are less proficient in language they are trying to acquire, therefore English MLD can be considered as a type of dictionary that translates hard English into easy English (Humblé, 2001: 34). Kernerman (Online 3) claims that such dictionaries can be characterized by:

- a limited wordlist (2000 – 3500 words) to define the relevant headwords, derivatives, and idioms;
- introducing sentences and other explanatory material representing the most regular uses;
- supplying comments and some other relevant facts that might be useful to the dictionary user.

Underhill introduces several advantages of using a monolingual English learners' dictionary. First of all, the users of MLDs are forced to think in English, which gives an advantage because it encourages a more rapid growth of passive vocabulary, additionally the users may benefit from the precisely formulated definitions and descriptions of different meanings (Underhill, 1985: 104). Another important asset of such dictionaries is the example sentences that illustrate typical usage and reveal how the word is used in context. Finally, Underhill states that MLDs gives a sense of satisfaction for learners who on their own are capable of solving various language problems and difficulties (ibid.).

In comparison to regular bilingual dictionaries, MLDs are more comprehensive and reliable from encoding perspective, and according to Jackson there are two principal ways in which learner's encoding needs can be satisfied (2002: 135). Firstly, MLDs provide substantial grammatical information enabling the users to form natural sentence constructions in English. Such information may include the distinction between countable and uncountable

uses of nouns, indication of inflectional possibilities of adjectives, and limitations regarding the syntactic positioning of adjectives (ibid.). Secondly, MLDs supply information about collocations, idioms, and other modes of expression (ibid.: 137-138). Such dictionaries usually reflect the most typical collocations with the help of their entry definitions. Apart from grammar and lexical patterning which are two primary types of information, Jackson also mentions indication of sense relations (synonymy, antonymy, and sometimes even hyponymy) and labelling of usage (ibid.: 139). Labelling refers to the indications of slang and taboo.

The main takeaway is that MLDs offer a large assortment of lexicographic practices and therefore are highly appreciated by language learners all around the world.

1.4. The difference between encoding and decoding in language learning

Another significant point that should be observed with regard to language learning is the distinction between encoding and decoding.

Based on the information provided by Jackson (2002: 83), 'a learner, or indeed a native speaker, may consult a dictionary when engaged in one of two broad types of language task'. During a process of reading or listening a learner may come across a word or phrase that causes confusion and whose meaning he or she cannot work out even from the context: in this case the dictionary is utilized as an assistance tool for decoding the linguistic unit. Alternatively, the dictionary is used for encoding purposes when a learner during a process of writing or preparing to speak needs to see how a known word may be used in the particular context (ibid.: 83-84). A language learner might want to consult a dictionary for encoding purposes to learn about spelling, inflections, pronunciation, collocations, behaviour of a word into grammatical structure and whether there are any restrictions in terms of its usage in social context, Jackson maintains that on this account learners' dictionaries are required to consider not only language learners' decoding needs but especially their encoding needs (ibid.). Humblé (2001: 61-62) sustains the idea by claiming that the difference between examples for encoding and the ones for decoding lies in the fact that decoding refers to the meaning of a lexical unit, while encoding is concerned with a word's syntactic features and collocates. At the present time, language learners demand more improved tools for encoding since this requires more information than decoding (ibid.: 39). According to Béjoint (1982: 210), for decoding purposes it is better to use a dictionary that offers as many entries as possible, whereas for encoding purposes the scholar suggests using a dictionary that provides the most

complete information on syntax and collocations. A language learner usually decodes the information to understand or translate the linguistic unit (Humblé, 2001: 100). The former case represents the main function of a dictionary and suggests that a language learner is reading a text, or searching for a word he or she heard, only to understand the idea, yet translation requires a practice of encoding which comes after the practice of decoding. Decoding is also less difficult than encoding, especially for advanced learners, since it is generally a smoother process than encoding (ibid.: 101). From an encoding perspective a differentiation between beginners and more advanced learners has to be made (ibid.: 124). When encoding information with bilingual dictionary, a language learner may face the subsequent problems:

- inability to understand what the word is in the target language;
- inability to remember it actively even after recognizing it;
- inability to be certain of its collocates or syntax after understanding the meaning of the word.

In Humblé's view (ibid.) the problems mentioned above will disturb any foreign language learner every now and then, nevertheless the first problem is mostly faced by beginner level learners, and the last two problems are more common for advanced level learners.

The information presented in this sub-chapter suggests that a language learner usually uses a dictionary when engaged in at least one of the following tasks: decoding or encoding. It also reveals that the difference lies in the fact that decoding is concerned with the meaning of a word, while encoding is interested in providing more complete material about a word's syntactic features and collocates.

2. CORPUS LINGUISTICS - THE STUDY OF LANGUAGE

Corpus linguistics is a study of language which has made a great contribution in various fields, but especially in lexicography. Large computerized databases of texts designed for linguistic research enables lexicographers to learn about how certain words or phrases function in regular everyday language, therefore the purpose of this chapter is:

- to define the concept of text corpora;
- to describe the key characteristics of corpus linguistics and its historical development;
- to explain how corpus linguistics can be used particularly in lexicography.

2.1. The concept of text corpora

Before providing any information about corpus linguistics or the use of corpus data in contemporary English dictionaries it is crucial to define the concept of text corpora.

According to Meyer (2004: 11-12), a corpus can be defined as collection of texts or parts of texts which can help to carry out some overall linguistic analysis, the scholar supports this idea by making it clear that there is no such thing as a corpus of proverbs to investigate proverbs, or a corpus of relative clauses to study relative clauses, alternatively a corpus is a tool which people can use to research proverbs or relative clauses. A more definite definition of corpus as the medium of communication is introduced by Paltridge who puts forward that it is typically presumed that a corpus is a collection of written or spoken real life texts which represents a specific field of language use (2012: 144). ‘A corpus is usually computer-readable and can be assessed with tools such as concordances which are able to find and sort out language patterns’ (ibid.).

McCarthy (2014: 1), not only stresses the fact that a corpus consists of written or spoken texts but also explains that corpora may vary depending on their size, for example, consisting of only 50,000 words or consisting of several billions of words. McCarthy also clarifies that the plural form of corpus is corpora. The definitions provided by Meyer, Paltridge, and McCarthy are also sustained by Sinclair who states that ‘a corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language’ (1991: 171). Based on the information presented by Björkenstam (Online 4), in the best possible way, a corpus is an assortment of language production examples created to represent a language via thorough selection of data. Language examples produced by both women and men, of different ages, and from different language production regions should be involved to build a general corpus (ibid.).

An interesting idea is offered by Hunston who specifies that a corpus by its own nature can do nothing and functions only as a collection of real life language production examples, the scholar maintains that in order to benefit from the corpus it is necessary to make use of the access software that can re-arrange that collection and analyse it in ways that are usually impossible (2002: 3). Assuming that a corpus serves as a tool for portraying a speaker's experience of language, the access software re-arranges that experience and shows how to investigate it in numerous ways (ibid.).

During the development of corpora, lexicographers have designed various types of corpora intended for different purposes. The table below lists the main types of corpora, as well as provides a brief description of each type (based on Hunston, 2002: 14-16).

Table 2.1 **Types of corpora**

Type	**Description**
Specialised corpus	A collection of texts which serves as a representative of a specific kind of text, such as geography textbooks, newspaper editorials, lectures, etc. Texts in such corpora are usually restricted to a time frame or to a social setting. Two of the most recognized specialised corpora are Cambridge and Nottingham Corpus of Discourse in English (CANCODE) and the Michigan Corpus of Academic Spoken English (MICASE).
General corpus	A collection of texts of various types, which does not serve as a representative of any particular text type. In comparison to specialised corpus, it is much larger and is used as a material to observe differences between general and specialised corpora. The British National Corpus and the Bank of English fall into the category of general corpora.
Comparable corpora	A collection of texts consisting of at least two corpora in distinct languages (such as English and Spanish), or distinct varieties of a language (such as Indian English and Canadian English). It is generally used to compare differences and similarities between languages and language varieties. A good example of comparable corpora is International Corpus of English (ICE).
Parallel corpora	A collection of texts consisting of at least two corpora in distinct languages, in which each text has been translated from one language into the other (such as a novel translated from English into Spanish, and one from Spanish into English). Such corpora are utilized not only to compare differences and similarities but also to identify possible equivalent expressions in both languages.
Learner corpus	A collection of texts produced by language learners. Learner corpus is used to identify difference between language learners and native speakers. The best-known learner corpus is the International Corpus of Learner English (ICLE).
Pedagogic corpus	A corpus comprising all the language a learner has been exposed to during the course of language acquisition. Such corpus is used not only as a tool to raise awareness but also as a comparative source to see how learner's language differs from naturally occurring language.

Historical or diachronic corpus	A corpus consisting of texts from different time periods. Such corpus looks at how a language has developed over time. Helsinki Corpus is a well-known example of historical corpus.
Monitor corpus	In comparison to the previous one, monitor corpus aims to trace present-day changes in a language. Monitor corpus is constantly enriched with new language examples, so it continually becomes greater in size.

Finally, it can be concluded that a text corpus is a collection of real language production examples which consists of written and spoken texts and serves as a means of conducting a specific type of language investigation.

2.2. Key characteristics and historical development of corpus linguistics

The origins of corpus linguistics can be traced back to the early 1960s when Winthrop Nelson Francis together with Henry Kučera created the first computer corpus, also known as the Brown Corpus which received quite a lot of scepticism from generative grammarians due to its incapacity to follow the rules of acceptable linguistic practice. (Meyer, 2004: 1). Despite this negative perception, Nelson and Kučera are considered to be founders and pioneers of corpus linguistics even today.

Although the Brown Corpus marked the beginning of electronic-corpora studies, this methodology was used long before the 1960s. Scholars refer to this period as pre-electronic corpora studies. According to Kennedy (1998: 13), pre-electronic corpora contributed to the development of linguistic investigation in five different fields:

- **Biblical and literary studies:** From the 18th century when Bible was used as a corpus to produce lists and concordances to show factual consistency of its parts.
- **Lexicography:** The use of corpus lexicography can be traced back to the 17th century. A good example is Samuel Johnson's *Dictionary of the English Language*, which consisted of approximately 40,000 headwords in conjunction with 150,000 explanatory citations (ibid.: 14). Also, *Oxford English Dictionary (OED)* published in 1928 was corpus-based and consisted of literary written English examples (ibid.).
- **Dialect:** In the 19th century when linguists assembled and interpreted different types of corpora to study linguistic variation in regional dialects. Joseph Wright's *The English Dialect Dictionary* (1898 – 1905) and Alexander J. Ellis's *The Existing Phonology of English Dialects* (1889) are considered to be two of the most significant dialect related studies in the United Kingdom (ibid.: 15).

- **Language education:** Pre-electronic corpora studies in language education first appeared in 1921, when Thorndike created 4.5 million-word corpus to improve the quality of study materials used for teaching literacy (ibid.). After that several other corpora were introduced.
- **Grammatical:** In the first half of the 20th century, newspapers and novels were used as a supply of illustrative examples for grammatical features and constructions. Otto Jespersen, Etsko Kruisinga, and Henrik Poutsma are three of the best-known grammarians who contributed in this field (ibid.: 16).

Although pre-electronic corpora studies have had a huge impact on the further development of corpus linguistics, in modern days corpus is for the most part equivalent with electronic corpus, which, as reported by Lindquist is 'a collection of texts which is stored on some kind of digital medium and used by linguists to retrieve linguistic items for research or by lexicographers for dictionary-making' (2009: 3).

Corpus linguistics differs from other branches of linguistics such as sociolinguistics, psycholinguistics or neurolinguistics in the sense that it is not even a branch of linguistics, since it does not inform what exactly is investigated, but simply tells that a specific methodology is applied (Lindquist, 2009: 1). Lindquist clarifies that corpus linguistics is a methodology, encompassing several relevant methods which can be applied to research various theoretical studies. Bonelli (2001: 1), on the other hand, claims that 'corpus linguistics goes well beyond this methodological role', Bonelli sustains this idea by arguing that corpus linguistics has some theoretical significance, and therefore it can be considered as a separate discipline (ibid.). Although the opinions differ in this respect, most scholars believe that corpus linguistics is not a separate discipline. McEnery, Xiao and Tono (2006: 7) agree with Lindquist and assert that corpus linguistics is a methodology, not a separate branch of linguistics like syntax, phonetics, semantics or pragmatics, for example. The main reason for this is because these latter studies actually characterize and clarify a particular feature of language use, while corpus linguistics is not limited and can be engaged to research almost any linguistic area (ibid.). Also, McEnery and Hardie (2011: 1) confirm that corpus linguistics is a methodology that is not intended to study any specific aspect of language, instead it is designed to concentrate on various procedures and language study methods, the scholars also talk about how these procedures and methods are still evolving, and therefore remain incomprehensible in terms of their outline. As suggested by McEnery and Hardie, corpus linguistics has the ability to:

- change the direction of the study of language;
- improve and reformulate several language theories;

- empower us to utilize different language approaches that were problematic to research before corpora and other relevant advancements were introduced;
- explore new theories of language (ibid.).

According to Meyer (2004: 11), specialists in language subjects have learnt that corpus can be a very effective tool for proceeding different types of studies, for example, lexicographers who have discovered how to compile dictionaries more efficiently by using large corpora databases. Nevertheless, lexicography is not the only linguistic discipline where corpora have been introduced. In total Meyer lists eight other disciplines, namely: grammatical studies of specific linguistic constructions, reference grammars, language variation, historical linguistics, contrastive analysis and translation theory, natural language processing, language acquisition, and language pedagogy (ibid.: 11-27). The use of corpus in grammatical studies provides linguistic information of different grammatical constructions regarding their forms, frequency, specific contexts in which they take place, as well as their communicative capability (ibid.: 11-12). In addition to that it is possible to use this linguistic information as the foundation for writing a reference grammar of English (ibid.: 13). Corpus linguistics can also be useful when investigating language variation, which studies how language changes under the influence of several variables such as: gender, social class, and age (ibid.: 17-18). Historical linguistics, on the other hand, can be applied to see not only how language changes on the basis of these variables in earlier periods of English, but also how language has changed from the past to the present (ibid.: 20). Corpora allow various types of contrastive analyses and progress advancements in translation theory, however, to make such studies possible the software must be capable to align corpora sentences of two different languages to see the differences of the object under study in the source and in the target language (ibid.: 22-23). For explanatory purposes Meyer introduces the study conducted by Hasselgård (1997) in which she observed the differences of sentence opening grammatical structures in English and Norwegian. A slightly different, more computational rather than linguistic related discipline where corpus can be employed is natural language processing whose aim is to improve the development in such areas as information retrieval, parsing, tagging and speech recognition (ibid.: 24). The second to last discipline is language acquisition in which corpora ease the study of first- and second-language acquisition (ibid.: 26). To ensure an effective functioning of language learning, several scholars have begun improving learner corpora: corpora consisting of written and spoken materials produced by people studying English as their second language. The final field in which corpus linguistics can make a major contribution is language pedagogy; in this area, the above-mentioned learner corpora provide useful information to improve the quality of several teaching methods (ibid.: 27). For instance, Altenberg's and Tapper's research in 1998

helped to discover that the best method for teaching the use of connectives to Swedish learners of English is the exposure of a greater scope of registers and a more substantial preparation in expository writing, due to the excessive use of informal connectives in their writing.

2.3. Corpus linguistics in lexicography

The previous sub-chapter looked at eight major disciplines in which corpus linguistics has made a significant contribution. Since the main research object of this study is a monolingual learners' dictionary, the purpose of the sub-chapter 2.3 is to explain how exactly text corpora influence dictionary making.

According to Béjoint (2000: 97), lexicographers started employing corpus as a source of authentic texts from the middle of the eighteenth century. The corpus during this period was mainly used to provide quotations, and the person who should be mentioned as the main initiator of this practice is Samuel Johnson (ibid.). Johnson used his own collections of quotations for various reasons:

> to illustrate the meaning of words in context, to establish that a word had been used by a reputable authority, to display how words were used by the best authors, to show the language as it was at an earlier era before it was contaminated by foreign influences, and to impart useful lessons and moral instruction. (Morton, 1989: 154-155, discussed in Béjoint, 2000: 97-98)

The use of authentic texts as a source of quotations was a principal revolution in lexicography, however, despite the fact that it has affected the present-day lexicography axiomatically, the entire ideology has been changed (Béjoint, 2000: 98). The difference lies in the fact that the eighteenth-century lexicographers preferred to work with carefully selected texts which would help them to show good usage, modern dictionary makers, on the other hand, have preferred to work with non-selective texts to reflect more objective and reliable description (ibid.). A similar idea is also provided by Atkins and Rundell who confirm that corpus does not promote pure language, and that choosing texts based on their quality is essentially in conflict with the principles of corpus linguistics (2008: 55). Atkins and Rundell (ibid.: 45) also suggest that any well-compiled dictionary provides a reliable information about word and its behaviour in normal everyday life when taking part in real communicative activities. In order to effectively illustrate this information, the application of corpus is an indispensable part.

There are different approaches of using corpora in dictionary making. Many publishers prefer adapting, updating, and editing already existing dictionary according to corpus data (ibid.: 97). Although it undeniably enriches the content of the dictionary, it also can hinder the

editing team's work especially when inadequate budget does not allow for effective corpus analysis, therefore it is advisable to start over from the beginning and advance systematically from corpus to dictionary (ibid.). Atkins and Rundell assert that the most cost-effective way of writing a monolingual dictionary is by using a twofold approach which consists of analysis and synthesis processes, the scholars maintain that each lexicographer is different: one person might be more proficient at analysis, while some other person at dictionary-entry writing (synthesis) (ibid.: 98).

Before looking at how corpus data can be applied in contemporary lexicography, which is the main part of the study, it is crucial to understand what analysis and synthesis processes involve. Analysis is a process in which lexicographers from an already existing corpus compile a database that will systematically store all the chosen facts about the word to facilitate the task for dictionary editors who are responsible for the fashioning of the final dictionary entry (ibid.: 100-101). Database entries are similar to dictionary entries, yet they are more thorough and comprise a large number of corpus examples displaying the headword in various contexts, as well as illustrating a variety of its meanings and patterns. Such database not only enables dictionary compilers to perform computerized searching and filtering, but also enables other application builders to use the carefully selected information in preparation of other systems such as information retrieval tools and machine-assisted translation systems (ibid.).

The next stage of dictionary making is the synthesis part, during which experienced and skilled lexicographers produce the final entry (ibid.: 102). Generally, the synthesis phase is fairly uncomplicated, nevertheless both skills and experience are absolutely crucial attributes when dealing with more challenging tasks such as deciding on the dictionary senses and formulating their definitions (ibid.). Atkins and Rundell particularly highlight the fact that every user with sufficient level of knowledge about dictionary use should not have problems when dealing with any dictionary entry, if problems arise then it is the fault of the responsible lexicographer not the dictionary user (ibid.: 103).

In order to perform analysis and synthesis of information in lexicography two types of systems known as *a Dictionary Writing System (DWS*) and *a Corpus Query System (CQS)* are used. *DWS* is a software system used for the purpose of information synthesis and serves as a tool to compile dictionary text onscreen (ibid.: 113). One of the easiest input tools utilized by many dictionary compilers is a generic XML editing program Emacs which can be specifically adapted for work in the field of lexicography (ibid.). Emacs allows lexicographers to insert information about a word between different pairs of tags as well as helps to automate habitual tasks and administrative processes, thus boosting lexicographer's productivity (ibid.:

114-115). *CQS,* on the contrary, is a program responsible for the analysis process and it can be used to collect data about any of the following categories of information:

- headwords, variant forms, and their meanings
- word senses and their order
- combinational nature (syntactic preferences and collocations)
- labels of items that are typical of a given region, time period, or style (ibid.: 103-104).

The choice on what to include in the dictionary will depend on how easily the *CQS* enables the retrieval of the aforementioned data, therefore the *KWIC* concordance *(keyword in context)* and the lexical profiling tool *(the 'Word Sketch')* are invaluable assets of corpus lexicography (ibid.). For illustrative purposes Atkins and Rundell provide an example of a concordance list for the verb *taste,* which has been produced from the *BNC (British National Corpus).* The authors explain that *CQS* allows to look at the node word *(taste)* in the surrounding context of the corpus sentences and helps in performing activities such as lemmatization and POS-tagging (ibid.: 105). Keyword in context tool automatically extracts all the inflected forms of the verb *taste*: *tastes, tasting,* and *tasted.* In lexicography this process is known as lemmatization. POS-tagging, which is another feature of the *BNC,* makes it possible to separate all the occurrences of the word *taste* as a noun and as a verb. Apart from lemmatization and POS-tagging that Atkins and Rundell state to be the main benefits of *CQS,* also frequency information is of great importance when working with this program. Any well designed *CQS* will provide a large spectrum of frequency information, showing the number of cases of a particular phrase, headword, or recurring pattern such as *taste* + an adjective (ibid.: 108) The frequency information is extremely important in the dictionary compilation process because it helps to decide upon many linguistic judgments, especially when determining what to incorporate, how much to report about it, and in what order to arrange it (ibid.). The second asset of corpus lexicography is *the Word Sketch* tool which works as a statistical summary database and discloses the most noticeable aspects about the combinational nature of the word (ibid.: 109). In other words, *the Word Sketch* is employed to identify collocational patterns and to give revealing apprehension of a word's behaviour and uses (ibid.: 110).

Although some of the ways how the corpus data can be applied in monolingual learners' dictionaries have already been addressed, the intention of the following sub-chapters is to tackle each issue in more detail.

2.4. Application of corpus data in a dictionary

As mentioned previously, text corpora have a significant influence on dictionary making and based on the information provided by Karpinska (2015: 38-40), the following bullet points list seven most pertinent ways of applying corpus data when compiling a monolingual dictionary:

- corpus provides the frequency information about a word and displays a number of different word forms (also known as lemmatization as discussed before);
- corpus reveals how a word functions in natural language, and therefore is capable of detecting various senses of that particular word;
- corpus illustrates combinational behaviour of word co-occurrence (collocations and phraseology);
- corpus supplies authentic examples of usage;
- corpus offers information about any social or regional area in which the usage of a word is restricted;
- corpus indicates the frequency of different spelling patterns;
- corpus gives data and statistics about grammatical patterns of a word.

More detailed information will be provided only for the first four corpus application ways, since the investigation of the selected dictionaries focuses solely on those.

2.4.1. Lemmatization and frequency

As discussed previously, lemmatization is the process through which a collection of lemmas is being produced using a concordance list of all the possible word forms (Online 5). Durkin asserts that lemmatization provides the opportunity to automatically filter, search, and list the information about the particular lemma, without the need to identify all the word forms manually (2016: 77). Atkins and Rundell (2008: 88) explain that the automatic filtering, searching, and listing is done by a special lemmatization program which arranges different word forms accordingly to the lemma they apply to, the scholars maintain that because of the simple morphology of English, the process described above can be carried out more easily than in other languages. A description of how a single query produces all the instances for a specific lemma is not needed because the process has already been introduced in sub-chapter 2.3. It is quite obvious that lemmatization is a very important process when compiling dictionary entries, since a large number of monolingual electronic dictionaries add a separate section of all the possible word forms.

Another considerable benefit given by the corpus is the frequency information that has been widely used by many dictionary enterprises. According to Jackson (2013: 197) such information is crucial when determining the frequency of different vocabulary items, meanings, and patterns. Summers (Online 6) emphasizes the necessity of such information even more by indicating that every single detail of lexicography is affected by frequency and that such information could be used as an assistance material when making various linguistic decisions. As an example, Summers presents the CLAWS tagging system which provides the opportunity to generate frequency apportionments for words that function as nouns and verbs (such as *pull* – the verb and *pull* – the noun) (ibid.). Such information can be particularly useful when deciding upon the order and arrangement of different definitions in the dictionary entry. The importance of frequency data and its impact on the dictionary entry content will be addressed more in the following sub-chapters.

2.4.2. Senses and sense ordering

Sense is a fundamental element of a dictionary and can be defined as one of the most essential parts of the entry (Piotrowski, 1994: 21 and Lew, 2013, discussed in Jackson, 2013: 284). Additionally, senses are listed in a sequence using a particular number, sign, or symbol (ibid.). At first it may seem that the sense ordering is random, yet looking more closely it is evident that the order often reveals certain tendency because the senses are arranged according to some specific standard. Atkins and Rundell introduce three most commonly used ordering patterns (2008: 250). The first ordering method is the historical order in which the senses are arranged in the sequence starting with the earliest ones and then followed by the most recent ones. The second ordering system is the semantic order. When employing this ordering system, it is first necessary to identify the core meaning of the word which will be the first sense in the dictionary entry, afterwards the core meaning is followed by other senses that are semantically closest (ibid.: 251). The final and the most important system for this particular thesis is frequency order approach in which the senses are arranged according to their frequency in corpus. In comparison to the two previous methods, frequency order is more objective and appropriate for the language learner since the most recurrent senses are the ones people are probably looking for. Lew, on the other hand explains, that such ordering system is not always the best alternative, especially for advanced learners who will more likely be interested in less popular senses, because the most frequent ones are probably already known (Lew, 2013, discussed in Jackson, 2013: 292). All the relevant information

about sense ordering in *LDOCE, MEDAL,* and *OALD* can be found in sub-chapter 3.2 of the thesis.

2.4.3. Combinational nature and collocations

Apart from lemmatization, frequency, and sense ordering, corpus also retrieves the information about collocations, thus helping lexicographers to study different patterns of words and their co-occurrence (Otlogetswe, 2011: 44). According to Sinclair (1991: 170), a collocation is 'the occurrence of two or more words within a short space of each other in a text'. Otlogetswe explains, that in lexicography collocation retrieval is carried out via the previously mentioned computer concordance list, which displays an account of the most regular collocations or words that come together and establish fixed relationships, the scholar maintains that such account also helps to make a decision concerning the order in which collocations should appear in the dictionary entry, since concordances usually provide statistics about relative frequency (ibid.). A large number of lexicographers such as Howard Jackson, Susanne Handl, and Philippe Humblé support the inclusion of the most typical collocations in contemporary learners' dictionaries and the practical application of how such information is obtained and represented will be described in Chapter 3.

2.4.4. Corpus examples

Since many lexicographers emphasize the importance of authenticity, the final sub-chapter describes the application of corpus examples in contemporary English dictionaries. Atkins and Rundell (2008: 452) claim that example sentences lexicographers work with during the synthesis stage are a crucial constituent of the dictionary compilation process, furthermore the database described in sub-chapter 2.3 provides raw materials lexicographers can use when composing a dictionary entry. The three main functions of examples are as follows:

- **Attestation:** The main purpose of this function is to determine the origins and track the evolution of a word. One of the dictionaries that selects the examples on the basis of this function is *Oxford English Dictionary* (ibid.: 453). It is also suggested that in such cases corpus examples may be shortened.
- **Elucidating meaning:** This function implies that a well-selected example can elucidate sense dissimilarities when dealing with polysemous words, put differently

corpus examples must ensure that the entry word can be easily understood without its definition (ibid.: 454).

- **Illustrating contextual features:** The final function suggests that examples are largely responsible for depicting the contextual range of a word including syntax, register, or collocation. Without such representation a language learner may encounter difficulties in understanding the proper usage of a word (ibid.).

To illustrate how a bad example looks like, Atkins and Rundell introduce the following two samples from the first edition of *COBUILD:*

- **Gravitate** He gravitated naturally, to Newmarket.
- **Grudge** (verb) Not that she grudged it. (ibid.: 457).

Although the examples are authentic, they do not clearly explain the meaning of both verbs and do not provide a comprehensible context that would help to decipher it, therefore, apart from authenticity Atkins and Rundell also stress the importance of naturalness, typicality, informativeness, and intelligibility (ibid.: 457-458). The task of the lexicographer is to select and modify the corpus examples in accordance with all the aforementioned principles.

To sum up, corpus provides a variety of corpus application opportunities in contemporary lexicography. Some of the most important ones, and the ones selected for the particular study include frequency statistics, information about lemmatization, sense ordering, and collocations, apart from that corpus also contributes in the selection of illustrative examples. It is in the interest of each dictionary publisher to build a team of proficient lexicographers who are capable of creating high quality dictionaries using different corpus analysis systems and tools.

3. THE ANALYSIS OF CORPUS DATA IN THREE ENGLISH MONOLINGUAL LEARNERS' DICTIONARIES

3.1. Methodology

Sub-chapter 3.1 introduces the research methodology applied in the present study which is undertaken to investigate the application of corpus data in three English monolingual learners' dictionaries – *Longman Dictionary of Contemporary English (LDOCE), Oxford Advanced Learner's Dictionary (OALD),* and *Macmillan English Dictionary for Advanced Learners (MEDAL).*

The goal of a Bachelor Thesis is to investigate application of corpus data in *Longman, Oxford,* and *Macmillan* English monolingual learners' dictionaries.

The main reason why exactly these dictionaries were selected for the analysis is because their practice is heavily influenced by various corpus application approaches.

To be more precise, the main focus of the study is to compare the application of corpus data in the aforementioned dictionaries. The following five corpus application methods were selected for the investigation: word frequency, lemmatization, collocations, sense ordering, and corpus examples.

Since the study analyses application of corpus data in *Longman, Oxford,* and *Macmillan* electronic dictionaries it is important to briefly introduce each dictionary separately. *LDOCE* was first published in 1978 by Longman publishing company (Cowie, 1999: 105). Currently there are six editions of *LDOCE,* the sixth one first published in 2014. *OALD* was first published in 1948 by A.S. Hornby (Cowie, 2008: 398). Currently there are nine editions, the last one first published in 2015. *MEDAL* was first published in 2002 by Macmillan Education (Durkin, 2016: 29). At the moment, there are two editions, the most recent one first published in 2007. For this study, only online versions of selected dictionaries were employed, since they are more up-to-date.

Another objective of the research was to show how corpus data is retrieved, therefore the *British National Corpus (BNC)* was employed. *BNC* is a corpus of 100 million words that was created by Oxford University Press (Online 7). Additionally, *the Corpus Query Processor (CQP)* and *the Sketch Engine* tool were used to analyse corpus data and to show exactly how the data is obtained before composing the dictionary entry. The comparative analysis was used to compare the application of corpus data in three monolingual dictionaries:

Longman Dictionary of Contemporary English, *Oxford Advanced Learner's Dictionary* and *Macmillan English Dictionary for Advanced Learners.*

The following research procedure was applied to examine how corpus data is represented in the aforementioned MLDs:

- First, five criteria of analysis were selected (criteria: word frequency, lemmatization, collocations, sense ordering, and corpus examples)
- Then, a collection of ten entry words for the investigation of each criterion were chosen
- Further, *the British National Corpus (BNC)* of one hundred million words was employed to show how corpus data is retrieved
- After that all entry words were analysed depending on the relevant criterion
- Finally, the discussion of the results was introduced.

The selection of entry words for investigation was not random and was performed in an orderly fashion to make sure that the representation of the selected criteria is as comprehensible and effective as possible. A new set of ten headwords was selected for the analysis of each criterion. In order to research sense ordering, the selection of polysemous words is the only alternative, however a completely different set of words is needed when investigating the application of collocations, for example. The following table displays all five sets of entry words selected for each criterion of analysis.

Table 3.1 **Five sets of entry words selected for the investigation**

CRITERIA	***ENTRY WORDS***
Word frequency	*executive, weapon, enemy, werewolf, virtual, commercial, emotional, fix (verb), edit (verb), hinder*
Lemmatization	*write, economy, computer, school, water, stretch, music, break (verb), thunder (noun), intelligent*
Senses and sense ordering	*fan, palm, goal, summit, buck, pupil, blink, log, coke, letter*
Collocations	*independence, policy, breakfast, democracy, expense, moment, employment, bank, fame, emotion*
Corpus examples	*dangerous, exit (noun), doll, downtown, triangle, official (adjective), careless, parrot, sky, fluent*

3.2. Results and discussion

The remaining part of the thesis is devoted to the analysis of the findings concerning application of corpus data in the selected English MLDs. Since corpus data in contemporary

lexicography can be applied in a multitude of ways, it is crucial to first provide a general overview of how the five corpus data application ways selected for the investigation are represented in each dictionary. Table 3.2 provides such overview.

Table 3.2 **Application of corpus data in *Longman, Oxford,* and *Macmillan* MLDs**

CORPUS DATA APPLICATION WAYS	***LDOCE***	***OALD***	***MEDAL***
Corpus and its size	Longman Corpus Network of 330 million words	Oxford Corpus Collection	The World English Corpus of 200 million words
Representation of word frequency	In *LDOCE* the 3000 most frequent words are marked by using three dots, in addition *LDOCE* differentiates between the frequency of words in written and spoken language	In *OALD* the 3000 most frequent words are denoted via a small key	In *MEDAL* the 3000 most frequent words are denoted via three stars
Representation of different word forms (lemmatization)	The list of the entire word family (including all nouns, adjectives, verbs, and adverbs) is the first information introduced in the dictionary entry	*OALD* provides only plural forms for nouns and all the tense forms for verbs	*MEDAL* provides only plural forms for nouns and all the tense forms for verbs
Senses and their ordering	Senses are ordered according to their frequency in corpus, which means that the most frequent ones are at the top of the list	Senses are ordered according to their frequency in corpus, which means that the most frequent ones are at the top of the list	Senses are ordered according to their frequency in corpus, which means that the most frequent ones are at the top of the list
Representation of collocations	Collocations are located before the examples and sometimes highlighted in bold, additionally the most typical collocations are presented in a separate collocation box	The most typical collocations are introduced in a separate section *Oxford Collocations Dictionary*	The most common collocations are identified using the *Word Sketch Engine* and are listed in a separate collocation box and sometimes highlighted in bold
Representation of corpus examples	Modified and unmodified examples are presented directly under each	Modified and unmodified examples are introduced under each definition, apart	Modified and unmodified examples are presented only under each definition

	definition, additionally a multitude of longer corpus examples are provided separately in a section *Examples from the Corpus*	from that a list of other examples is provided in a separate section *Extra examples*	

Most of the information about the application of corpus data was found in the front matter of each dictionary, yet some conclusions and judgments were made based on the analysis of the lexicographic material. The main purpose of the following sub-chapters is to investigate each criterion selected for analysis in more detail.

3.2.1. Representation of word frequency

As mentioned previously, almost every single detail in lexicography is affected by frequency, therefore the goal of sub-chapter 3.2.1 is to indicate how the information about word frequency is shown in each dictionary. The following table clearly illustrates the frequency of the selected entry words.

Table 3.3 **Frequency representation of the selected entry words**

LDOCE	***OALD***	***MEDAL***
executive ●●○ W3	*executive*	*executive* ★★
weapon ●●● W2	*weapon*	*weapon* ★★★
enemy ●●● W2	*enemy*	*enemy* ★★★
werewolf -	*werewolf* -	*werewolf* -
virtual ●●○ AWL	*virtual* -	*virtual* ★★
commercial ●●○ S3 W3	*commercial*	*commercial* ★★★
emotional ●●● S3 W3	*emotional*	*emotional* ★★★
fix (verb) ●●● S2 W2	*fix(verb)*	*fix(verb)* ★★★
edit (verb) ●●○	*edit (verb)* -	*edit (verb)* ★★
hinder ●○○	*hinder* -	*hinder* -

All three dictionaries selected for the research use some specific symbol to show the word frequency, however in some dictionaries such information is presented more comprehensively than in others. For example, *OALD* highlights the 3000 most frequent words with a small key symbol, *MEDAL*, on the other hand, indicates the 7500 most frequent words in red, additionally dividing such words in smaller categories by using three-star rating approach. The more stars are given the more frequent the word. *LDOCE* uses similar method to *MEDAL* and indicates the 3000 most frequent words via three-dot rating approach. *LDOCE* is more advanced in comparison to *OALD* and *MEDAL* because it also distinguishes between

the frequency of words in written and spoken language. For instance, W2 means that a word is among the 2000 most frequent words in written language, on the contrary, S3 means that a word is among the 3000 most frequent words in spoken language. *OALD* lags behind *LDOCE* and *MEDAL* because it does not divide the most frequent words in finer categories.

The material in Table 3.3 also show that the word frequency in both *LDOCE* and *MEDAL* is similar for the most part. The reason why the word frequency slightly differs in some cases is because different corpus is used by each dictionary.

3.2.2. Representation of different word forms (lemmatization)

As discussed previously, corpus also reveals different word forms. This is done by using a *Sketch Engine* which generates a variety of concordance lines. The concordance lines show a display of every instance of a specified word. Since *the British National Corpus* is lemmatized such information can be retrieved quite easily. *The Sketch Engine* for the *BNC (https://the.sketchengine.co.uk)* is used to show exactly how such material is obtained. In total ten entry words were selected for this part of the analysis, however only two will be described in more detail.

Figure 3.1 **The concordance list for the first twenty results of the lemma '*write*'**

Written bo...	the Woolley or the Horsman Fellowship should	**write**	to the Principal enclosing a s.a.e.for an
Written bo...	and return before December. 1913 Gladys Hill	**writes**	(aged nearly 96) that she enjoys reading a wide
Written bo...	balance. Her condition is irreversible and she	**writes**	, 'this is not so much news as an appeal to
Written bo...	from Latin into Esperanto the Somerville song	**written**	in 1903 by Helen Darbishire, Margaret Moor and
Written bo...	British Academy. Lucia Glanville (Mrs Turner)	**wrote**	to say that, sadly, her elder daughter, Rosalie
Written bo...	Studies, Lancaster University. She has	**written**	various articles on medieval theatre. 1955
Written bo...	awareness of world population problems. She	**writes**	'I would like very much like to hear from any
Written bo...	(Mrs Raza) has since 1986 been a journalist	**writing**	for a current affairs publication. Claire O'
Written bo...	and career.' Janice Flook (Mrs Boniface)	**writes**	'after a varied career during which I migrated
Written bo...	of Islamic Art. 1969 Jacqueline Clements	**writes**	'In 1989 I became a partner in the Lincoln's Inn
Written bo...	ode Hail Bright Cecilia. Ruth Thompson	**writes**	'I was promoted to Grade 5 (Assistant Secretary
Written bo...	with the Met. and moved to St. Albans. She	**writes**	, 'We live very happily here with not enough
Written bo...	Nottingham. 1974 Clare Lawrence (Mrs Hatcher)	**writes**	, 'Since qualifying as a solicitor
Written bo...	a G.L.C. funded organisation to combat it, and	**wrote**	a book about it. In 1988 she was called to the bar
Written bo...	In May she came to Oxford for her D. Phil viva. She	**writes**	, 'The children are thriving and our jobs give us
Written bo...	Maxwell's office as the Corporate Analyst. She	**writes**	, 'The work is interesting, ever changing and
Written bo...	and take Second B. Mus. there, and afterwards	**write**	my Exercise - as that was financially possible
Written bo...	restricted motor traffic. Mary Humphrys	**wrote**	: 'I fell in love - with Oxford itself and with its
Written bo...	dons worked extremely hard. Barbara Davies	**writes**	: 'It is astonishing how Somerville, as well as
Written bo...	in her garish red-gilt rooms in 41 St. Giles',	**writes**	Margaret Brown). That we were grateful to Miss

Figure 3.1 shows how a regular concordance list looks like. The first twenty concordance lines already reveal five different inflected verb forms for the query lemma *write.* Based on such information, lexicographers can create a dictionary entry and include all the identified word forms, which in this particular case are *write, writes, written, wrote,* and *writing. The Sketch Engine,* therefore serves as an indispensable tool a lexicographer can use to save both time and effort.

Table 3.4 provides a comparison of how such information for the same query lemma is indicated in the selected dictionaries.

Table 3.4 **Representation of different word forms for the entry word *write* in the selected dictionaries**

LDOCE	***OALD***	***MEDAL***
Word family (noun) *writer, writing, rewrite* (adjective) *written ≠ unwritten* (verb) *write, rewrite*	present tense I/you/we/they *write* he/she/it *writes* past simple *wrote* past participle *written* -ing form *writing*	present tense I/you/we/they *write* he/she/it *writes* present participle *writing* past tense *wrote* past participle *written*

As shown in Table 3.4, *OALD* and *MEDAL* clearly use similar approach and offer all the relevant verb forms. *LDOCE*, on the other hand, indicates a list of inflected nouns, adjectives, and verbs under a separate section called *word family*. Representation of various tense forms is definitely more superior in *OALD* and *MEDAL*, however *LDOCE* provides more substantial information about different parts of speech, therefore it is difficult to tell which method is better.

A slightly different system is applied to display different word forms for nouns, adjectives, and adverbs. Table 3.5 is intended to reveal how the application of various word forms differs in the case of the entry word *economy*.

Table 3.5 **Representation of different word forms for the entry word *economy* in the selected dictionaries**

LDOCE	***OALD***	***MEDAL***
Word family (noun) *economics, economist,* (adjective) *economic, economical ≠ uneconomic(al)* (verb) *economize* (adverb) *economically ≠ uneconomically* (singular) *economy* (plural) *economies*	(singular) *economy* (plural) *economies*	(singular) *economy* (plural) *economies*

In this example *LDOCE* introduces two different noun forms, three different adjective forms, one verb form, two adverb forms, as well as singular and plural forms. Such information in *OALD* and *MEDAL* is far less substantial since both dictionaries present only singular and plural forms. A similar tendency was observed when investigating representation of various forms for other words (see Appendix 1).

The main conclusion to be drawn from the above observations is that in most cases *LDOCE* evidently provides a more comprehensive account of different word forms, whereas tense form representation of verbs is more intelligible in *OALD* and *MEDAL*.

3.2.3. Senses and sense ordering

In the theoretical part of the Thesis, three most commonly applied sense ordering patterns were introduced. To identify which pattern is used by each dictionary ten polysemous entry words were selected. The investigation confirmed that all three MLDs use the frequency order approach in which senses are arranged according to their frequency in corpus. Of all the selected dictionaries only *LDOCE* introduced an account of what sense ordering pattern is used in its dictionary entries. Knowing that senses in *LDOCE* are arranged according to their frequency in corpus, it was easy to draw conclusions about the two other dictionaries.

Table 3.6 **Sense arrangement for the entry word *palm* in the selected dictionaries**

LDOCE	***OALD***	***MEDAL***
1. **(part of hand)** the inside surface of your hand, in which you hold things 2. tree **(also palm tree)** a tropical tree which grows near beaches or in deserts, with a long straight trunk and large pointed leaves at the top	1. the inner surface of the hand between the wrist and the fingers 2. **(also palm tree)** a straight tree with a mass of long leaves at the top, growing in tropical countries.	1. the inside part of your hand, between your fingers and your wrist 2. a palm tree, or a large plant similar to a palm tree

In order to investigate the application of this criterion of analysis the entry word *palm* was selected. *Palm* is a polysemous word with two different senses: *palm as a part of hand,* and *palm as a large tropical tree.* According to Table 3.6 both senses of the noun entry *palm* in each dictionary are arranged identically, accordingly the logical conclusion is that all three dictionaries use the same ordering pattern.

One example is not enough to justify that the same practice is indeed employed in all three dictionaries, therefore the following table shows the arrangement of senses for another polysemous noun entry *fan.*

Table 3.7 **Sense arrangement for the entry word *fan* in the selected dictionaries**

LDOCE	***OALD***	***MEDAL***
1. someone who likes a particular sport or performing art very much, or who admires a famous person 2. a) a machine with turning blades that is used to cool the air in a room by moving it around	1. a person who admires somebody/something or enjoys watching or listening to somebody/something very much 2. a machine with blades that go round to create a current of air 3. a thing that you hold in your hand and wave to create a current of cool air	1. someone who likes watching or listening to something such as a sport, films, or music very much, or who admires a famous or important person very much 2. a machine with blades that turn and move the air in a room to make it feel less hot 3. a flat object that you move backwards and

b) a flat object that you wave with your hand which makes the air cooler		forwards in front of your face in order to make yourself feel less hot

Fan is another polysemous noun that has three meanings: *fan as a person who likes and admires something/someone very much, fan as a machine that is used to cool the air,* and *fan as a flat object people can use to cool the air.* Table 3.7 clearly shows that all three senses in each dictionary follow the same sequence, which only sustains the previously discussed finding. The only difference lies in the fact that *LDOCE* divides the last two senses in sub-senses *a)* and *b),* whereas *OALD* and *MEDAL* distinguish both as two different senses.

In three out of the ten selected entries, the sense ordering sequence slightly differed. Table 3.8 reveals such discrepancy in the case of the entry word *summit* which is a polysemous word with two different senses: *summit as an important meeting or series of meetings,* and *summit as the top of the mountain.*

Table 3.8 **Sense arrangement for the entry word *summit* in the selected dictionaries**

LDOCE	*OALD*	*MEDAL*
1. an important meeting or set of meetings between the leaders of several governments 2. the top of a mountain	1. the highest point of something, especially the top of a mountain 2. an official meeting or series of meetings between the leaders of two or more governments at which they discuss important matters	1. a meeting or series of meetings between leaders of two or more countries 2. the top of a mountain 3. the highest level of achievement in something

LDOCE and *MEDAL* clearly use the same arrangement pattern, however *OALD* introduces both senses in the opposite order. As has already been mentioned, only three ordering deviations have been identified, but strangely, such differences were recognized only in *OALD* entries. *MEDAL* and *LDOCE* in all ten cases use identical approach. Such conclusion suggests that *OALD* not always follows the frequency order entirely, and probably combines it with semantic ordering pattern in which the core sense will always be the first one in the dictionary entry, followed by other senses that are semantically closest.

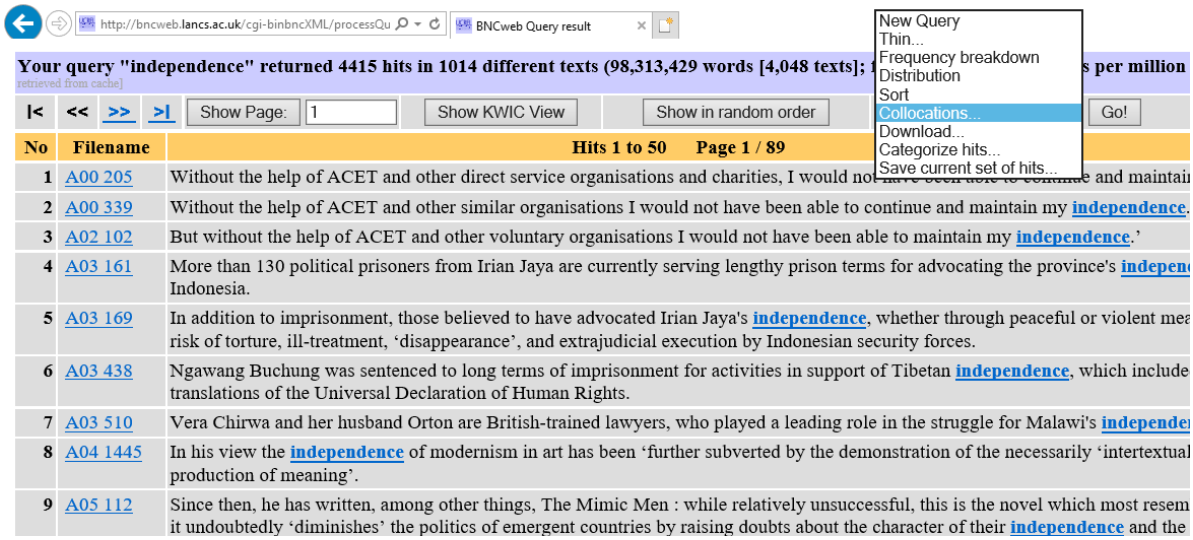
3.2.4. Collocations

With the help of various corpus analysis tools lexicographers can study different patterns of words and their co-occurrence, also known as collocations. *The Corpus Query Processor (CQP)* for the *BNC (http://bncweb.lancs.ac.uk)* is used to explain how such patterns are recognized. It is important to note that the *BNC* is quite old in comparison to more recent and larger corpora used by the compilers of the selected dictionaries, yet it was chosen because it is still one of the most popular corpora in the world. The lack of accessibility is the main

reason for not choosing the corpora of the selected dictionaries. In sub-chapter 3.2.4 only two of the ten entry words will be investigated more thoroughly (see Appendix 3).

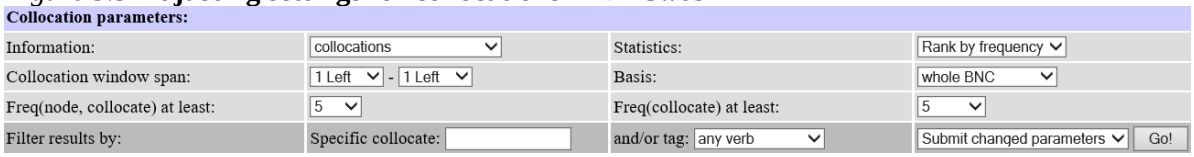
The first task of the lexicographer during the analysis process of any lexical item is to select the relevant criterion from the dropdown menu, which in this case is *Collocations.* Figure 3.2 demonstrates how such selection is performed.

Figure 3.2 **The selection of the necessary criterion in *BNCweb***



After that *the Corpus Query Processor* automatically creates a collocation database for the selected query word, which in this instance is *independence.* Certain parameters can be adjusted by the lexicographer to either expand or narrow the database content. Figure 3.3 demonstrates how such adjustments can be made.

Figure 3.3 **Adjusting settings for collocations in *BNCweb***



In the example of the query word *independence*, only few changes were made. First of all, collocation window span settings were changed so that only the nearest collocations from the left-hand side are displayed in the collocation list, apart from that also statistics settings were altered to make sure that all the retrieved collocations are ranked by frequency. Finally, the required word class was selected to single out only those collocations that function as verbs. Figure 3.4 shows the final result after all the described adjustments.

Figure 3.4 **The list of collocates for the query word *independence* after the adjustments**

No.	Word	Total No. in whole BNC	Expected collocate frequency	Observed collocate frequency	In No. of texts
1	achieved	7,821	0.304	45	15
2	gained	3,616	0.141	30	14
3	declared	4,210	0.164	14	11
4	achieve	6,710	0.261	9	8
5	want	54,762	2.128	9	7
6	seeking	4,557	0.177	7	7
7	achieving	1,820	0.071	6	4
8	declare	926	0.036	6	6
9	granted	4,541	0.176	6	6
10	promote	3,142	0.122	6	6
11	grant	1,258	0.049	6	6
12	following	11,380	0.442	5	5
13	gain	3,658	0.142	5	5
14	gaining	1,217	0.047	5	4
15	retain	2,522	0.098	5	5

The corpus query database automatically generates a list of all the collocates that match the selected criteria, additionally the database provides information about expected and observed collocate frequency. In this particular instance the corpus query system has generated a list of fifteen collocations arranged in the order of frequency, however in reality there are only nine, since some verbs in the list are displayed in different tense forms and therefore recur. Taking this into account the nine most frequent verb collocations for the query word **independence** are *achieve, gain, declare, want, seek, grant, promote, follow,* and *retain.* When such results are retrieved, the lexicographer's task is to create a dictionary entry. Table 3.9 is introduced to compare the obtained results with the collocations included in the selected dictionaries.

Table 3.9 **Verb collocations for the entry word *independence* in the selected dictionaries**

LDOCE	*OALD*	*MEDAL*
1. get independence 2. gain/achieve/win independence 3. declare independence 4. grant something independence 5. bring independence to something 6. move towards independence	1. have independence 2. lack independence 3. enjoy independence	1. achieve independence 2. declare independence 3. gain independence 4. lose independence 5. maintain independence 6. preserve independence 7. recognize independence 8. retain independence 9. win independence

The information about various verb collocations in each dictionary differs. *MEDAL* offers nine different verb collocations, *LDOCE* ranks second with eight collocations, and *OALD* provides only three. *LDOCE* and *MEDAL* are more superior than *OALD* also because their selection of collocations corresponds more precisely to the results obtained previously from *BNCweb.* There is a perfect match of four collocations in both *LDOCE* and *MEDAL,* whereas there is no match of any collocation in *OALD.* Apart from that the selection in *OALD* is quite random, yet the selection is quite the opposite in *LDOCE* and *MEDAL*, which

seemingly have arranged all the collocations in the order of frequency. One of the reasons why the correspondence between the database results from the *BNCweb* and the collocations represented in Table 3.9 is not completely identical is because different corpora were used by each dictionary. All three dictionaries tend to highlight the most frequent collocations within the corpus example sentences, especially *LDOCE* which provides the largest number of such examples.

The second entry word selected for the analysis is *policy.* The same adjustments were implemented, however unlike the previous example, in this case adjective collocations will be addressed in more detail. According to *BNCweb,* the fifteen most frequent adjective collocations for the query word **policy** are *foreign, social, economic, monetary, public, agricultural, regional, industrial, new, fiscal, national, environmental, general, American,* and *domestic.*

Table 3.10 **Adjective collocations for the entry word *policy* in the selected dictionaries**

LDOCE	*OALD*	*MEDAL*
1. government/public/state policy 2. party policy 3. company/hospital/university policy 4. foreign policy 5. economic/fiscal policy 6. defence/energy/housing policy 7. deliberate policy 8. clear policy 9. coherent policy	1. clear policy 2. coherent policy 3. explicit policy Additionally, two collocations were highlighted in bold in example sentences: 4. foreign policy 5. domestic policy	Highlighted in bold before example sentences: 1. foreign policy 2. housing policy

In this example, the results are slightly different. Previously *MEDAL* offered the largest number of collocations, however in this instance *MEDAL* provides only two collocations. *OALD* and *LDOCE* offer five and sixteen different collocations respectively. The highest number of matches between dictionary results and *BNCweb* results can be found in *LDOCE* (four matches), followed by *OALD* (two matches), and *MEDAL* (one match).

Although the application of collocations in the selected dictionaries is diverse, some conclusions can be drawn. The analysis of the entire set of entry words (see Appendix 3) revealed that in the vast majority of cases *LDOCE* provides the largest number of collocations, additionally the layout of collocations in *LDOCE* is more accurate in comparison to *OALD* and *MEDAL,* which introduce the collocations in a more scattered fashion. In seven out of the ten selected headwords, the highest number of matches between dictionary results and *BNCweb* results were discovered in *LDOCE.*

3.2.5. Corpus examples

As discussed in sub-chapter 2.4.4, corpus also contributes to the selection of illustrative examples, which is one of the main characteristic features of a monolingual learners' dictionary. Corpus provides access to authentic examples lexicographers can work with during the synthesis stage of a dictionary compilation process, therefore the purpose of sub-chapter 3.2.5. is to present how corpus examples are introduced in each dictionary. For the analysis of this criterion ten entry words were selected, and the three picked out for a more thorough analysis are *dangerous, exit,* and *parrot.*

The corpus examples in *LDOCE* are slightly modified versions of authentic sentences carefully selected from the Longman Corpus Network and therefore are more realistic than in other dictionaries (Longman, 2005: x). *MEDAL* (Online 8) introduces a similar approach and explains that the example selection process is done by means of Word Sketch and concordances which identify the most typical features and facts about the word. Just like *LDOCE,* also *MEDAL* often introduces shortened forms of sentences especially when they are too long. *OALD* is the only dictionary that has not given any information about corpus example application in its dictionary entries, but hopefully the results of the study will provide the necessary answers.

Table 3.11 **Corpus example sentences for the entry word *dangerous* in the selected dictionaries**

LDOCE	*OALD*	*MEDAL*
1. laws about dangerous dogs 2. Some of these prisoners are extremely dangerous. 3. It was a highly dangerous situation. 4. The crumbling sidewalks are dangerous for old people. 5. The virus is probably not dangerous to humans. 6. It's dangerous for a woman to walk alone at night. 7. The powdered milk was not as good as breast milk, and was **downright dangerous** (=actually dangerous) when it was mixed with unclean water.	1. a dangerous road/illness/sport 2. dangerous levels of carbon monoxide 3. The prisoners who escaped are violent and dangerous. 4. The situation is **highly dangerous**. (British English) 5. a conviction for dangerous driving 6. The traffic here is very dangerous for children. 7. dangerous for somebody to do something 8. It would be dangerous for you to stay here.	1. a dangerous dog 2. a dangerous stretch of road 3. Air pollution has reached dangerous levels in some cities. 4. an exciting but highly dangerous sport 5. Children are taught to avoid potentially dangerous situations. 6. It is not yet known whether these chemicals are dangerous to humans. 7. It's dangerous to walk around here on your own at night.

As shown in Table 3.11, the application of corpus example sentences in the selected dictionaries is quite similar. All dictionaries use a mixture of complete and incomplete

sentences. From the first moment it may seem that the shortened sentences are less useful since they do not help learners to decipher the meaning of the relevant headword, however having carefully observed all the sentences it seems that such examples have a different function. The aim of the incomplete sentences is to illustrate the most typical collocations. For instance, sentences 1, 2, and 4 in the case of *MEDAL* clearly indicate the most typical collocations *(a dangerous dog, a dangerous stretch of road, a dangerous sport),* the full sentences, on the other hand, provide more context and create a more authentic effect. *LDOCE* and *OALD* in comparison to *MEDAL* supply a list of extra corpus examples. *LDOCE* introduces additional corpus examples in a section *Examples from the Corpus,* whereas *OALD* provides supplementary list of examples in a section *Extra Examples.* The selection of sentences in all three dictionaries is not random and is done with intent to illustrate the most typical contextual features including syntax, register, and collocations.

Table 3.12 **Corpus example sentences for the entry word *exit* in the selected dictionaries**

LDOCE	*OALD*	*MEDAL*
1. We made for the nearest exit. 2. an exit door 3. Two men were blocking her exit.	1. Where's the exit? 2. There is a **fire exit** on each floor of the building. 3. The **emergency exit** is at the back of the bus.	1. I was wandering round Belgrade Airport looking for the exit. 2. Passengers should leave the plane by the nearest emergency exit.

For comparison purposes, Table 3.12 displays a list of corpus examples for the noun entry *exit.* As before, the example sentences are selected and modified based on the four criteria introduced in sub-chapter 2.4.4, which are naturalness, typicality, informativeness, and intelligibility. So far, it has been found that the two primary functions of example sentences are to indicate the most typical contextual features and to help learners to decipher the meaning of the relevant entry word, yet often there are sentences that do not perform any of the aforementioned functions. The first sentence in *OALD (Where's the exit?)* serves as an excellent example of such sentences, since it does not provide a comprehensible context and does not give any information about the most typical collocations, yet dictionary editors have considered it necessary to include such example in the dictionary entry because it is frequently used in everyday, real life conversations. Such observation suggests that corpus example selection is another domain where frequency information plays an important role.

The final headword selected for a more detailed analysis is *parrot.* It was found that none of the selected dictionaries have included a single corpus example sentence directly after the definition. *LDOCE* is the only dictionary that has provided a list of seven sentences in the section *Examples from the Corpus.* Such a trend was noted in relation to some other headwords as well (see Appendix 4). Inclusion and non-inclusion of corpus examples in

certain cases depends on various aspects such as frequency in corpus, insufficiency of collocability potential, and transparency of syntactic behaviour.

The main inference to be drawn is that all three dictionaries are almost identical in terms of sentence quality and structure, the only difference is that *LDOCE* and *OALD* in comparison to *MEDAL* supply a larger number of extra examples.

CONCLUSIONS

The goal of this Bachelor Thesis was to investigate application of corpus data in *Longman*, *Oxford*, and *Macmillan* English monolingual learners' dictionaries. The investigation provided answers to the following research questions:

1. How the application of corpus data is applied in each dictionary selected for the investigation?
2. Which MLD applies the corpus data most effectively?

The investigation of the theoretical material revealed that MLDs provide a large assortment of lexicographic practices which are heavily influenced by many corpus-based approaches. Over the years the rapid development of corpus lexicography has helped to perfect the dictionaries and make them better with every subsequent edition. The main conclusion to be drawn from the theoretical material is that corpus in combination with different analysis systems and tools provides a considerable number of corpus application opportunities, and therefore is an indispensable part of contemporary lexicography.

The findings showed that application of corpus data in the selected dictionaries is, for the most part, similar, but certain differences can be identified with regard to almost every corpus application criterion selected for the analysis. The investigation of word frequency representation is a good example of such inference because *LDOCE* and *MEDAL* use three-symbol rating approach to divide the most frequent words in smaller categories. *LDOCE* for example uses three red points to indicate whether the word is among the three thousand, two thousand, or one thousand most frequent words. *OALD* in comparison to *LDOCE* and *MEDAL* does not divide the most frequent words in finer categories. Additionally, it was noted that *LDOCE* is the only dictionary that distinguishes between the frequency of words in written and spoken language. Such observation leads to the conclusion that *LDOCE* provides the most comprehensive account of word frequency.

LDOCE is superior also because it offers the largest number of collocations and presents them more accurately than *OALD* and *MEDAL*, which exhibit different collocations in a more disorderly fashion. To illustrate how the most typical collocations are retrieved from the corpus *the Corpus Query Processor* for *the British National Corpus (BNCweb)* was employed. The results indicated that in seven out of the ten selected headwords, the largest number of matches between *BNCweb* most typical collocations and collocations identified in dictionary entries were discovered in *LDOCE*.

The study revealed that it is quite difficult to determine which dictionary illustrates different word forms most effectively, mainly because it really depends on what the learner is looking for. *LDOCE* certainly offers more complete information about various word forms

including nouns, adjectives, adverbs, and verbs, whereas *OALD* and *MEDAL* provide a more comprehensible information about different tense forms of verbs.

Some major differences have also been identified with respect to senses and sense ordering of the entry words. Knowing that senses in *LDOCE* are arranged according to their frequency in corpus, it was easy to draw conclusions about the two other dictionaries. The results of the research revealed that *MEDAL* and *LDOCE* follow the same ordering pattern, however in the case of *OALD* three ordering deviations were recognized. Such deviations appeared because in some cases *OALD* combines the frequency order with the semantic order method in which the core sense will always be presented as the first one, followed by other senses that are semantically closest.

The final criterion of analysis was the application of corpus examples. During the investigation it was concluded that all three dictionaries use a quite similar approach and include both complete and incomplete sentences. The complete sentences tend to provide more context, and therefore help learners to decipher the meaning of the relevant headword, additionally such sentences create an authentic effect. The incomplete sentences offer a more explicit illustration of the most typical collocations. The only reason why the application of corpus examples in the selected dictionaries slightly differs is because *LDOCE* and *OALD* in comparison to *MEDAL* supply a larger number of extra corpus examples.

The results of the study indicate that *LDOCE* applies the corpus data most effectively. *LDOCE* is an unquestionable leader in three of the five criteria selected for the analysis, whereas the two other dictionaries lag behind and do not stand out as a clear leader in any of the selected criteria, especially *OALD* which is seemingly the weakest dictionary in terms of word frequency representation, and therefore could present such information more thoroughly by dividing the 3000 most frequent words in smaller categories. This could be done using three-symbol rating approach as in the two other dictionaries. Although the application of corpus data in some dictionaries is better than in others, it does not mean that they are overall inferior.

Since the author of the thesis has selected only three MLDs for the analysis, it is recommended to conduct further studies regarding application of corpus data in other dictionaries such as *CALD, Cobuild,* and *MWALED.*

THESES

1. The main feature which distinguishes monolingual dictionary from any other dictionary type is its capability to provide detailed grammatical, stylistic, and semantic information.
2. The application of corpus data has made a very important contribution towards the creation of various English monolingual learners' dictionaries.
3. Corpus is defined as a collection of real language production examples which can help to perform various types of language analysis.
4. Apart from lexicography, corpora have also been introduced in other linguistic disciplines including grammatical studies of specific linguistic constructions, reference grammars, language variation, historical linguistics, contrastive analysis and translation theory, natural language processing, language acquisition, and language pedagogy.
5. Various corpus analysis systems and tools such as *the Corpus Query Processor* and *the Sketch Engine* significantly facilitate the task of dictionary editors and lexicographers during the processes of data analysis and synthesis.
6. *LDOCE* provides the most effective representation of corpus application in three of the five criteria selected for the analysis, whereas *MEDAL* and *OALD* do not surpass in any way.
7. *LDOCE* provides the most comprehensive representation of word frequency, since it is the only dictionary that distinguishes between the frequency of words in written and spoken language. *OALD* and *MEDAL*, on the other hand, do not provide such distinction. *OALD* is the least effective in terms of word frequency representation because it does not clearly indicate which words are among the three thousand, two thousand, and one thousand most frequent words.
8. *OALD* is the only dictionary that combines the frequency sense order with the semantic ordering pattern in which the core sense is always presented as the first one, followed by others that are semantically closest.
9. Lexicographers must be proficient and particularly attentive in selecting and modifying the corpus examples. Such sentences must provide not only a comprehensible context but also accurate information about the most typical collocations.
10. In some cases, such aspects as frequency in corpus, collocability potential, and transparency of syntactic behaviour determine the inclusion or non-inclusion of corpus examples in the dictionary entries.

11. Mostly the quality of corpus application in *LDOCE, OALD,* and *MEDAL* differs only slightly, however the only dictionary that is far behind the two other dictionaries in terms of representation of word frequency is *OALD.*

REFERENCES

1) Atkins, B. T. S. and Rundell, M. (2008) *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.
2) Atkins, B. T. S. (1985) *Monolingual and Bilingual Learners' Dictionaries: A Comparison.* New York: Collins Publishers Ltd.
3) Béjoint, H. (1981) The foreign student's use of monolingual English dictionaries: a study of language needs and reference skills. *Applied Linguistics,* 2 (3): 210.
4) Béjoint, H. (2000) *Modern Lexicography: An Introduction.* Oxford: Oxford University Press.
5) Bogaards, P. (2010) *The Evolution of Learners' Dictionaries and Merriam-Webster's Advanced Learner's English Dictionary*. In Kernerman, I. and Bogaards, P. (eds.), (2010) *English Learners' Dictionaries at the DSNA 2009.* Jerusalem.
6) Brumfit, C. J. (ed.), (1985) *Dictionaries, Lexicography and Language Learning.* Oxford: Pergamon Press Ltd.
7) Cowie, A. P. (1999) *English Dictionaries for Foreign Learners: A History.* Oxford: Oxford University Press.
8) Cowie, A. P. (ed.), (2009) *The Oxford History of English Lexicography.* Oxford: Oxford University Press
9) De Cock, S. and Granger, S. (2004) *Computer Learner Corpora and Monolingual Learners' Dictionaries: the Perfect Match*. In Teubert W. and Mahlberg M. (eds.), (2005) The Corpus Approach to Lexicography. *Lexicographica: International Annual for Lexicography*, 20: 72.
10) Durkin, P. (ed.), (2016) *The Oxford Handbook of Lexicography.* Oxford: Oxford University Press.
11) Fontenelle, T. (2009) *Linguistic Research and Learner's Dictionaries: the Longman Dictionary of Contemporary English*. In Cowie, A.P. (ed.), (2009) *The Oxford History of English Lexicography.* Oxford: Oxford University Press.
12) Humblé, P. (2001) *Dictionaries and Language Learners.* Frankfurt: Haag & Herchen.
13) Hunston, S. (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.
14) Jackson, H. (2013) *The Bloomsbury Companion to Lexicography.* London and New York: Bloomsbury Publishing Plc.
15) Jackson, H. (2002) *Lexicography: An Introduction.* London: Routledge.
16) Karpinska, L. (2015) *English-Latvian Lexicographic Tradition: A Critical Analysis.* Berlin and Boston: Walter de Gruyter GmbH.

17) Kennedy, G. D. (1998) *An Introduction of Corpus Linguistics.* London and New York: Routledge.
18) Lew, R. (2013) *Dictionaries and Technology.* Poznań: Adam Mickiewicz University.
19) Lew, R. (2004) *Which Dictionary for Whom?: Receptive Use of Bilingual, Monolingual and Semi-bilingual Dictionaries by Polish Learners of English.* Poznań: Motivex.
20) Lindquist, H. (2009) *Corpus Linguistics and the Description of English.* Edinburgh: Edinburgh University Press Ltd.
21) McCarthy, M. (2004) *Touchstone: From Corpus to Course Book.* Cambridge: Cambridge University Press
22) McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-based Language Studies: An Advanced Resource Book.* London and New York: Routledge.
23) McEnery, T. and Hardie, A. (2011) *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press.
24) Meyer, C. F. (2004) *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.
25) Morton, H. C. (1989) Gove's rationale for illustrative quotations in Webster's Third New International. *Dictionaries,* 11: 154-5.
26) Otlogetswe, T. J. (2011) *Text Variability Measures in Corpus Design for Setswana Lexicography.* Newcastle upon Tyne: Cambridge Scholars Publishing
27) Paltridge, B. (2012) *Discourse Analysis: An Introduction.* London and New York: Bloomsbury Publishing Plc.
28) Siddiek, A. G. (2013) *Monolingual & Bilingual Dictionaries as Effective Tools of the Management of English Language Education.* Finland: Academy Publisher.
29) Sinclair, J. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.
30) Sterkenburg, P. G. J. (2003) *A Practical Guide to Lexicography.* Amsterdam and Philadelphia: John Benjamins Publishing Company.
31) Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work.* Amsterdam and Philadelphia: John Benjamins Publishing Company.
32) Underhill, A. *Working with the Monolingual Learners' Dictionary.* Hastings: International House.
33) Zgusta, L. (1971) *Manual of Lexicography.* The Hague: Mouton.

Internet sources:

1) Available from https://bloch.umkc.edu/students/student-services/documents/Writing-A-Comparative-Analysis.pdf [Accessed on 5 May 2018]
2) Mairs, J. (2013) *Monolingual and bilingual dictionaries.* Available from http://www.learnersdictionary.com/qa/monolingual-and-bilingual-dictionaries [Accessed on 14 January 2018]
3) Kernerman, L. (1996) *English Leraners' Dictionaries: How Much Do We Know about their use?* Available from http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%201996%20Part%201/045_Lionel%20Kernerman%20-English%20Learners%20Dictionaries_%20How%20Much%20do%20we%20Know.pdf [Accessed on 5 May 2018]
4) Björkenstam, K. N. *What is a corpus and why are corpora important tools?* Stockholm: Stockholm University. Available from https://nordiskateckensprak.files.wordpress.com/2014/01/knb_whatisacorpus_cph-2013_outline.pdf [Accessed on 22 December 2017]
5) Lehmann, C. (2017) *Lemmas and lemmatization.* Available from https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/index.html?https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/lemmatization.html [Accessed on 22 April 2018]
6) Summers, D. (2005) *Corpus Lexicography – The importance of representativeness in relation to frequency.* Available from www.pearsonlongman.com/dictionaries/pdfs/corpus-lexicography.pdf [Accessed on 22 April 2018]
7) Available from https://corpus.byu.edu/bnc/
8) Available from http://www.macmillandictionaries.com/features/from-corpus-to-dictionary/
9) Available from https://www.sketchengine.eu/

Dictionaries:

1) *Longman Dictionary of Contemporary English*, sixth edition (2014) London: Pearson Longman.
2) *Longman Dictionary of Contemporary English*, fourth edition (2005) London: Pearson Longman.
3) *Oxford Advanced Learner's Dictionary,* ninth edition (2015) Oxford: Oxford University Press

Electronic versions of the selected dictionaries:

1) *Longman Dictionary of Contemporary English.* Available from https://www.ldoceonline.com/
2) *Oxford Advanced Learner's Dictionary.* Available from https://www.oxfordlearnersdictionaries.com/
3) *Macmillan English Dictionary for Advanced Learners.* Available from https://www.macmillandictionary.com/

Appendix 1

Representation of different word forms (lemmatization)

LDOCE	*OALD*	*MEDAL*
Entry word: computer Word family (noun) *computer, computing, computerization* (verb) *computerize* (singular) *computer* (plural) *computers*	**Entry word: computer** none	**Entry word: computer** (singular) *computer* (plural) *computers*
Entry word: school Word family (noun) *school, pre-school* (adjective) *pre-school* (verb) *school* (singular) *school* (plural) *schools*	**Entry word: school** (singular) *school* (plural) *schools*	**Entry word: school** (singular) *school* (plural) *schools*
Entry word: water Word family (noun) *water, waters* (adjective) *underwater, water, waterless* (verb) *water* (adverb) *underwater* (singular) *water* (plural) *waters*	**Entry word: water** (singular) *water* (plural) *waters*	**Entry word: water** (singular) *water* (plural) *waters*
Entry word: stretch Word family none	**Entry word: stretch** present tense I/you/we/they *stretch* he/she/it *stretches* past simple *stretched* past participle *stretched* -ing form *stretching*	**Entry word: stretch** present tense I/you/we/they *stretch* he/she/it *stretches* present participle *stretching* past tense *stretched* past participle *stretched*
Entry word: music Word family (noun) *music, musical, musician, musicianship, musicology, musicologist* (adjective) *musical, unmusical* (adverb) *musically*	**Entry word: music** none	**Entry word: music** none
Entry word: break Word family (noun) *break, outbreak, breakage* (adjective) *breakable ≠ unbreakable, broken ≠ unbroken*	**Entry word: break** present tense I/you/we/they *break* he/she/it *breaks* past simple *broke* past participle *broken* -ing form *breaking*	**Entry word: break** present tense I/you/we/they *break* he/she/it *breaks* present participle *breaking* past tense *broke* past participle *broken*

(verb) *break*		
Entry word: thunder (noun) Word family none	**Entry word: thunder (noun)** none	**Entry word: thunder (noun)** none
Entry word: intelligent Word family (noun) *intelligence, intelligentsia, intelligibility* (adjective) *intelligent ≠ unintelligent, intelligible ≠ unintelligible* (adverb) *intelligently, intelligibly*	**Entry word: intelligent** (opposite / antonym) *unintelligent*	**Entry word: intelligent** (adverb) *intelligently*

Appendix 2

Senses and sense ordering

Sense arrangement for the entry word *goal* in the selected dictionaries (ordering sequence: distinct)

LDOCE	*OALD*	*MEDAL*
1. something that you hope to achieve in the future 2. the area between two posts where the ball must go in order to score in games such as football or hockey 3. the action of making the ball go into a goal, or the score gained by doing this	1. (in sports) a frame with a net into which players must kick or hit the ball in order to score a point 2. the act of kicking or hitting the ball into the goal; a point that is scored for this 3. something that you hope to achieve	1. something that you hope to achieve 2. the net or structure that you try to get the ball into in games such as football and basketball a. the action of putting a ball into a goal b. the point or points that you score by putting a ball into a goal

Sense arrangement for the entry word *buck* in the selected dictionaries (ordering sequence: identical)

LDOCE	*OALD*	*MEDAL*
1. a US, Canadian, or Australian dollar 2. a male rabbit, deer, and some other male animals 3. a young man	1. a US, Australian or New Zealand dollar 2. a male deer, hare or rabbit 3. a young man	1. a dollar 2. the male of some animals such as rabbits or deer 3. a young man

Sense arrangement for the entry word *pupil* in the selected dictionaries (ordering sequence: identical)

LDOCE	*OALD*	*MEDAL*
1. someone who is being taught, especially a child 2. the small black round area in the middle of your eye	1. a person who is being taught, especially a child in a school 2. a person who is taught artistic, musical, etc. skills by an expert 3 the small round black area at the centre of the eye	1. someone, especially a child, who goes to school or who has lessons in a particular subject a. a barrister who is completing their training by working with an experienced barrister 2. the black round part in the middle of your eye

Sense arrangement for the entry word *blink* in the selected dictionaries (ordering sequence: identical)

LDOCE	*OALD*	*MEDAL*
1. to shut and open your eyes quickly 2. if lights blink, they shine unsteadily or go on and off quickly	1. when you **blink** or **blink your eyes** or **your eyes blink**, you shut and open your eyes quickly	1. to close your eyes for a very short time and quickly open them again 2. if a light blinks, it goes on and off continuously

	2. to shine with an unsteady light; to flash on and off	

Sense arrangement for the entry word *log* in the selected dictionaries (ordering sequence: identical)

LDOCE	*OALD*	*MEDAL*
1. a thick piece of wood from a tree 2. an official record of events, especially on a journey in a ship or plane 3. a logarithm	1. a thick piece of wood that is cut from or has fallen from a tree 2. (also **logbook**) an official record of events during a particular period of time, especially a journey on a ship or plane 3. logarithm	1. a thick piece of wood cut from a tree 2. a written record of things that happen, especially an official record of a journey on a ship or in a plane 3. a logarithm

Sense arrangement for the entry word *coke* in the selected dictionaries (ordering sequence: distinct)

LDOCE	*OALD*	*MEDAL*
1. the drink Coca-Cola, or a bottle, can, or glass of this drink 2. cocaine 3. a solid black substance produced from coal and burned to provide heat	1. cocaine 2. a black substance that is produced from coal and burnt to provide heat	1. a type of sweet brown fizzy drink (=with gas in it), or a glass of this drink 2. the drug cocaine 3. a solid black substance similar to coal that people burn to produce heat

Sense arrangement for the entry word *letter* in the selected dictionaries (ordering sequence: identical)

LDOCE	*OALD*	*MEDAL*
1. a written or printed message that is usually put in an envelope and sent by mail 2. a sign in writing or printing that represents a speech sound 3. a large cloth letter that you sew onto a jacket, given as a reward for playing in a school or college sports team	1. a message that is written down or printed on paper and usually put in an envelope and sent to somebody 2. a written or printed sign representing a sound used in speech 3. a sign in the shape of a letter that is sewn onto clothes to show that a person plays in a school or college sports team	1. a message that you write on a piece of paper and send to someone 2. a written symbol that is used to represent a sound used in speech

Appendix 3

Collocations

Adjective collocations for the entry word *breakfast* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. big breakfast 2. good/proper breakfast (=healthy) 3. hearty breakfast 4. small/light breakfast 5. English/full breakfast 6. cooked/fried breakfast 7. buffet breakfast 8. continental breakfast 9. quick/hasty/hurried breakfast 10. long/leisurely breakfast 11. early/late breakfast 12. working breakfast	1. big breakfast 2. full breakfast 3. good breakfast 4. hearty breakfast 5. light breakfast	1. cooked breakfast 2. English breakfast 3. continental breakfast 4. full breakfast

The fifteen most frequent adjective collocations for the query word **breakfast** are *English, some, big, continental, good, cooked, any, early, new, late, fried, own, hearty, full,* and *light.*

Matches in *LDOCE: 11*

Matches in *OALD: 5*

Matches in *MEDAL: 3*

Adjective collocations for the entry word *democracy* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. parliamentary democracy 2. multiparty democracy 3. Western democracy 4. new democracy 5. emerging/fledgling democracy	1. genuine democracy 2. real democracy 3. true democracy 4. parliamentary democracy 5. Western democracy	1. new/emerging/fledgling democracy 2. social democracy 3. industrial democracy

The fifteen most frequent adjective collocations for the query word **democracy** are *liberal, industrial, parliamentary, social, local, new, multiparty, representative, participatory, political, direct, greater, French, socialist,* and *true.*

Matches in *LDOCE: 3*

Matches in *OALD: 2*

Matches in *MEDAL: 3*

Verb collocations for the entry word *expense* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. incur an expense	1. go to expense 2. incur expense 3. involve expense	1. incur expense 2. cover expense 3. meet expense 4. go to expense 5. reimburse expense

The ten most frequent adjective collocations for the query word **expense** are *achieve, go to gain, increase, make, save, provide, grow, maintain,* and *incur.*

Matches in *LDOCE: 1*

Matches in *OALD: 2*

Matches in *MEDAL: 2*

Table 3.9 **Adjective collocations for the entry word *moment* in the selected dictionaries**

LDOCE	*OALD*	*MEDAL*
1. right/perfect moment	1. brief moment	1. critical moment
2. opportune moment	2. fleeting moment	2. crucial moment
3. historic moment	3. passing moment	3. exact moment
4. memorable moment	4. one moment	4. key moment
5. exact/precise moment	5. very moment	5. precise moment
6. very moment	6. right moment	6. right moment
7. present moment		7. very moment
8. important moment		
9. critical/crucial moment		
10. defining moment		
11. finest moment		
12. proudest moment		
13. worst moment		

The fifteen most frequent adjective collocations for the query word **moment** are *one, last, very, right, same, long, brief, particular, next, present, first, given, precise, possible,* and *crucial.*

Matches in *LDOCE: 5*

Matches in *OALD: 4*

Matches in *MEDAL: 4*

Verb collocations for the entry word *employment* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. give/offer employment	1. look for employment	1. find employment
2. provide employment	2. seek employment	
3. create employment	3. find employment	
4. find/get employment	4. affect employment	
5. obtain/secure employment	5. prevent employment	
6. look for/seek employment		

The fifteen most frequent verb collocations for the query word **employment** are *find, seek, provide, create, get, obtain, increase, offer, secure, affect, generate, boost, stimulate, take,* and *raise.*

Matches in *LDOCE: 8*

Matches in *OALD: 3*

Matches in *MEDAL: 1*

Noun collocations for the entry word *bank* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. bank account	1. bank loan	1. bank manager
2. bank balance	2. bank manager	2. bank loan
3. bank card	3. bank account	3. bank robbery
4. bank charges	4. bank balance	4. bank card
5. bank clerk	5. bank deposit	5. bank rate
6. bank loan		6. bank draft
7. bank note		7. bank account
8. bank statement		8. bank balance
9. bank manager		9. bank holiday
10. bank robber/robbery		10. bank statement

The fifteen most frequent noun collocations for the query word **bank** are *holiday, account, manager, loan, deposit, balance, estimate, lending, manager, clerk, credit, rate, note, official,* and *base.*

Matches in *LDOCE: 6*

Matches in *OALD: 5*

Matches in *MEDAL: 6*

Verb collocations for the entry word *fame* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. win/gain fame	1. enjoy fame	1. rise to/shoot to fame
2. achieve/find fame	2. achieve fame	
3. bring fame	3. come to fame	
4. rise to fame	4. win fame	
5. shoot to fame	5. rise to/shoot to fame	
6. seek fame		
7. enjoy fame		

The nine most frequent verb collocations for the query word **fame** are *find, achieve, gain, shoot to, bring, rise to, enjoy, win,* and *come to.*

Matches in *LDOCE: 8*

Matches in *OALD: 6*

Matches in *MEDAL: 2*

Adjective collocations for the entry word *emotion* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. strong/intense emotions	1. deep emotions	1. uncomfortable emotions
2. powerful emotions	2. extreme emotions	2. strong emotions
3. deep emotions	3. intense emotions	3. mixed emotions
4. painful emotions	4. mixed emotions	
5. overwhelming emotions		
6. positive emotions		
7. negative emotions		
8. mixed/conflicting emotions		
9. pent-up emotions		
10. great emotions		

11. real emotions 12. raw emotions 13. human emotions		

The fifteen most frequent adjective collocations for the query word **emotion** are *conflicting, strong, human, mixed, negative, different, powerful, deep, uncomfortable, pent-up, positive, complex, many, real,* and *suppressed.*

Matches in *LDOCE: 10*

Matches in *OALD: 2*

Matches in *MEDAL: 3*

Appendix 4

Corpus examples

Corpus example sentences for the entry word *doll* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. a small wooden doll	1. a rag doll	No examples

Corpus example sentences for the entry word *downtown* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. I have to go downtown later.	1. a downtown store	1. the streets of downtown Las Vegas 2. Let's go downtown.

Corpus example sentences for the entry word *triangle* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. a triangle of land	1. (British English) a right-angled triangle 2. (North American English) a right triangle 3. Cut the sandwiches into triangles.	1. a triangle of land 2. four small triangles of bread

Corpus example sentences for the entry word *official* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. an official investigation into the causes of the explosion 2. the official policy on education 3. official statistics about illegal drug use 4. You will have to get official permission first. 5. Finally the letter of appointment came, making it all official.	1. an official announcement/decision/statement 2. according to official statistics/figures 3. the official biography of the President 4. An official inquiry has been launched into the cause of the accident. 5. The country's official language is Spanish. 6. I intend to lodge an official complaint (= to complain to somebody in authority). 7. The news is not yet official.	1. There will be an official investigation into last week's accident. 2. You'll have to get official permission from the headteacher.

Corpus example sentences for the entry word *careless* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. It was careless of him to leave the door unlocked. 2. a careless mistake 3. careless driving 4. He's careless with his glasses and has lost three pairs. 5. Careless talk can be disastrous for a business.	1. It was careless of me to leave the door open. 2. Don't be so **careless about/with** spelling. 3. a careless worker/driver	1. Try not to be so careless next time! 2. It was very careless of you to leave the medicine where the children could get it.

Corpus example sentences for the entry word *sky* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. The sky grew dark, and a cold rain began to fall. 2. A shooting star sped across the night sky. 3. There wasn't a cloud in the sky.	1. What's that in the sky? 2. The sky suddenly went dark and it started to rain. 3. the night sky 4. a cloudless sky 5. cloudless skies 6. a land of blue skies and sunshine 7. The skies above London were ablaze with a spectacular firework display.	1. At noon the sun is directly above us in the sky. 2. Air pollution is clearly visible in the skies over the city. 3. a clear blue sky 4. a patch of blue sky among the clouds

Corpus example sentences for the entry word *fluent* in the selected dictionaries

LDOCE	*OALD*	*MEDAL*
1. She was fluent in English, French, and German.	1. She's fluent in Polish. 2. a **fluent speaker/reader** 3. 'Can he speak German?' 'Yes, he's fluent.'	1. I'm fluent in three languages. 2. Steve speaks fluent Japanese.

Dokumentārā lapa

Bakalaura darbs “Application of Corpus Data in Contemporary English Learners’ Dictionaries” (Korpusa datu izmantošana mūsdienu mācību vārdnīcās angļu valodā) izstrādāts LU Humanitāro Zinātņu fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Mārtiņš Jānis Možeiko 24.05.2018.

Rekomendēju/nerekomendēju darbu aizstāvēšanai

Vadītāja: docente Dr. philol. Laura Karpinska 24.05.2018.

Recenzents:

Studiju metodiķe: Samanta Matecka 24.05.2018.

Darbs iesniegts Anglistikas nodaļā 24.05.2018.

Darbu pieņēma:

Darbs aizstāvēts bakalaura gala pārbaudījuma komisijas sēdē

2018. gada..... jūnijā, prot. Nr. ....., vērtējums .....

Komisijas sekretāre: