Role of Language Corpora in Today's Linguistic Study of the English Language

Barbora Šudová

Bachelor Thesis
2009

Tomas Bata University in Zlín
Faculty of Humanities

Univerzita Tomáše Bati ve Zlíně
Fakulta humanitních studií
Ústav anglistiky a amerikanistiky
akademický rok: 2008/2009

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Barbora ŠUDOVÁ**
Studijní program:	**B 7310 Filologie**
Studijní obor:	**Anglický jazyk pro manažerskou praxi**
Téma práce:	**Role lingvistických korpusů v současném lingvistickém studiu anglického jazyka**

Zásady pro vypracování:

Teoretická část:
Definice lingvistických korpusů
Typy a struktura lingvistických korpusů
Britský národní korpus
Využití lingvistických korpusů
Praktická část:
Sestavení dotazníků
Analýza využitelnosti anglických korpusů ve výukové praxi
Zpracování dotazníkového šetření
Zhodnocení role anglických korpusů ve výuce anglického jazyka

Rozsah práce:

Rozsah příloh:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

McCarthy, Michael and O'Dell, Felicity. English Collocation in Use. Cambridge: Cambridge University Press, 2005.
McEnery, Tony and Wilson, Andrew. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996.
Meyer, Charles F. English Corpus Linguistics. Cambridge: Cambridge University Press, 2002.
Nat Bartels. Applied Linguistics and Language Teacher Education. Boston: Springer Science + Business Media, Inc. 2005
Aarts J., de Haan P. and Oostdijk N. English Language Corpora: Design, Analysis and Exploitation: Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992. Amsterdam & Atlanta: Rodopi, 1993.
Bowker, Lynne and Pearson, Jennifer. Working with Specialized Language: A Practical Guide to Using Corpora. London: Routledge, 2002.

Vedoucí bakalářské práce: **Mgr. Lenka Drábková, Ph.D.**
Ústav anglistiky a amerikanistiky

Datum zadání bakalářské práce: **30. listopadu 2008**

Termín odevzdání bakalářské práce: **15. května 2009**

Ve Zlíně dne 11. února 2009

L.S.

prof. PhDr. Vlastimil Švec, CSc.
děkan

doc. Ing. Anežka Lengálová, Ph.D.
vedoucí katedry

PROHLÁŠENÍ AUTORA BAKALÁŘSKÉ PRÁCE

Beru na vědomí, že

- odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby [1)];
- beru na vědomí, že bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí;
- na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3 [2)];
- podle § 60 [3)] odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- podle § 60 [3)] odst. 2 a 3 mohu užít své dílo – bakalářskou práci - nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tj. k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům.

Ve Zlíně ....

---

1) zákon č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, § 47b Zveřejňování závěrečných prací:

(1) Vysoká škola nevýdělečně zveřejňuje disertační, diplomové, bakalářské a rigorózní práce, u kterých proběhla obhajoba, včetně posudků oponentů a výsledku obhajoby prostřednictvím databáze kvalifikačních prací, kterou spravuje. Způsob zveřejnění stanoví vnitřní předpis vysoké školy.

(2) Disertační, diplomové, bakalářské a rigorózní práce odevzdané uchazečem k obhajobě musí být též nejméně pět pracovních dnů před konáním obhajoby zveřejněny k nahlížení veřejnosti v místě určeném vnitřním předpisem vysoké školy nebo není-li tak určeno, v místě pracoviště vysoké školy, kde se má konat obhajoba práce. Každý si může ze zveřejněné práce pořizovat na své náklady výpisy, opisy nebo rozmnoženiny.

(3) Platí, že odevzdáním práce autor souhlasí se zveřejněním své práce podle tohoto zákona, bez ohledu na výsledek obhajoby.

2) zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, § 35 odst. 3:
(3) Do práva autorského také nezasahuje škola nebo školské či vzdělávací zařízení, užije-li nikoli za účelem přímého nebo nepřímého hospodářského nebo obchodního prospěchu k výuce nebo k vlastní potřebě dílo vytvořené žákem nebo studentem ke splnění školních nebo studijních povinností vyplývajících z jeho právního vztahu ke škole nebo školskému či vzdělávacího zařízení (školní dílo).

3) zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, § 60 Školní dílo:
(1) Škola nebo školské či vzdělávací zařízení mají za obvyklých podmínek právo na uzavření licenční smlouvy o užití školního díla (§ 35 odst. 3). Odpírá-li autor takového díla udělit svolení bez vážného důvodu, mohou se tyto osoby domáhat nahrazení chybějícího projevu jeho vůle u soudu. Ustanovení § 35 odst. 3 zůstává nedotčeno.
(2) Není-li sjednáno jinak, může autor školního díla své dílo užít či poskytnout jinému licenci, není-li to v rozporu s oprávněnými zájmy školy nebo školského či vzdělávacího zařízení.
(3) Škola nebo školské či vzdělávací zařízení jsou oprávněny požadovat, aby jim autor školního díla z výdělku jím dosaženého v souvislosti s užitím díla či poskytnutím licence podle odstavce 2 přiměřeně přispěl na úhradu nákladů, které na vytvoření díla vynaložily, a to podle okolností až do jejich skutečné výše; přitom se přihlédne k výši výdělku dosaženého školou nebo školským či vzdělávacím zařízením z užití školního díla podle odstavce 1.

ABSTRAKT

Účelem této bakalářské práce je zhodnotit roli jazykových korpusů v lingvistickém studiu anglického jazyka, především českém školství.

Teoretická část začíná definicí korpusové lingvistiky, která se dá chápat jako úvod do do celé problemtaiky. Dále obsahuje definice jazykových korpusů z pohledu různých lingvistů, výčet známých a velkých korpusů. Poslední kapitola vysvětluje využití jazykových korpusů k různým účelům.

Praktická část je založena na dotaznících, které byly zaslány vysokoškolským a středoškolským pedagogům. Jejich odpovědi jsou zpracovány do jednotlivých kapitol a na závěr zhodnoceny.

Klíčová slova:
Jazykový korpus, korpusová lingvistika, konkordance, frekvence slov

ABSTRACT

The aim of this bachelor thesis is to evaluate the role of language corpora in the linguistic study of the English language, predominantly in the Czech school system.

The theoretical part starts with a definition of corpus linguistics which could be understood as an introduction to this area. Further, this part contains definitions of language corpora from the point of view from various linguists, and then a list of well-known and large language corpora is provided. The last part explains the use of language corpora for various purposes.

The practical part is based on the questionnaires which were answered by university and secondary school lecturers. Their answers are described in separate chapters and finally evaluated.

Keywords:
Language corpus, corpus linguistics, concordance, word frequency

ACKNOWLEDGEMENTS

I would like to thank my supervisor Mgr. Lenka Drábková, Ph.D., for valuable methodological conduction, all the consultations and important suggestions. I would also like to thank my family for their support during my studies.

DECLARATION OF ORIGINALITY

I hereby declare that the work presented in this thesis is my own and certify that any secondary material used has been acknowledged in the text and listed in the bibliography.

May 15, 2009

……

CONTENTS

INTRODUCTION **10**
1 CORPUS LINGUISTICS **11**
1.1 Approaches to Corpus Linguistics 12
2 LANGUAGE CORPORA **13**
2.1 The British National Corpus 15
2.1.1 The Design of the British National Corpus 15
2.1.2 Written and Spoken Sections 16
2.2 The Brown Corpus 17
2.3 The Cobuild Project 18
2.4 The Bank Of English 19
3 THE USE OF LANGUAGE CORPORA **20**
3.1 Lexical Studies 20
3.1.1 Dictionaries 20
3.1.2 Collocations 21
3.2 Corpora in Language Teaching 22
3.3 Corpora and Grammar 23
4 ANALYSIS OF THE QUESTIONNAIRES' RESULTS **24**
4.1 Analysis of the Answers from the Closed Questions Part 25
4.1.1 Question No. 1 25
4.1.2 Question No. 2 26
4.1.3 Question No. 3 27
4.1.4 Question No. 4 28
4.2 Analysis of the Answers from the Open Questions Part 30
4.2.1 Question No. 5 30
4.2.2 Question No. 6 32
4.2.3 Question No. 7 34
4.2.4 Question No. 8 36
4.2.5 Question No. 9 37
4.2.6 Question No. 10 39
CONCLUSION **41**
BIBLIOGRAPHY **42**
APPENDICES **43**

INTRODUCTION

Language corpora are one of the most interesting and valuable means in the study of language. Although corpus linguistics is quite a new branch, it is still developing and significant in linguistics studies.

There are many reasons why I am interested in this topic. The first reason is that I have always been involved in linguistics and its branches. The second one is that language corpora are not much known among students and I wanted to cast light on this phenomenon. Another reason is that corpus linguistics seems to be a very attractive discipline to me since it offers many possibilities to study language from new points of view, with new methods and by new means.

The theoretical part concerns with the main facts about corpus linguistics and language corpora. I described large language corpora, because I considered them as the best known ones among linguists and students. At the end, I concentrated on the use of language corpora. Since the number of possible use of language corpora is countless, I depicted the main ones.

In the practical part I aimed to outline the real situation of the use of language corpora in order to evaluate their role. Hence, questionnaires were compiled and sent to university and secondary school lecturers with questions concerning their own experience with language corpora. This survey offers many interesting results and ideas which are described in the practical part.

1 CORPUS LINGUISTICS

Elena Tognini-Bonelli (Tognini-Bonelli, 2001) describes corpus linguistics as a methodology which examines the use of language and does linguistic analysis on the basis of language corpora. Corpus linguistics is a relatively new discipline. It appeared in the 1960s almost simultaneously as Noam Chomsky, who had dismissed the corpus as a dependable source for his research, introduced his new approach to language studies. Chomsky saw corpora as an irrelevant source of data for linguistic research, mainly because linguists could not rely on corpora's poor informative value. He claimed that any corpus would be just a mere list of imprecise data and that the system would be distorted because its random word selection since some sentences could be all omitted or false.

M. A. K. Halliday (Halliday et al, 2004) writes about Chomsky's widely disputed text, *Syntactic Structures* which emerged in 1957 and his other work *Aspects of the Theory of Syntax* published in 1965, discussions about the standard views on theoretical linguistics were initiated. The language started to be examined thoroughly when linguists became to be discontented about the language theories of that time because the information seemed insufficient and new data had to be discovered. Due to that language corpora started to be compiled. With these large bodies of various text types of written or spoken language it became possible to make lists of the most frequent word phrases. New grammatical rules could be traced, the former ones could be modified and improved.

Graeme Kennedy (Kennedy, 1998) describes the situation in the 1970s. It was rather time-consuming to find a concordance consisting of a frequently used word, (e.g. *when* or *that)* since the performance of computers was not so high. The mainframe computing was a part of corpus linguistics until the mid-1980s. The 1990s meant a rapid progress for corpus linguistics because language corpora could run on personal high-powered computers as well.

Sampson (Sampson et al, 2004) sees modern corpus linguistics as ‘electronic corpus linguistics’ since this methodology depends heavily on advanced computer technology. Research could be done thanks to electronic processing which makes it easier to work with corpora. As a consequence, it is a computer science from the modern viewpoint.

Tserdanelis (Tserdanelis et al, 2004) points out that since the language corpora are collections of linguistic materials intended for specific purposes, corpus linguistics also deals with annotating and designing of the materials.

Sampson (Sampson et al, 2004) comes up with the idea that language can be analyzed thanks to corpus linguistics on the basis of discourse. It is impossible to access all existing texts because of the immense complexity of a discourse but the analysis is facilitated by corpus linguistics, concretely by language corpora. The samples clarify meanings of words but understanding is rather personal because everyone explains words differently on the basis of the concrete experience with a word. Corpus linguists inclined to the idea that a word is not innate in a language, therefore the position of a word is not considered to be crucial. The meaning of words has been constantly changing over the time, i.e. the meaning is transitory.

Generally, it is assumed that corpus linguistics helps to find out about the use of words and understanding their meanings when collecting a discourse. Due to the fact that corpora are always finite and can provide only a limited view on the language corpus linguistics is sometimes criticized. Nonetheless, it is acknowledged that the most frequent and important phenomenon will appear in corpora. (Semino et al, 2004)

1.1 Approaches to Corpus Linguistics

Graeme Kennedy (Kennedy, 1998) distinguishes scholars into three groups. The first group of scholars shapes a corpus and as its compilers they are mainly interested in its structure and they prepare data to be stored. The second group concerns with tools for the corpora analysis, i.e. software tools. Both the groups of linguists contribute to the further corpora development. The third group consists of descriptive linguists whose aim is to describe the lexicon and grammar of corpora – not directly what it is but more likely how often the specific forms are used, i.e. frequency of words. This is the largest group. The fourth group which deals with corpus linguistics is the newest one. There are linguists who use corpus-based linguistics for language teaching and learning.

2 LANGUAGE CORPORA

The language corpora are connected with computers very often but there is a significant pre-electronic corpora tradition. There had been some corpora before the publication of the first officially renowned corpora or certain types of collections of words and phrases slightly similar to the nowadays corpora. The authors of such texts did not intend to compile corpora in a present-day sense. They did not even have to know they were creating a sort of a new system of recording language. Older dictionaries or collections of texts published earlier were not electronically stored but handwritten.

Charles F. Meyer (Meyer, 2002) describes the word ‘corpus’. It is of Latin origin and it was adopted as ‘body’ by English. Used in connection with linguistic terminology it is an integral and complex unit with interrelated parts. He also presents other linguists' theories in order to clarify the ideas about language corpora. In his opinion, linguists prefer restricted definitions of ‘corpus’ rather than calling it just a collection of many samples (any text type, newspapers, novels, poetry, drama, spoken language, etc.). However, Meyer also inclines to the use of the phrase ‘a collection of texts or parts of texts’.

According to Graeme Kennedy (Kennedy, 1998), language corpora are large collections of texts which can be used for further linguistic analysis. He also emphasizes the fact they are recorded in an electronic form which helps to compare them with text archives to realize what corpora are. The difference consists in systematization and constitutions of these systems. While the language corpora are compiled methodically, text archives are collected randomly. The language corpora, unlike text archives, represent a language, identify the elements and patterns of a particular language and map the rules of its usage.

Tony McEnery and Andrew Wilson (McEnery and Wilson, 1996) point out four main characteristics: sampling and representativeness, finite size, machine-readable form and a standard reference.

Sampling and representativeness are acquired when an abstract of a concrete language is made. To shape a representative corpus it is advised to draw from more sources. It is impracticable to analyze all possible texts and utterances that the language offers thus when researching corpora it is functional to choose the most characteristic models of a linguistic phenomenon.

Finite size is a typical feature of the language corpora but it is not a rule. Nevertheless, most of the language corpora were previously planned to be finite so the assumption that entries in corpora are of a finite size is confirmed.

Corpora are compiled using a computational method, although some corpora are still available also in printed forms, e.g. *A Corpus of English Conversation* (Svartvik and Quirk 1980) or *The Survey of English Usage* (Quirk 1968). Machine-readable corpora have several advantages over the printed ones, such as speed of searching, way of editing entries or the method of recording the spoken texts.

Finally, corpora can serve as a standard reference for researchers since they represent the language variety.

According to Michael Stubbs (Stubbs, 2002) corpora are designed by linguists who gather records of performance from various sources, a sample of the language use of many speakers. Data are collected from extensive texts, then rules about a word's meaning and use are derived. Stubbs mentions a term '*meaning in use*' which is related to the definitive understanding of words and phrases. The point is words could be understood differently in a variety of social and linguistic contexts. If this approach together with corpus semantics is interconnected, words can acquire entirely new meanings or the previous ones can be enlarged from their frequent co-occurrence with other words.

Elena Tognini-Bonelli (Tognini-Bonelli, 2001) gives another view on the language corpora. She claims that the language corpora are assumed to be a representative part of each language hence they are suitable for a scientific research. At the same time the samples co-occur naturally in the concrete language and they are produced spontaneously. Generally, it is assumed that corpora depict an authentic language.

Sampson (Sampson et al, 2004) expresses another idea. He supposes that the explanatory dictionaries may be viewed as some kind of small corpora as well. There is a word and a few meanings taken from miscellaneous contexts. After studying such an entry the use of the word is clearer. Dictionaries show the most typical examples and cases of the word's use and thus we can imagine the basic idea how a word works but it may not be always a clear perception of the meaning. Corpora are more elaborated than dictionaries because there are up to thousands entries of one concrete word in many contexts.

These were only some of many definitions. Each linguist interprets the meaning of the language corpora in a different way but they all have main features in common. To conclude, there is a definition from Graeme Kennedy (Kennedy, 1998) which says that language corpora are wide lists of millions of words extracted from various sources – from

literary pieces, poetry as well as prose, newspapers, educational literature, plays or from spoken language. It is evident that sources are practically unlimited since the language is still developing and flourishing.

2.1 The British National Corpus

The British National Corpus (BNC) is a body of written and spoken texts. It is monolingual and it is not dedicated to any other language or dialects than to British English. The British National Corpus does not deal with historic development of a language, only with the second half of the twentieth century, i.e. it is synchronic.

Graeme Kennedy (Kennedy, 1998) says that this intention to compile a huge collection of spoken and written British English originated in the 1990s. The BNC was published in 1994 and as a unique and well-organized system it has become a new standard in corpus design and compilation.

Further he refers to the fact that the corpus was aimed to be the representation of the British English. Several organizations came together to work on this challenging project - Oxford University Press, Longman Group (UK) Ltd, W. & R. Chambers, the British Library and the Universities of Oxford and Lancaster and this project was partly sponsored by the British government.

This corpus reached a size of more than 100 million words of modern British English and it is still growing because the language itself is still developing and modern computers are capable of storing a higher volume of data.

2.1.1 The Design of the British National Corpus

As Graeme Kennedy (Kennedy, 1998) mentioned before, the British National Corpus is a collection of spoken and written texts with a fixed structure and it is considered as one of the most influential and crucial repository of knowledge about nowadays British English.

He also describes the rate of the written texts to the spoken ones is unbalanced. Only about 10% of the 100 million words are from spoken sources. The reason is that it is easier to store the written texts in database unlike the spoken texts. The texts are kept electronically; hence they could be searched and edited promptly. It is a necessary condition for a transparent and well-arranged system.

2.1.2 Written and Spoken Sections

Following the content of the texts in the BNC, Graeme Kennedy (Kennedy, 1998) explains that the written sources consist of about 75% of ´informative´ prose, published after the year 1975, and about 25% of ´imaginative´ prose (literary works), published after the year 1960. It is taken from various kinds of prints (books, periodicals, published plays or printed speeches).

He also claims that the selection of the sources for these corpora is very challenging because it is impossible to draw from all renowned works and cover all extant sources. For that reason about a half of the texts is selected randomly. The language used in the corpus is not just formal language, e.g. academic or technical texts, but it is also informal language (e.g. slang, dialects), used by specific groups of people, including also very 'low' style. Archaic expressions or words from dead languages which are still being used are recorded as well.

Graeme Kennedy points out that the spoken part consists of 10 million words. The first type of recordings is collected from the sources such as educational lectures, tutorials, news reports, consultations or interviews. The second type of the sources consists of thousands of hours of recordings made by more than one hundred volunteers from various environments (socio-economic groups, males and females, aged between 15 and 60), all from the United Kingdom. Those people systematically recorded all their conversations. All the recordings were thoroughly monitored, even with pauses or repetitions, etc. No phonetic features were added and there was no phonetic analysis in the BNC.

The significant difference between the spoken and written texts is also in the use of the same words in various contexts or in a different understanding according to their occurrence either in spoken or written language.

2.2 The Brown Corpus

Charles F. Meyer (Meyer, 2002) outlines the origin of the first electronic corpus which has about one million running words - the Brown Corpus of American written English, which was compiled by native Americans W. Nelson Francis and Henry Kučera at Brown University in Rhode Island and published in 1964 in the United States. H. Kučera and W. Nelson Francis were two of the first compilers of language corpora in the 1960s. Working on language corpora was not much supported since there was a little tolerance for anything but generative grammar and corpus linguistics was in opposition. W.N. Francis and H. Kučera's effort to create this corpus are valued today but it was a daring attempt at that time. Moreover, some linguists characterized the compilation of the Brown Corpus as "a useless and foolhardy enterprise".

Charles F. Meyer further describes that the Brown Corpus was prepared as a part of the programme known as 'Project English'. The authors chose books, periodicals or anonymous materials from the year 1961 and further. They emphasized the fact that the corpus aimed to introduce a representative American English but since some authors of the used sources were unlisted, this is not acknowledged. Its sources, the overall size, the structure or the number of categories were agreed in advance since the corpus creation had been planned in a systematic way. The origin of the Brown Corpus gives an illustrative example about the situation between the corpus linguists and the generative grammarians, as well as the development of corpus linguistics. Today it is considered as the most balanced corpus and a type of benchmark in corpus linguistics.

W. Teubert (Halliday et al, 2002) describes the Brown Corpus as a data-oriented project which was unique in its easy comprehensive use. Nonetheless, the number of entries was not sufficient enough for research concerning both grammar and lexicon because one million words is just a tiny fraction of the whole discourse which is needed for further analysis. Even though the Brown Corpus served as a popular source for linguistic studies in Europe, it started to be undervalued in America after some time.

The corpus is available for further studying in Contact International Computer Archive of Modern English (ICAME) or in Norwegian Computing Centre for the Humanities in Bergen, Norway.

2.3 The Cobuild Project

The corpus-based lexicographical Cobuild project began in the year 1980. The name 'COBUILD' comes from cooperation between the publisher Collins and a research team from the University of Birmingham (*Collins Birmingham University International Language Database*). The COBUILD Project is known also as the *Cobuild Corpus* or as the *Birmingham Corpus*.

Graeme Kennedy (Kennedy, 1998) names John Sinclair as the leading force in Cobuild Corpus project. The compilers of this corpus used only real data. The aim of the Cobuild Corpus was to represent the English language as a teaching material for teachers, learners and researchers. The corpus, which consists of about 25% of spoken texts, represents rather general language than technical discourse. The authors focused on current usage of standard dialects. Both written and spoken texts were collected from people aged from 16 years and over, mainly British (70%), then American (20%) and other nationalities.

Charles F. Meyer (Meyer, 2002) suggests that this corpus is valuable especially for lexicographers for their work with word's meanings or collocations.

The COBUILD Project resulted in creation of the *Collins Cobuild English Language Dictionary*, as Halliday claims (Halliday et al, 2004). This was the first dictionary based exclusively on a language corpus. Nevertheless, language corpora do not contain all common or less common words that use to appear in dictionaries. There are predominantly words which are used by members of concrete discourse groups. In that case some infrequent words are not listed in the dictionary, e.g. *apo(ph)thegm*.

The Collins COBUILD project drew on the Bank of English fulfilling purpose of shaping descriptive grammar or compiling concordance for use in schools.
(Ghadessy et al, 2001)

2.4 The Bank Of English

The Bank of English was created by COBUILD at the University of Birmingham in 1991. The corpus expanded to 524 million words and the texts are added constantly. All the data in The Bank of English are stored electronically and the corpus is a collection of modern English language.

The Bank of English also draws from written (newspapers or books) or spoken (recorded speech from television or radio) sources, as Charles F. Meyer (Meyer, 2002) points out. It is an everyday discourse used by ordinary people.

Since there is a similarity between COBUILD and the Bank of English, the purpose of use which Charles F. Meyer refers to is the very same – the creation of dictionaries. All the sections of this corpus were designed to be the primary source for the *BBC English Dictionary*. But that was not the only dictionary which drew from this corpus. The Bank of English served also as the basis of the Collins COBUILD English Dictionary.

Baker (Baker et al, 2006) describes another feature of this corpus. The Collins Cobuild Bank of English is also called “dynamic corpus”, which is useful when monitoring language.

The corpus is accessible not only in full-version but also in forms of smaller sub-corpora. There is a sub-corpus available on the Internet or as a data-base stored on CD-ROM (Cobuild 1995b), which consists of 200 million words. Thousands of headwords collections or random samples, all in context, are available in this data-base.

3 THE USE OF LANGUAGE CORPORA

It is essential to make the right choice of what specific kind of language corpus should be used when studying language with language corpora. Each type of corpus is predetermined for different purposes. The main areas of the use of language corpora which will be outlined are these: lexical studies, teaching of language and grammar.

3.1 Lexical Studies

Graeme Kennedy (Kennedy, 1998) notices that language corpora have been crucial for lexicographical research since the 1960s when lexicography started to be heavily dependent on corpora. Before the origin of electronic corpora lexicographers had had to compile lexicon without statistical information. Among the main uses of language corpora belongs compiling of dictionaries, defining collocations or idioms.

Tony McEnery (McEnery and Wilson, 1996) says that the way which changed lexical studies lies in facilitation of the lexicographer's work. Millions of entries are sorted, edited and applied very quickly and they have a logical implication. As a consequence of that the information is more complex, precise and up-to-date.

3.1.1 Dictionaries

Tserdanelis (Tserdanelis et al, 2004) sees the use of dictionaries as widespread and he thinks that many people rely only on the interpretation which is written in a dictionary. Dictionaries represent just a tiny fraction of discourse while corpora represent millions of words of a particular language.

Charles F. Meyer (Meyer, 2002) refers to small-scaled corpora. To obtain as much information about words as possible when compiling a dictionary it is essential to explore such corpora which are not small and specialized but rather large and general. This precondition of an effective research is shown on the frequency of words. To get the most or the least frequent words in vocabulary a quantity of samples is necessary. A small corpus will not provide lexicographers with complete information which would lead to relevant results.

Baker (Baker et al, 2001) explains another way of compiling a dictionary. A corpus product called lexicon, a list of words which are kept in electronic form, can be used.

Graeme Kennedy (Kennedy, 1998) talks about electronic dictionaries that are also published today and which serve for further researches thanks to their machine-readable

form. The electronic forms of today's modern dictionaries are technically advanced programmes. Due to that all entries are easily located when seeking for definitions and easily searchable due to specific morphological features, word classes or borrowed words. Among the best known dictionaries belong the *Oxford English Dictionary* (OED) or the *Longman Dictionary of Contemporary English* (LDOCE).

Charles F. Meyer (Meyer, 2002) further notes that many dictionaries are still being produced on the basis of the *Collins COBUILD Project* because this corpus in not static and it admits new words constantly. Sections of the *Bank of English* are also used as a primary source for dictionaries of various types (e.g. *BBC English Dictionary* or the *Collins COBUILD Dictionary*). The British National Corpus also served as the basis of the *Longman Dictionary of Contemporary English*.

3.1.2 Collocations

Collocations can be easily found when using language corpora, as Baker mentions (Baker et al, 2006). By the methods of corpus linguistics researches easily define the frequency of collocations using statistics. It is useful to realize which collocations are yet low-frequent and which are used more often in a discourse.

Michael Stubbs (Stubbs, 2002) points out that a corpus seeks a collocate and the more times it occurs the more it is probable it would depict a lexical relation between two or more words appearing in a running text.

Another fact which is mentioned by Michael Stubbs is that researches show that words predominantly appear in routine phrases and words tend to be a part of collocations, i.e. they will not function independently but only in the conventional ways of a discourse.

A different tool that Michael Stubbs describes is KWIC (Key Word in Context) or word list generation - the basic medium for lexical studies and corpus analysis. This function allows researchers to find and display required phrases very quickly among corpus data and to allocate the most frequent collocations.

The reason for using language corpora when deciding which collocations should be put in dictionaries is clear – the choice of collocation would be random or dependant only on lexicographers. In that case the collocations would not be representative.

3.2 Corpora in Language Teaching

According to Graeme Kennedy (Kennedy, 1998) language teaching theories and methods have been developing through the last decades. Nowadays the change lies in the way of teaching. To learn a language in an efficient way is considered to be the goal of both teachers and students and the most efficient way is to focus on each learner's needs. The emphasis is put on communication, not particularly on analyzing traditional theories, systematic learning of vocabulary and grammar. When this was clarified language corpora started to be used in pedagogy more. They could provide learners with information about language means which are used the most and least frequently by native speakers.

Graeme Kennedy further mentions that information from language corpora means radical changes for pedagogy. Primarily, corpora studies transform the system of teaching completely when organizing what to be learned, how to teach or to define priorities of a subject matter. Secondly, teachers can focus only on the most useful things, e.g. frequently appearing phenomenon.

Another important fact which Graeme Kennedy points out is that in spite of the wide range of possible uses of language corpora it is necessary to work with corpora judiciously since they do not fully represent standard language. Still it is recommended to use corpora only as one of many possible sources of searching information when teaching a language.

Elena Tognini-Bonelli (Tognini-Bonelli, 2001) sees the significance of language corpora in teaching of a language in its double role – corpora bring new methodology to language teaching and also the theoretical level is innovated. These two aspects mean a new possibility how to teach a language. From a theoretical point of view, new facts are revealed using corpora. And from the methodological point of view, new ways of teaching are revealed, e.g. a teacher seeks the information in corpora or students themselves work with corpora.

Elena Tognini-Bonelli explains that it used to be predetermined what was to be taught but language corpora can help significantly to improve the way how to teach a language. For example, a student's own research can be innovated when using corpora. Students themselves identify words in another contexts, they examine the existing rules on concrete examples or they learn how to avoid common mistakes.

Mohsen Ghadessy (Ghadessy et al, 2001) suggests that even though larger corpora are more reliable when seeking for information, small specialized corpora are more appropriate for students. Small-scale analysis is easier and more proper for their needs.

There are many easily accessible tools for learners using corpora, e.g. frequency lists, concordance programmes or analysis of collocations.

Tony McEnery and Diana McCarthy (McEnery, Wilson, 1996) describe that since teaching of languages can be divided into two approaches (empirical and rationalist), textbooks are divided into two groups as well. The first group of texts depends heavily on established rules and examples, i.e. it is a rather theoretical approach. The books draw on already published materials and due to them the books can be innovated. The second group is consists of books which are based on empirical approach and it does not rely on theory much. Such books are created e.g. by the Collins COBUILD Project. Content of these textbooks is formed on the basis of language corpora. Empirical data collected thanks to corpora are essential in learning a foreign language because students acquire knowledge from real communicative situations. It was even proved that students who learned from the traditional textbooks relying on well-known theories did not comprehend more complex statements which often occured in corpora.

McEnery further points out that language corpora also serve as critique on ESL textbooks (English as a second language). Many scholars (e.g. Kennedy, Holmes and Mindt) criticized these textbooks using language corpora to prove the textbooks' deficiencies. They agreed that ESL textbooks should rely on authentic examples more. Another thing that they recommended is that the authors of the textbooks should concentrate on the ways of expressing specific language means more (collocations, idioms) or they should innovate vocabulary to be more up-to-date. The conclusion they made is that the non-empirical textbooks could be misleading and language corpora should definitely be used for textbook compilation.

3.3 Corpora and Grammar

Research concerning with grammar also take advantage of language corpora. Firstly, corpora are important for grammatical studies because of the number of grammatical phenomena they represent. Secondly, empirical data in corpora are also valuable for grammar theories. Most quantitative analyses which are carried out depend on corpora as a basis for their results. (McEnery, Wilson, 1996)

4 ANALYSIS OF THE QUESTIONNAIRES' RESULTS

I prepared a questionnaire for university and second school lecturers about language corpora and corpus linguistics, their real use in practice. The research is based on questions concerning the teachers' own experience. The questionnaire is divided into two parts. The first part consists of closed questions; therefore the statistical information can be derived from answers. The second part consists of open questions referring to the use of language corpora. The intention was that the open questions should be answered by those who have experience with corpora in connection with their work. There are two charts for questions 1-6; the following questions have too varied answers to be demonstrated in charts. The questionnaire was compiled in two language versions – Czech and English versions, which are enclosed as appendices.

The questionnaires were distributed to 209 (177 Czech and 32 native speakers) university lecturers and to 212 (200 Czech and 12 native speakers) secondary school teachers. It was intended to address as many respondents as possible to get a sufficient amount of answers for concrete results. University lecturers who answered the questionnaires work on philosophical faculties and faculties of education. I chose grammar schools as the most appropriate type of secondary schools.

The Chart No. 1 shows the overall number of respondents and the ratio between the university and secondary school teachers. From all the addressed respondents 33 university lecturers and 10 secondary school teachers sent back filled questionnaires. On the whole, 25 women and 18 men answered. Except the ten respondents another seven teachers replied that they did not work with corpora and thus they would not answer the questionnaire at all. The Chart No. 2 shows the ratio between the university and secondary school lecturers. There were 9 native speakers among the 43 respondents. Seven native speakers were from universities, only two were from secondary schools.

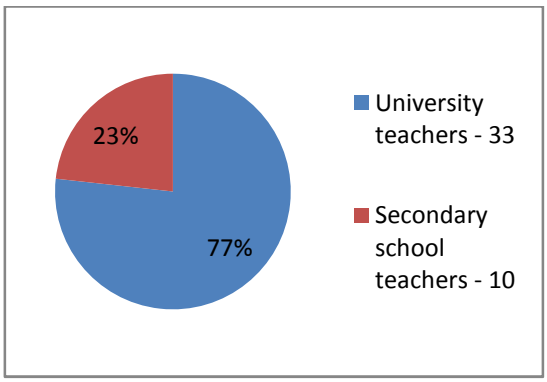

Chart No. 1 - Overall number of respondents

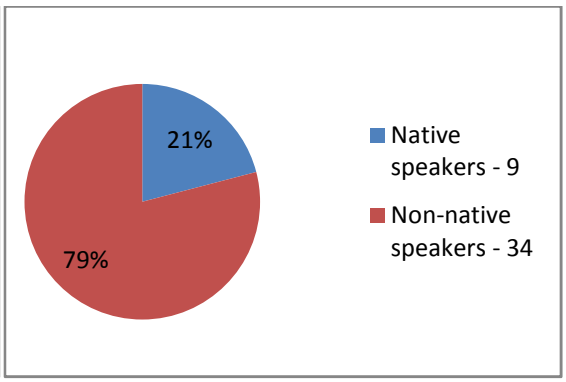

Chart No. 2 - Ratio between native and non-native speakers

4.1 Analysis of the Answers from the Closed Questions Part

The first four questions were answered by all the respondents. Lecturers who did not work with language corpora or those who were not interested in corpus linguistics and corpus studies did not answer the open questions in the second part of the questionnaire. On the other hand, those who were engaged in corpus linguistics described their opinions very thoroughly, although some of them did not have access to any corpus software.

4.1.1 Question No. 1

The first question was “Have you ever come across the term "language corpora" in connection with your work?”

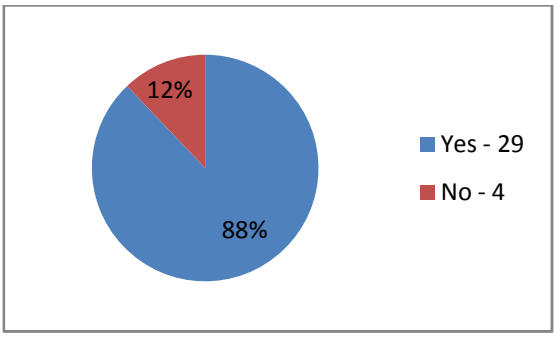

Chart No. 3 – University lecturers

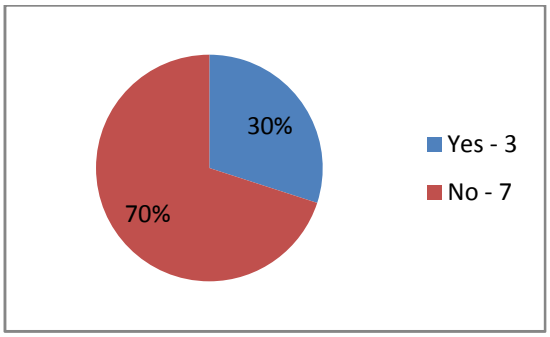

Chart No. 4 – Secondary school teachers

The Chart No. 3 shows that 29 (88%) out of 33 university lecturers have come across the term "language corpora" and only four have not. Two of the 33 university teachers did not teach linguistics but literature. Even these two teachers knew what language corpora were about as well as what their purpose was, although it was not in their specialization.

The Chart No. 4 shows that secondary school teachers knew the term "language corpora" less, only three (30%) out of 10 respondents answered affirmatively. It is worth mentioning that two of the three respondents who were familiar with the term were native speaker teachers. Corpora software is available more at universities than at secondary schools, mainly because of scientific researches and deeper study of language. Hence, I reckon that the level of familiarity that teachers have with language corpora depends on the accessibility of corpora software.

4.1.2 Question No. 2

The second question was "Do you know the purposes of language corpora? Can you name any?"

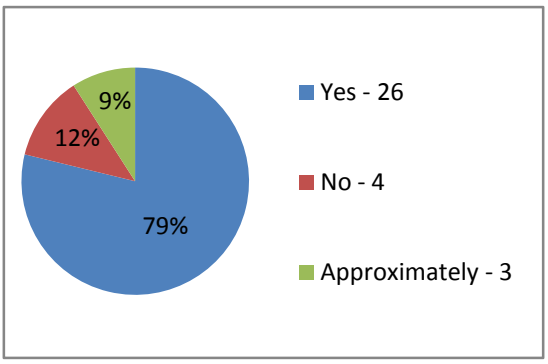

Chart No. 5 – University lecturers

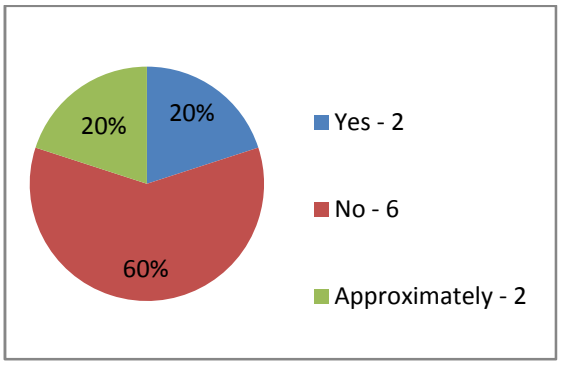

Chart No. 6 – Secondary school teachers

The Chart No. 5 shows that 26 (79%) out of the 33 addressed university lecturers knew the purposes of language corpora. Three (9%) of them knew approximately their purpose and four (12%) teachers did not know the purpose at all. There also were various kinds of corpora that are known among the respondents, e.g. The British National Corpus was mentioned by five teachers, Collins Wordbanks, MICASE corpus , LOB corpus, LUND corpus and the Brown Corpus were mentioned each once. Teachers also knew Czech corpora, e.g. four teachers knew Český národní korpus and one teacher named Pražský závislostní korpus.

Three respondents mentioned their ideas about the use of language corpora, e.g. to ascertain the lexis, to verify correctness and distribution of words and collocations in various contexts (journalistic or technical styles), to teach collocations and idioms or for discourse analysis, semantic analysis or for dictionaries compilation. One respondent referred to seminars and conferences held by Cambridge University Press (CUP). After visiting one of these courses the teacher learned about the corpora which are made by CUP in order to compile new textbooks and dictionaries, e.g. the Cambridge International Corpus or the Cambridge Learner Corpus.

The Chart No. 6 shows that six (60%) out of ten secondary school teachers answered negatively, two (20%) knew the purposes approximately and two (20%) of them were familiar with the purposes. Surprisingly, the two respondents who answered affirmatively were native speaker teachers. They claimed that language corpora are useful for linguists who look at language performance, not just at grammatical utterances. They also used

corpora to find collocations, to determine current usage patterns and frequencies of words, grammatical and idiomatic expressions, and colloquialisms of the language or common mistakes.

I suppose that this question is connected to the first one. If teachers know the term "language corpora", they are also able to name concrete corpora and to specify their purposes.

4.1.3 Question No. 3

The third question is “Do you have access to language corpora at your workplace? If yes, could you specify to which ones?”

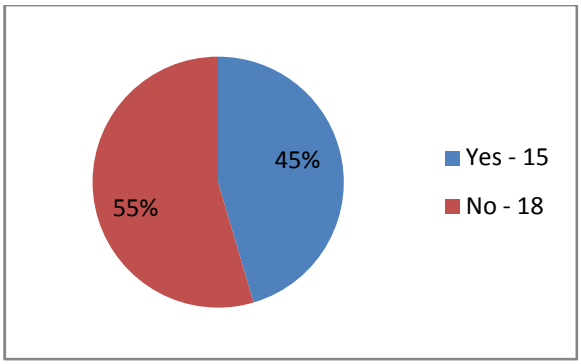

Chart No. 7 – University lecturers

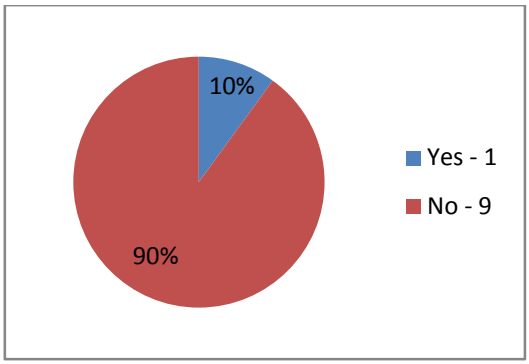

Chart No. 8 – Secondary school teachers

The Chart No. 7 indicates that 15 (45%) out of 33 university lecturers had access to language corpora at their workplace and 18 (55%) did not have access to official corpora software. Among the language corpora which were at the teachers' disposal were the British National Corpus and its SARA online software which was mentioned by 11 teachers, the Cambridge International Corpus, MICASE corpus, ICAME corpus, COCA corpus and BASE corpus were mentioned by one teacher. The ANC corpus is also used by one teacher from the respondents. Also Czech language corpora are used at universities, such as Český národní korpus which was mentioned by six teachers, or translational corpora Kačenka or K2, those two were mentioned by one teacher. Online corpora which are freely accessible on the Internet were mentioned by four teachers. Two of the addressed teachers also had their own corpora which had been created by them and their students for various purposes, e.g. to compile a list of articles on informatics or for their own scientific research.

The Chart No. 8 shows that nine (90%) teachers from secondary schools did not have access to any corpora software. Only one teacher (10%), native speaker, claimed that he

had access to language corpora but these were only extracts from the course books and dictionaries.

In my point of view, grammar schools do not participate in projects which would make use of language corpora thus teachers in fact do not need access to language corpora.

4.1.4 Question No. 4

The fourth question is "Are you interested in corpus linguistics and corpus studies?" Findings in these charts show that interest in corpus linguistics differs among university and secondary school teachers.

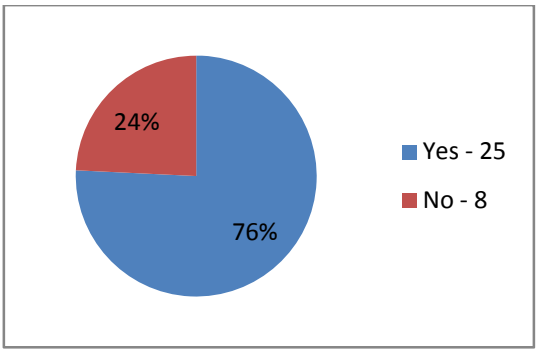

Chart No. 9 – University lecturers

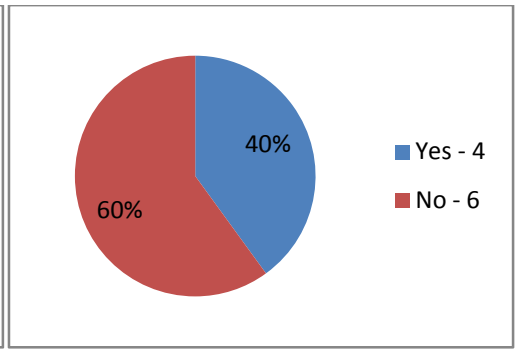

Chart No. 10 – Secondary school teachers

The Chart No. 9 indicates that 25 (76%) university were mostly interested, although 18 lecturers out of these 25 did not have access to software or they even did not work with them at all. These lecturers were interested in the use of corpora for many reasons, e.g. it is a benefit when teaching a foreign language. Eight lecturers (24%) who answered negatively did not come across this term or they did not know the purpose of corpora and thus they saw no sense of learning about them. Another fact why lecturers were not interested in corpora was that they could not use them in their current work but they could imagine that there might be projects in which corpora could be useful.

The Chart No. 10 shows that four (40%) out of ten secondary school teachers were interested in language corpora and corpus linguistics even though all of them did not have personal experience with any kind of software. The six teachers (60%) not interested in language corpora did not have access to corpora and they could not think of the purpose of them.

Findings from these charts give evidence of the existing link between the access to corpora software and the interest in corpus linguistics. I assume that respondents cannot imagine practical use of corpora unless they have chance to work with them.

4.2 Analysis of the Answers from the Open Questions Part

The second part of the questionnaire was meant to be filled in by the teachers who have access to language corpora or had experience with them. However, this was not a pre-condition thus the second part was filled in also by the teachers who did not use corpora for their work but who were interested in corpus studies.

This part gave the respondents a chance to express their opinions. It was aimed to learn more about language corpora and their position in pedagogy by descending from the general to the particular since each teacher had her/his specific experience which was essential for my research. Due to that it was possible to process all the answers and get the idea about the real situation of the role of language corpora at universities and secondary schools. Finally, the answers helped to evaluate the future prospects of language corpora both at universities and secondary schools.

4.2.1 Question No. 5

The fifth question was: “Do you think that studies of language corpora are more suitable for secondary schools or for universities?”

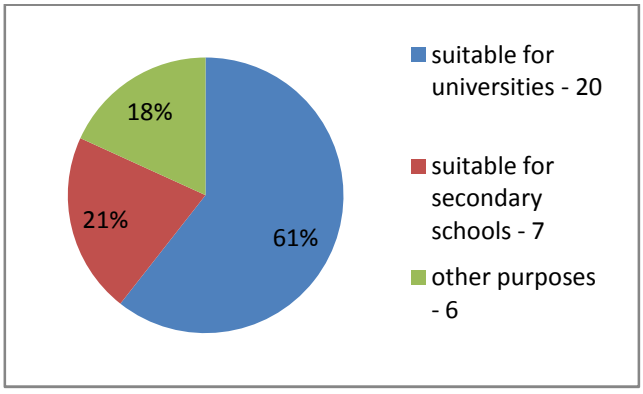

Chart No. 11 – University lecturers' opinion

I would like to express the main division of the answers with the Chart No. 11. It is clear that most of the university lecturers thought that the use of language corpora at universities would be effective; it was 20 (61%) out of 33 respondents. Seven lecturers (21%) thought it is also suitable for secondary schools and six lecturers (18%) gave other examples.

The general opinion was that language corpora are very specialized tools for studying languages and it is proper only for universities, mainly in English branches. One lecturer thought that it is not suitable for secondary schools because students at secondary schools

should learn prescriptive language, not descriptive. She also indicated that language corpora depict the development of language throughout the time, thus it is rather descriptive and corpora are not much useful for teaching secondary school students. She suggested that corpora are appropriate for universities where they have a wide use, e.g. they are useful in teaching, in scientific research or in writing thesis. One teacher wrote that he used corpora for linguistic analysis or for pedagogical use, e.g. which words to teach first.

One teacher expressed her opinion that it is not a question of age or a type of school if a student is able to work with a corpus, but of a learner's abilities. Thus it depends on such factors as intelligence, ability to use a corpus properly or a type of course which students participate in. In that case corpora could be suitable for both types of schools and it is up to every individual how she/he can make use of a corpus.

One teacher mentioned that using the Internet in a way a corpus is used can be also useful at elementary schools. For example when students have to build a phrase and they are unsure about it, they can use the Internet and they can find real existing phrases via search engines.

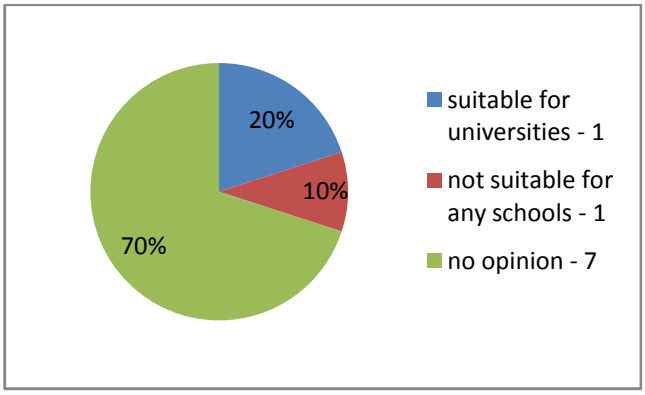

Chart No. 12 - Secondary School Teacher's Opinion

Answers from secondary school teachers were not so descriptive. According to the Chart No. 12 we can see that only three teachers (30%) out of ten expressed their opinion. One native speaker thought that language corpora were not suitable for any kind of schools at all but only for research. And the two remaining teachers thought that language corpora were suitable only for universities in conjunction with theoretical linguistics. However, one of these two teachers, a native speaker, mentioned that corpora were perfectly suitable for secondary school students who were interested in language and they might use dictionaries compiled on the basis of corpora.

According to the answers I assume that language corpora are indeed more suitable for university students because they can use them for their research or for studying linguistics rather than general English. Besides, lecturers at universities have more freedom to teach what they want and how they want. Nevertheless, I would suggest that secondary school students should be at least informed about language corpora and their possibilities of use. I think that language corpora could be used by secondary school teachers but their students need to be conducted when working with them.

4.2.2 Question No. 6

The sixth question was: "Should the work with language corpora be a part of compulsory education or an elective subject? Why?"

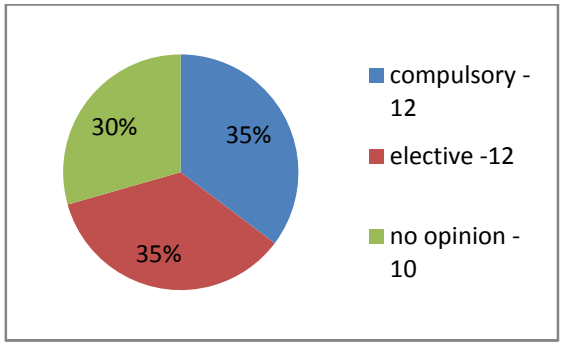

Chart No. 13 - University lecturers' opinion

The Chart No.13 shows that the answers were very balanced. Ten lecturers (30%) out of 33 did not give their opinion, twelve lecturers (35%) thought the work with language corpora should be a part of compulsory education, twelve (35%) lecturers considered language corpora to be an elective subject. The reasons were various.

The twelve lecturers who wanted language corpora to be a compulsory subject suggested that this subject would be appropriate for learners studying philological branches (i.e. courses dealing with translation, grammar or lexicology). They claimed that linguistic theories could not be explained without knowing basic facts about a corpus. One teacher thought that language corpora could not be omitted because they had become largely important for linguistic studies. Further, she also claimed that students would not manage to study properly without language corpora. There was also an opinion that nowadays empirically based research could not be made without such detailed databases as language corpora are. Some universities already offer compulsory subjects using corpora or dissertations are assigned on the basis of a corpus.

One teacher was strongly for the idea that work with corpora should be a part of compulsory education. Moreover, he suggested that corpora could be integrated into current courses, e.g. in practical use of language or in teaching how to write in a foreign language.

Further detailed study of corpora can be an elective subject but every university student should know what language corpora are and what particular benefit can be drawn from a corpus. Access to corpora software at universities is also important – students of a linguistic branch should have a possibility of working with a corpus.

The remaining twelve teachers thought that work with corpora should be an elective subject. The main reason was that a skill to work with a corpus is beyond the bounds of higher education and it is too academic and specialized.

Many universities also offer elective subjects on corpus linguistics, e.g. Masaryk University in Brno or University of Ostrava.

One of them even claimed that the quality of students had gone down in the last decade and they had to work on their proficiency in the first place, even though a corpus could show trends in language and be useful for linguistic purposes.

Two university teachers also proposed that it could be useful to have an elective subject concerning language corpora at secondary schools because some students were involved in corpus linguistics but it should be voluntary.

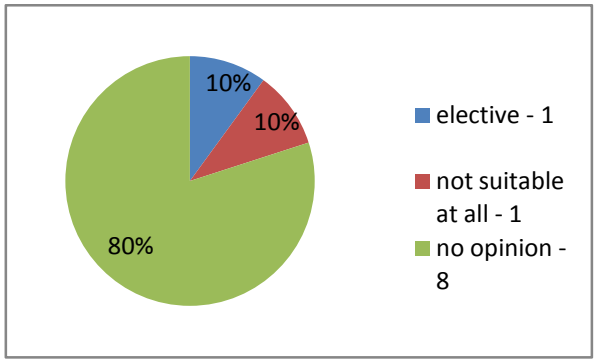

Chart No. 14 - Secondary School Teacher's Opinion

Findings in the Chart No. 14 show that only two out of ten secondary school teachers expressed their opinion. It is worth mentioning that those two were native speakers. The first one is the teacher who thought corpora were not suitable for schools at all thus she answered this question similarly. She suggested that corpus linguistics should not participate in any school's curriculum. The second teacher thought that the use of corpora

made the most sense in the context of linguistic theory, e.g. for testing the theories but it seemed unnecessary for secondary school students. He mentioned that he used Collins Cobuild English Grammar as a useful reference tool for language learners but that deep individual study of language corpora was not necessary at secondary schools.

All the answers proved that corpus linguistics is considered to be a matter of universities. Since the same quantity of university teachers stood either for compulsory education or for an elective subject there is no clear conclusion and I would suggest further studies in order to understand teachers' attitude. In my point of view corpus linguistics should be a part of university curriculum but only as an elective subject. Definitely more in-depth investigation into the topic would be needed if we were to decide whether corpus linguistics should be a compulsory or an elective subject.

4.2.3 Question No. 7

The seventh question was: “What are the advantages and disadvantages of the work with language corpora in your opinion?” This question elicited a lot of response. All the university teachers named many advantages as well as disadvantages.

The most frequently mentioned advantage was the large amount of information in one collection of texts and its easy availability. It has ever been so easy to have an access to so many examples from authentic discourse, all systematically arranged. Due to that it is possible to look for concrete information very quickly, to analyze the huge amount of data, to gather documents for a research, to do a statistical survey or to search morphological and syntactic structures, collocations and idioms.

One respondent claimed that context is the most important factor when explaining language rules. Thus he worked with language corpora a lot because they provided him and his students with hundreds of examples how each word appears in difficult contexts.

The respondents also mentioned a corpus's advantage for their students. Also weaker students are able to sort data statistically without making theoretical analysis (which had to be made formerly) and they have access to the living language and current vocabulary. Another teacher, a native speaker, indicated that work with a corpus is a funny and adventurous way how to learn a language. Due to that students become more independent when they acquire linguistic knowledge in this kind of way.

Only two secondary school teachers, native speakers, answered this question. One of them was very enthusiastic about the use of language corpora and found no disadvantage at all. The main advantage that he thought of was preparation of course books and

dictionaries. The second native speaker mentioned that it is advantageous to use texts or reference materials which had been prepared and analyzed by universities. His students got a solid sense of preferred usage and they could see how the language is actually used.

Let us now turn on disadvantages which were mentioned. They were more varied because each respondent named many disadvantages that he had experienced on his own. Anyway, there were three main disadvantages that the respondents had in mind: time, price and representativeness. Firstly, it is often time-consuming to sort the data out and to analyze them. Secondly, corpora software is quite expensive and not every school can afford to buy it. Finally, the language in corpora does not have to be always representative, it is just a fraction of discourse and no one can fully rely on that or make rules on the basis of a corpus.

Two teachers mentioned a connection with the Internet. Big search engines can function as corpora as well. For example Google can do similar work and it can occasionally supersede the idea of pure text searching. Thus language corpora are not as unique tools for linguistic analysis as they suppose to be.

Another teacher complained about the impossibility to find more information about the sources that corpora had drawn from. She would like to know more about the phrases in a corpus, about the speakers and the concrete situations.

Among commonly named disadvantages was the problem of an insufficient tagging of semantic categories, e.g. denotation, expressivity or emotionality. These factors limit researchers when they carry out a survey.

Only one of the two native speakers from secondary schools found disadvantages about language corpora. He considered corpora to be unnecessary and useless for his students. Corpora consist of millions of words, which is too demanding and not particularly user-friendly for students at secondary schools. He preferred to select a set of texts for his students to work with.

In order to evaluate all the answers I think the main advantage is that language corpora are useful tools for linguists and students which help to understand a language, its evolution and variety. I think the main disadvantage is that corpora are expensive and thus they are not much available at universities and secondary schools. For those who are deeply involved in corpus linguistic and their university cannot afford buying corpus software I would suggest trying to obtain a grant. Many corpora softwares are also available on the Internet.

4.2.4 Question No. 8

The eighth question was: “Could you think of spheres of language corpora use in English language teaching?” This question is also based on personal experience thus there were many various answers, all with common signs.

Among the things that the university teachers agreed on is a wide use of language corpora. One teacher answered that there are unlimited possibilities of use, e.g. a lector is able to demonstrate grammatical phenomena, idioms, phraseology, irregularities, the use of lexicon, the acquisition of knowledge about frequency of concrete words, or collocations in context on many examples which corpora offer. Students using corpora can verify morphological and syntactical structure or prepositional phrases.

Among the other reasons was also the form of learning. One respondent mentioned that work with a personal computer could be entertaining and not stereotyped. I think that this kind of studying is very attractive mainly for younger students. Moreover, a student works as a researcher when searching information and it is a creative work, unlike learning theories by memory.

One respondent from a pedagogical faculty prepared future teachers to use corpora, how to work with them in correcting written work in particular, and extracting useful illustrative sentences for a variety of teaching purposes. He also uses corpora for analyzing discourse or for stylistic analysis.

Another teacher, also from pedagogical faculty, had quite a different opinion of using corpora in English language teaching. She claimed that it is too demanding to create materials for teaching with language corpora. She proposed that it would be good if their department had one non-teaching member of the staff who would just help them to create activities, using corpora, to suit their needs. But it would be impossible at small universities.

One respondent claimed that he doubted that the methodology of work with language corpora is well-elaborated. New methods should be used in order not to waste time and to increase efficiency of the work with corpora.

A different idea appeared among the answers and it was about the real use of language corpora for advanced students. One respondent considered the well-known benefits of corpora to be useful but he doubted that they are helpful for students unless they could really apply results of corpora studies in practical language use.

One native speaker gave an interesting example from his own experience. He wrote about the possible use of language corpora. He would draw from them when showing his

students examples from discourse which they could not remember. For example, this respondent corrected the use of the expression ‘on the market’ to ‘in the market’. To him, ‘on the market’ meant ‘for sale’. However, ‘on the market’ seemed so widely used in Europe that he decided that he should give up. To conclude, this teacher would like to use corpora to confirm this concrete example and many others.

This question was answered by the two native speakers teaching at secondary schools as well as at the previous questions. The first respondent did not think that language corpora are applicable at secondary schools, at least not yet. Anyway, he would be happy if the A-level exams were replaced by an exam that properly tests pupils' abilities to understand, speak and write accurate, comprehensible English. Also, if English literature was taught in secondary schools, then it would be interesting to apply language corpora of different authors during lessons. The second respondent named grammar books and dictionaries which were corpus-informed as useful for reference.

In conclusion, use of language corpora in English language teaching is unlimited and it is hard to describe all possibilities because every teacher prefers a different method. However, I think that students should participate in corpora compiling and they should use corpora by themselves in order to learn how to work with them. One of the opinions was that the potential of language corpora is not fully realized in education and that corpora offer more possibilities than it is really used. Hence, I would propose that teachers could involve their students more into work with corpora and lead them, so they will not be passive. I do not think that corpora are inapplicable in education, even at secondary schools. It depends on every school, consequently on the stuff, how they will exploit the possibilities that language corpora offer.

4.2.5 Question No. 9

The ninth question was: “What specific function of corpora do you use in your work?” The aim of this question was to find out how language corpora are useful for teachers and to evaluate respondents' answers in order to learn about the real usage of corpora in practice.

Seven university teachers out of 31 answered that they used corpora firstly to look for collocations and subsequently they verified if the phrases existed in given contexts.

Further the respondents used corpora to make comments about stylistics, register, language variation, to examine data credibility, to prepare teaching materials, e.g. exercises and worksheets based on corpora, to examine various genres of spoken and written communication, to find word sketches or illustrative sentences.

The respondents mentioned that there is also a possibility of using corpora for lexical and grammatical analysis, to gather materials for researches, to make discourse analysis, sociolinguistic analysis, and pragmatic analysis, to generate charts on the basis of corpora or to add morphological derivations (e.g. word forms, inflected forms).

Two respondents used translational corpora. One of them worked as a translator and he gave lessons of CAT (Computer Aided Translation), thus he used special bilingual corpora with a function called translation memory. The second respondent used translational corpora for analysis of tendencies in translation. He also used Český národní korpus for his translations, mainly to search codification and the use of lexical items in the Czech language.

Another teacher used corpora very intensively. As she claimed, she tried to derive maximum benefit from a corpus and she used everything a given corpus offered, mainly concordances and a word frequency. Furthermore, she used corpora to create her own terminological vocabularies.

One teacher did not teach linguistics but literature and she would like to use corpora at her lessons but she was unsure if there some literary applications existed. I think that if one is deeply involved in language and corpus linguistics it does not matter she/he does not teach linguistics. Corpora could be useful for everybody.

Six teachers did not use any kind of corpus but they mentioned possibilities which the Internet offered. They used search engines as language corpora in order to verify phrases, collocations and word frequency.

Again, only two secondary school teachers, native speakers, expressed their opinion. The first native speaker claimed that he used corpora only when he compiled course materials for his students. The second native speaker mentioned that he used corpora to collect online texts to teach literature, i.e. only small corpora for specific purposes.

So far, it seems that the primary purpose for using a corpus is to find out more about collocations and word frequency. Other functions which corpora provide are countless since every teacher needs various information. Then it is worth mentioning that the Internet holds the function of a kind of a corpus which is easily accessible. Further studies are needed to find out how much the Internet and its search engines are reliable and substitutable. I suggest that both corpora and search engines have their specific advantageous functions and it is up to every individual which way she/he chooses.

4.2.6 Question No. 10

The tenth question was: “Are your students able to work with language corpora, possibly to create small corpora and work with them?” Asking this question I wanted to know if respondents compile corpora together with their students and if this helps them to conduct a lesson in a new way. Provided that students do their own investigation when compiling corpora I think it could be an interesting change in learning a foreign language.

The respondents split up into three groups – the first group were teachers whose students are able to work with corpora or even able to create them, the second group of respondents were those who refused this option and the first group were those who did not express their opinion.

Fifteen out of 33 university teachers thought it is possible to use a corpus in their lessons and they acknowledged that corpus linguistics is a frequently discussed topic among their students. Three teachers spoke about advanced students and how they compiled their own language corpora during their work on a dissertation and a bachelor thesis. These advanced students worked with specialized and smaller corpora or corpora which had been made by them, students studying for their first degree at university often used large and well-known corpora such as the British National Corpus or Český národní korpus.

One respondent was really enthusiastic about this methodology of conducting a lesson because she had an excellent corpus builder at her workplace so it was very easy to make corpora with students. She was very satisfied with the progress her students made. Students proved interest in corpus linguistics, they led discussions and they looked for commonly occurring errors, which was of greatest interest for them.

One respondent described his students' work during lessons. They learnt how to construct a small scale corpus during the courses at school, e.g. lexicology or grammar courses. Another respondent, who gave lessons on translation, used translation memories as small corpora and his students created their own collections of texts and terminological glossaries.

Among the fifteen respondents there were three teaches who believed that their students would be able to use corpora after a short training and thorough conduction.

Eight university teachers thought that their students were not able to work with language corpora or even to create their own small corpora. One of the reasons was that students were content when using search engines on the Internet and that was sufficient for

their school assignments. Another reason was that well-known language corpora are adequate for students' work and they did not have to create their own corpora.

Two native speakers who worked at secondary schools did not think that it would be useful for students to create their small corpora, definitely not on grammar schools. Besides, they could not imagine that there would be time for compiling small collections of texts at their workplace.

I consider the suggestion of compiling small-scaled corpora with students to be an excellent idea. Students can promote their language awareness and get to know their learning needs more. Students can make progress in learning a language while constructing a small specialized corpus. I suggest that the main problem is in the lack of time to conduct students in such a time-demanding work. No less important is the fact that language corpora are not much accessible at universities. In general, I think that teachers are enthusiastic about this possibility of learning a language.

CONCLUSION

This bachelor thesis provided a theoretical and practical introduction to the role of language corpora in linguistic study of the English language and the real use at universities and secondary schools.

The theoretical part aimed to clarify the problem of defining corpus linguistics and language corpora. Several definitions from various linguists were described. I hope these theories helped to comprehend the topic.

The practical part showed many interesting facts. First, I think that the most important factor for my survey is the number of respondents. Despite the large amount of questionnaires which were sent, only 33 university lecturers and 10 secondary school teachers answered. These data indicated that corpus linguistics is better known at universities. However, I was very glad that lecturers expressed their ideas, experience and suggestions, as well as with critique.

As I assumed, language corpora are not generally known at secondary schools and teachers are not much involved in the idea of using corpora during lessons. The reasons why language corpora are not either accessible or used are similar at both types of schools, e.g. schools cannot afford buying corpora software because of the high price, corpora are too large bodies of data and it would represent information overkill for students or the fact that there are easily accessible search engines on the Internet which could be used instead of language corpora. On the other hand, teachers were also conscious of many advantages of using language corpora at schools, e.g. corpora are useful tools when finding collocation, verifying phrases in given contexts or examining data credibility.

In order to evaluate the role of language corpora I reckon that a significant progress in the use of these collections of texts has been made. Students and teachers are interested more in corpus linguistics, they are aware of the usefulness of corpora in improving language proficiency. I think that teachers can also gain a better appreciation of their students when using such a creative tool for teaching a foreign language.

BIBLIOGRAPHY

Baker, Paul, Andrew Hardie and Tony McEnery. 2006. *A Glossary of Corpus Linguistics.* Edinburg: Edinburg University Press.

Ghadessy, Mohsen, Alex Henry, Robert L. Roseberry. 2001 *Small Corpus Studies and ELT: Theory and practice*. Amsterdam: John Benjamins.

Halliday, M.A.K, Wolfgang Teubert, Yallop Collin, Anna Čermáková. 2004. *Lexicology and Corpus Linguistics: An Introduction.* London: Continuum.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

McEnery, Tony, Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

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

O'Grady,William, John Archibald, Mark Aronoff, Janie Rees-Miller. 2005. *Contemporary Linguistics: An Introduction*. Boston: Bedford/St. Martin's.

Sampson, Geoffrey, Diana McCarthy. 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.

Semino, Elena, Mick Short. 2004. *Corpus Stylistics: Speech, writing and thought presentation in a corpus of English writing.* London: Routledge.

Stubbs, Michael. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell Publishing.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Tserdanelis, Georgios, Wai Yi Peggi Wong. 2004. *Language Files: Materials for An Introduction to Language and Linguistics*. The Ohio State University Press.

APPENDICES

Appendix I: The questionnaire

Appendix II: Dotazník

APPENDIX P I: QUESTIONNAIRE

My name is Barbora Šudová and I study at the Faculty of Humanities of Tomas Bata University in Zlín. This semester I am working on my bachelor thesis, the topic is "The Role of Language Corpora in Today's Linguistic Study of the English Language". The goal of my thesis is to study usage of language corpora in practice. I have prepared a questionnaire on this and I would like to ask you if you could answer the questions below. It is enough to cross or highlight your answers in the first part. The second part of the questionnaire are open questions. It would help my research a lot if you could give answers to these questions and express your opinions. All the data will be used only for the purposes of my bachelor thesis. Thank you for your time and willingness.

1. **Have you ever come across the term "language corpora" in connection with your work?**

   ☐ Yes

   ☐ No

2. **Do you know the purposes of language corpora? Can you name any?**

   ☐ Yes

   ☐ No

   ☐ Approximately

3. **Do you have access to language corpora at your workplace? If yes, could you specify to which ones?**

   ☐ Yes

   ☐ No

4. **Are you interested in corpus linguistics and corpus studies?**

☐ Yes

☐ No

Please answer the following questions if you use language corpora or have experience with them:

5. **Do you think that studies of language corpora are more suitable for secondary schools or for universities?**

6. **Should the work with language corpora be a part of compulsory education or an selective subject? Why?**

7. **What are the advantages and disadvantages of the work with language corpora in your opinion?**

8. **Could you think of spheres of language corpora use in English language teaching?**

9. **What specific function of corpora do you use in your work?**

10. **Are your students able to work with language corpora, possibly to create small corpora and work with them?**

APPENDIX P II: DOTAZNÍK

Jmenuji se Barbora Šudová a studuji na Fakultě humanitních studií, Univerzita Tomáše Bati ve Zlíně. Tento semestr zpracovávám bakalářskou práci na téma "The Role of Language Corpora in Today's Linguistic Study of the English Language". Cílem mé práce je zkoumat využití lingvistických korpusů v praxi. Připravila jsem na toto téma dotazník a ráda bych Vás požádala, zda-li byste mohli zodpovědět následující otázky. V první části stačí Vámi vybrané odpovědi zaškrtnout či zvýraznit. Ve druhé části jsou otevřené otázky. Velmi by mému výzkumu pomohlo, kdybyste odpověděli na tyto otázky a vyjádřili tak svůj názor. Všechna data budou použita pouze pro účely mé bakalářské práce. Děkuji za Váš čas a ochotu.

1. **Setkali jste se ve své praxi s termínem "jazykové korpusy"?**

   ☐ Ano

   ☐ Ne

2. **Víte, k jakým účelům se jazykové korpusy využívají? Můžete nějaké jmenovat?**

   ☐ Ano

   ☐ Ne

   ☐ Přibližně

3. **Máte na Vašem pracovišti přístup k jazykovým korpusům? Pokud ano, k jakým?**

   ☐ Ano

   ☐ Ne

4. **Zajímá Vás problematika jazykových korpusů?**

☐ Ano

☐ Ne

Následující otázky jsou určeny pro ty, kdo jazykové korpusy využívají. V případě, že k nim máte přístup, prosím odpovězte na následující otázky:

5. **Myslíte si, že jsou jazykové korpusy vhodnější pro výuku na středních školách nebo na vysokých školách?**

6. **Mělo by být studium a práce s jazykovými korpusy součástí povinné výuky nebo jako volitelný předmět? Proč?**

7. **Jaké jsou podle Vás výhody a nevýhody jazykových korpusů?**

8. **Jaké je podle Vás využití jazykových korpusů ve výuce angličtiny?**

9. **Jaké konkrétní funkce korpusů při své práci využíváte?**

10. **Jsou Vaši student schopni vytvořit jednoduchý korpus a pracovat s ním?**