universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Europarl Corpus HFWL Glossaries: A Set of Multilingual Glossaries Based on the Quantitative Corpus-Based Lexical Analysis of the Europarl Corpus

verfasst von / submitted by

Bc. Marek Paulovič, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 065 331 342
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Dolmetschen UG2002 Deutsch Englisch
Betreut von / Supervisor:	Univ.-Prof.Mag.Dr. Gerhard Budin

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Univ.-Prof. Mag. Dr. Gerhard Budin for his continuous and immense support, for his patience, motivation, and considerable knowledge in this field. His guidance has helped me throughout the duration of researching and writing of this thesis. I could not have imagined having a better advisor and mentor for my master thesis.

Besides my advisor, I would like to thank PhDr. Jana Rejšková and Mgr. Věra Kloudová, Ph.D. for their insightful comments and help with identification of the problematic internationalisms from the both didactic and linguistic perspective.

My sincere thanks also goes to Beth Garner for her immense help with proofreading of this master thesis. Without her encouragement, I might have not decided to write this master thesis in English.

I would like to also thank to my loved ones, family and friends who supported me throughout entire process, both by keeping me harmonious and helping me with putting the pieces together. Last but not least I would like to thank God for the good health and wellbeing that were necessary to complete this thesis.

Abstract

The main goal of this thesis was to experimentally create a set of multilingual glossaries (English/German/Czech/Slovak) of the most frequent words in the European Parliament using quantitative corpus-based lexical analysis of the Europarl corpus. The result was aimed to help advanced non-native speakers of English to acquire active vocabulary in their language combination for effective oral communication in the political discourse. Throughout the thesis, the inductive transdisciplinary approach is applied - creating a useful language learning tool represented the main research problem that was to be solved through application of knowledge from various disciplines such as English language teaching, speech science, lexicology, terminology, language for special purposes with a special focus on politolinguistics, computer assisted vocabulary learning, and finally, corpus linguistics as a working method for corpus analysis. Based on our theoretical review, it can be confirmed that: a) learning HFWL is the most effective learning method for vocabulary acquisition considering the amount of time given and the delivered results, b) learning vocabulary aimed at its automatization may lead to speech output improvement, and c) political language can be regarded to a certain extent as a language for special purposes. The results of literature review are summarized in the proposed cognitive model for spoken speech production. In the practical part, the theoretical knowledge of corpus linguistics was applied for creating a set of four multilingual glossaries. This thesis puts the case for using multilingual glossaries based on high frequency word lists as a didactic tool, but the hypothesis still remains to be validated in other experimental studies.

Keywords: Europarl corpus, glossary compilation, keyword analysis, LSP, political language, vocabulary learning, corpus-based lexical analysis, HFWL

Abstrakt

Die vorliegende Arbeit befasst sich mit der Erstellung 4 multilingualer Glossare (Englisch/Deutsch/Tschechisch/Slowakisch) der häufigsten Wörter im Europäischen Parlament, deren Auswahl auf einer quantitativen, korpusgestützten, lexikalischen Analyse des Europarl Korpus basiert. Das Ziel war es, anhand der Korpusanalyse ein Hilfsmittel für fortgeschrittene englische Nicht-Muttersprachler zusammenzustellen, das den Erwerb des aktiven Wortschatzes im politischen Diskurs erleichtert und folglich zur Verbesserung der mündlichen Sprachkompetenz führt. Die Masterarbeit basiert auf dem induktiven transdisziplinären Ansatz; Das Vorhaben, dieses Hilfsmittel zusammenzustellen, stellt das Forschungsproblem dar. Dieses wurde durch die Anwendung von Wissen aus unterschiedlichen akademischen Disziplinen gelöst, nämlich dem Sprachunterricht, der Sprachwissenschaft, der Lexikologie, der Terminologie, der Fachsprachenforschung, genauer genommen der Politolinguistik, dem computergestützten Sprachunterricht und zu guter Letzt der Korpuslinguistik als die Untersuchungsmethode. Auf der Grundlage der recherchierten Informationen konnte eruiert werden, dass a) das Erlernen von Häufigkeitswortschatz, gemessen an den Lernergebnissen und dem Zeitaufwand, die effektivste Lernmethode für den Wortschatzerwerb ist, b) ein Wortschatzerwerb, der auf die Automatisierung des aktiven Wortschatzes abzielt, zur Verbesserung der mündlichen Sprachkompetenz führen kann, und c) die politische Sprache in gewissem Maße als Fachsprache betrachtet werden kann. Die Ergebnisse der Literaturrecherche wurden beim Entwurf eines synthetisierten Models für die mündliche Sprachproduktion angewendet. Die in der Masterarbeit präsentierten vier Glossare bestehen aus einer quantitativen Korpusanalyse des Europarl Korpus und einer qualitativen Korpusanalyse der identifizierten 3000 häufigsten Wörter und der 1500 häufigsten Schlüsselwörter. In der Masterarbeit wird vorgeschlagen, das Lernen von Glossaren, die auf Häufigkeitswortschatz basieren, als didaktische Methode anzuwenden, aber diese Hypothese ist noch bei anderen experimentellen Studien zu verifizieren.

Schlüsselwörter: Europarl Korpus, Erstellung von Glossaren, linguistische Schlüsselwortanalyse, politische Sprache, Politolinguistik, Wortschatzerwerb, computergestützter Spracherwerb, korpusgestützte lexikalische Analyse, HFWL

TABLE OF CONTENTS

Introduction

When I pondered on a suitable topic for my master thesis, my first criterion was the practicality of the paper not only for me, but also for the interpreting community. I remember from my master studies of conference interpreting quite vividly, how insufficient vocabulary knowledge can thwart one's attempts for flawless interpreting. Bearing in mind that the highest accolade for an interpreter is to be employed by a supranational institution such as the European Parliament or the OSN and many speeches used in university interpreting training come from a political setting, I assumed that it would be of great interest for interpreting students, graduates and novice interpreters like me to grasp the political language used in this discourse.

After a preliminary study of research options, I came to the conclusion that it would be an asset to investigate the Europarl Corpus, which contains numerous proceedings from the European Parliament. The investigation would take place on a quantitative lexical level, and the most frequent words would be extracted. My search for suitable methods brought my attention to the area of corpus linguistics, a completely new and unknown field to me. Exploring this new area, I became increasingly aware of very blurred confines of translation studies as an academic discipline and the importance of transdisciplinarity which plays a major role in translation research. This led me to the thought that if translation studies is becoming so intertwined with other disciplines by borrowing their knowledge and methods for research, it is only natural to expect the research findings of translation studies also enriching and of value for other users beyond the interpreting community. Therefore, I suppose that creating a multilingual glossary based on the Europarl corpus for novice interpreters, as was my initial incentive, could also benefit other target groups such as students of political sciences, sociolinguists and applicants to various position in European institutions.

This thesis is based on an inductive approach: we firstly have defined a need or a problem, then looked for theoretical knowledge that would support our assumption. Next, we searched for suitable tools and methods for optimally reaching our research objective. Secondly, we attempted to create a useful glossary with pedagogical implications in accordance with terminological standards. Here we have to point out that without prior

knowledge about second language teaching, we would only stumble in the darkness in the correct glossary conception which would be based on intuition rather than solid scientific background knowledge. For this reason, we also decided to explore the usage of high frequency word lists (HFWL) in language teaching literature. We identified some pedagogical standards, but the literature review did not provide an answer to 2 inherent questions: firstly, why it is important to learn high frequent vocabulary in the first place, and secondly, how vocabulary acquisition reflects in second language production. We searched deeper and after inquiring into the fields of lexicology and speech production, we believed that we found answers closed the gap in the uncertainty about the purpose, usage and effectiveness of learning vocabulary from high frequency word lists.

We were also aware of the fact that creating a functional glossary is only the first part of the process. Even in a best-case scenario of creating a scientifically proven effective glossary for grasping the political language of the EU, it would not mean much if the glossary does not reach its target audience. For this reason, we had to also think about firstly finding an effective interface and secondly about dissemination of the glossary to its potential users. We sought to achieve an optimal solution by consulting CAVL or computer assisted vocabulary learning and available software for interpreters, identifying the key players and creating glossaries that could be imported to the software.

We sincerely hope that we have successfully accomplished this intrinsic transdisciplinary journey and managed to answer all important questions arising from our research.

Study aims

The aim of this study is to use the quantitative corpus-based analysis to identify the most common words that could help non-native English speakers improve their speech proficiency in the political discourse. We believe that this could be achieved by creating a glossary that represents the most frequent words used in MEPs' speeches in the European Parliament. To support this assumption, we aim to confirm 3 research hypotheses:

a) A glossary of the most frequent words can be a useful tool for language learning, AND

b) Learning vocabulary and its subsequent automatization may lead to better speech production, AND
c) Political language of the European Parliament is distinct through its special vocabulary and it can be classified as a language for special purposes.

Chapters Summary

This paper is divided into 2 parts: the theoretical part and the practical part. The theoretical part includes first 4 chapters, the practical part the last one.

The first chapter, *The Theoretical Framework for Inter- and Transdisciplinarity*, should introduce the reader to the concept of transdisciplinary research with a special emphasis on translation and interpreting studies. This should enlighten the need for transdisciplinarity in achieving our research objective.

The second chapter titled *Main Theoretical Framework for Glossary Compilation* deals in a great extent with high frequency word lists and the concept of speech production. It describes the usage and efficiency of high frequency word lists in language teaching, and explains, how automatization of vocabulary through learning high frequency word list could benefit the quality of spoken language. This required to consult language teaching and speech science as academic disciplines.

The aim of the third chapter *Supportive Theoretical Framework for Glossary Compilation* is to provide knowledge for proper glossary compilation and distribution. The area of focus ranges from defining political language as a language for special purposes, important insights from lexicology and terminology and exploring possibilities of computer-assisted language learning.

Finally, the forth chapter *Corpus Linguistics as the Method for Creating LSP Glossaries* covers the field of corpora and corpus linguistics as the methodology used for reaching our research objective. It discusses corpora, their types, usage, limitations, annotations, and corpus linguistics as the methodology with a great transdisciplinary overlap. Due to the scope of the paper, substantial attention is given to corpus analysis, main tools, statistical methods and their practicable usage.

The practical part of this paper discussed in chapter 5 describes the procedure followed in the corpus analysis of the Europarl Corpus and the subsequent glossary compilation based on corpus analysis findings. It also introduces 4 types of glossaries that

are presented as the result of optimal utilization of gather knowledge from the theoretical part.

Theoretical Framework for Inter- and Transdisciplinarity: Towards Transdisciplinarity in Translation Studies

Introduction

As outlined in the introduction, we consider it particularly important to elucidate on transdisciplinarity in translation studies in order to make a clear link between the research methods & knowledge used for the corpus analysis and possible application of its results.

We will firstly briefly define the terms *translation studies* and *transdisciplinarity*, and then illustrate the practical application in our study

Translation Studies

Translation studies or translatology, as presented and described by Baker, is the academic discipline which concerns itself with the study of translation. Interest in translation is old as human civilization and there is a vast body of literature on this subject which dates back at least to the first century BC. Yet, as an academic discipline, translation studies is relatively young and it was not until the second half of the 20th century that scholars began to discuss the need of conducting systematic research on translation and developing coherent translation theories. Before that, translation was in the academic field present only as a part of other academic disciplines such as comparative literature or contrastive linguistics. Nowadays, however, translation studies comprises a great number of fields, for instance technical or literal translation, subtitling, dubbing, conference, liaison or community interpreting and many others, which encompasses in essence both translation and formulation of written texts, and oral translation of spoken language (interpreting) Translation studies is also understood to cover the whole spectrum of research and pedagogical activities, from developing theoretical frameworks to conducting individual case studies to engaging in practical matters such as training translators and interpreters, and developing criteria for their assessment (Baker, 2001).

However, with the rising demand on various specialization fields in translation studies and through the academic endeavors to define these specializations strictly in order to justify them in regard to the others, there has come to unnecessary fragmentation

within translation studies in which disciplines seeking their individuality forgot that they should rather complement each other. An eloquent discussion on this topic can be found in Baker (2001):

> In the course of attempting to find its place among other academic disciplines and to synthesize the insights it has gained from other fields of knowledge, translation studies has occasionally experienced periods of fragmentation: of approaches, schools, methodologies, and even sub-fields within the discipline. At a conference held in Dublin in May 1995 for instance, some delegates called for establishing an independent discipline of interpreting studies, because theoretical models in translation studies by and large ignore interpreting and are therefore irrelevant to those interested in this field. This is true to a large extent, just as it is true that within interpreting studies itself far more attention has traditionally been paid to simultaneous CONFERENCE INTERPRETING than to other areas such as COMMUNITY INTERPRETING and liaison interpreting. However, the answer in both cases cannot lie in splitting the discipline into smaller factions, since fragmentation can only weaken the position of both translation and interpreting in the academy. The answer must surely lie in working towards greater unity and a more balanced representation of all areas of the discipline in research activities and in theoretical discussions (Baker, 2001, p. 279).

The attempt to establish an independent discipline is based on numerous reasonable and solid arguments (see Pöchhacker, 1994, 2004) that clearly distinguish interpreting, as rendering of a spoken language from translation, rendering of a written language. There might also be a purely semantic reason leading distinction attempts: English language has no unifying term for interpreting and translation that would not favor one term to another as opposed to German, Slovak or Czech (Translationswissenschaft: Übersetzen und Dolmetschen/Translatológia: prekladateľstvo a tlmočníctvo/Translatologie: překladatelství a tlumočnictví vs. translation studies: translation and interpreting). There can be no doubt that it is tremendously important to distinguish between interpreting and translation, but not to the extent that it would separate them completely from each other, as pointed out by Baker.

Baker calls translation scholars to unify and to recognize that no approach, however sophisticated, can provide the answers to all the questions raised in the

discipline, nor the tools and methodology required for conducting research in all translation studies areas. He points out that there can be no benefit in setting various approaches in opposition to each other, nor in resisting the integration of insights achieved through the application of various tools of research regardless of their origin (Baker, 2001). This complementarity rather than mutual exclusivity of individual areas of translation studies suggests that the application of transdisciplinarity, a widely accepted approach in natural and technical sciences, could be the key in evolving translation studies. It should help them address issues emerging in this academic discipline more effectively and practically.

Transdisciplinarity

Transdisciplinarity is a term describing application of various disciplines with an inductive approach. The research question comes outside from the real world and the researchers are expected to consult and apply knowledge from relevant scientific disciplines to address the issue. This approach is different from inter-disciplinary research in which the research problem lays in the common interest of two disciplines, or from a multi-disciplinary research in which multiple disciplines share a broad subject of investigation.

Harvard website for transdisciplinary research defines transdisciplinarity as:

> research efforts conducted by investigators from different disciplines working jointly to create new conceptual, theoretical, methodological, and translational innovations that integrate and move beyond discipline-specific approaches to address a common problem (Harvard, 2015, para. 1).

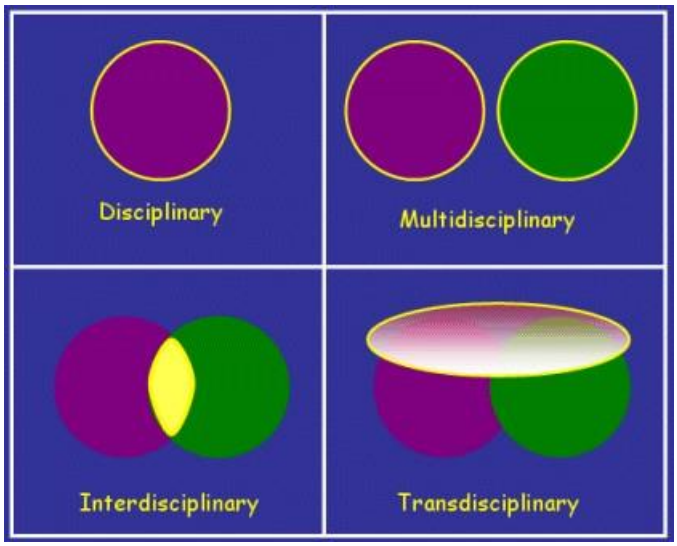
Because transdisciplinarity tends to be mistakenly mixed with interdisciplinarity, we offer a clear distinction from interdisciplinary research that can be seen in its definition provided by Aboelela, Larson, Bakken, Carrasquillo, Formicola, Glied and Gebbie.:

> Interdisciplinary Research is any study or group of studies undertaken by scholars from two or more distinct scientific disciplines. The research is based upon a conceptual model that links or integrates theoretical frameworks from those disciplines, uses study design and methodology that is not limited to any one field, and requires the use of perspectives and skills of the involved disciplines throughout multiple phases of the research process (Aboelela et al. 2007, p. 341).

A graphical explanation of disciplinary approaches can be seen in the image below. Examples of following overlaps:

a) Disciplinary: Epistemologies, assumptions, knowledge, skills, methods within the boundary of a discipline, e.g. Physics, History, Philosophy;
b) Multidisciplinary: Using the knowledge/understanding of more than one discipline. Physics and History;
c) Interdisciplinary: Using the epistemologies/methods of one discipline within another, e.g. Biochemistry, Astrophysics;
d) Transdisciplinary: Focus on an issue such as pollution or hunger both within and beyond discipline boundaries with the possibility of new perspectives (Holistic Education Network, 2011, para. 3).

Figure 1. Explanation of disciplinary approaches. Holistic Education Network of Tasmania. 2011. *Transdisciplinary Inquiry.* Retrieved from http://www.hent.org/transdisciplinary.htm.



Transdisciplinary research can be described as practically oriented and problem solving holistic approach that is not influenced by limitations within a disciplinary framework, but encourages thinking "out of the box" and bringing new perspectives. Nicolescu (1999) argues that transdisciplinarity is radically different from multidisciplinarity and interdisciplinarity for its goal and the understanding of the present world which cannot be accomplished in the framework of disciplinary research in which

the latter two remain. It is also not antagonistic but complementary to them (Nicolescu, 1999; Stavridou & Ferreira, 2010).

Transdisciplinarity in an academic discipline should not be understood as absence of disciplinary knowledge and as a hint showing its underdevelopment in regard to other disciplines, but rather as a possibility of broadening the scientific view and increasing academic self-reflection. This should lead to more flexibility and dynamics within the academic discipline (Bartoňková, 2002).

Transdisciplinarity in Transcultural Communication

The transdisciplinary approach can be very well illustrated on transcultural communication. Transcultural studies identified that the area of intercultural communication does not cover all needs arising in modern turbulent society full of a myriad of various cultures and subcultures, as suggested by Cooke (2007). It is evident that the globalization brought with itself new phenomena causing cultural "mingling". Consequently, intercultural communication, that investigated communication between cultures mainly on a national level, could not anymore address new communication needs of ever more complex society. It was rightly identified that communication should not only consider cultural differences on a national level, but also take a closer look at the individual situation, concerns, needs and communication style of "societies in society" which may comprise of different ethnic groups, subcultures, modern trend followers etc.

To communicate effectively means firstly to understand the communication partner and to look at the world through their eyes, and secondly to use this understanding and knowledge to adapt the communication style to their language. For instance, communication style of a hipster or a paleo enthusiast, much discussed cultural phenomena in 2015, might not be exceptionally different from mainstream society, but understanding them and adapting the language accordingly can increase the effectiveness of the communicated message. Adapting the language to the target audience is one of the most fundamental principles of media, PR and transcultural communication which effectively borrow knowledge from other academic disciplines to fulfill its purpose, namely to solve specific communication challenges.

Thus, transcultural communication as an academic discipline is not about merely collecting knowledge from other disciplines to build its theoretical knowledge base, but rather about finding their application in specific research questions.

There are many professional fields outside translation and interpreting, such as in media communication, PR, business and the corporate environment, as well as andragogy that approach to solving a problem with a similar inductive concept. Therefore, students of transcultural studies can easily apply their skills and knowledge in a variety of professions. This is supported by findings of Höller (2008) who interviewed graduates of translation studies working outside the translation field on the applicability and usefulness of acquired transcultural competencies in other job spheres.

Transdisciplinarity in Interpreting Studies

Taking interpreting also as an example of transdisciplinarity, it is reasonable to assume that transdisciplinarity played a key role by establishing interpreting as an academic field. Interpreting scholars firstly looked for addressing practical problems in interpreting performance and interpreting training, borrowing knowledge from translation studies, linguistics, psychology, stress research, cognitive sciences, rhetoric and many others (see Pöchhacker, 2004). However, with the academic discipline firmly established in the last decades, transdisciplinarity has been gradually replaced by interdisciplinarity.

Transdisciplinarity can be also identified in the core competences of an interpreter. Interpreters are except the relevant interpreting expertise (aka expert knowledge of source and target language and sufficient specialized field knowledge that may vary from assignment to assignment) also required to possess interpersonal, social, communicative and intercultural competences, as well as various personal traits and abilities[1] (Pöchhacker, 2004; Baumann, 2013; Krüger, 2013).

Baumann (2013) also points out that interpreting studies have broadened its focus as a result of globalization and highlights the importance of swift adaptation to new challenges and calls for more inter and transdisciplinarity in interpreting studies. These arguments suggest clearly that interpreting studies are also an inductive scientific disciplines that can greatly benefit from the transdisciplinary approach. Therefore, it might as well come back to transdisciplinarity by solving practical issues not only within, but also above and beyond translation studies.

[1] The importance of individual aspects varies from interpreting settings.

Conclusion

This theoretical overview served as an argument for the transdisciplinary approach used in this thesis. This could be outlined on the procedure of reaching our research objectives. The first inherent question was about creating a tool that could help non-native English speakers achieve better speaking proficiency in parliamentary language. We hypothesized that a multilingual glossary of the most frequent words from the political discourse could serve this purpose. After identifying from our literature review that there is no such frequency based glossaries for parliamentary language, we intended to create one. Initially, we searched for the optimal way to create the list of words for this glossary. This led us to the corpus linguistics and the Europarl corpus as the most relevant source. Next, to back our assumption about the usefulness of the glossary, we searched for supportive knowledge in relevant disciplines that brought us above and beyond what we initially expected to explore. Finally, after finding arguments supporting our assumptions, we consulted other scientific disciplines for proper compilation and presentation of the glossary.

We believe that our problem-oriented research procedures reflects as a transdisciplinary approach and hope that it will lead to broadening the scientific knowledge and some practical application. This thesis puts the case for using multilingual glossaries based on high frequency word lists as didactic tools, but this hypothesis still remains to be verified in other experimental studies.

As the reader will see in the next chapters, we have applied various knowledge and methods from corpus linguistics, language teaching, speech science, and language for special purposes, terminology, politolinguistics and interpreting studies that we synthetized in order to reach our research objective. The research was more oriented towards practical application than towards contributing to the large body of knowledge in translation studies. For this reason, we decided to apply the transdisciplinary approach that we tried to explain in this introductory chapter.

Main Theoretical Framework for Glossary Compilation

Introduction

In this chapter, we will consult various academic disciplines in order to find answers to practical questions arising from our intention to compile a multilingual glossary based on a quantitative lexical analysis of the Europarl corpus.

In the first part of this chapter, we will pay considerable attention to two main academic disciplines that can provide answers to the fundamental question *why it is important to create wordlists for second language acquisition and how it can benefit the learner*.

To answer these questions, we will firstly discuss the usage of world lists in pedagogy and second language acquisition to show pedagogical application of wordlists in second language acquisition. We will start with the history and usage of high frequency word lists (HFWL), explain the connection between the vocabulary size and the usage of HFWL discuss vocabulary categorization and acquisition, and finally present some limitations of HFWL and suggestions for correct vocabulary acquisition with help of HFWL. Secondly, we will take a closer look at the speech production itself in order to provide a scientific basis for pedagogical implications stated in the chapter above. This includes the speech production theory, explanation of automatization of language processes in special regard to bilingual lexicon and the introduction of the concept of shared attention. Then we will present a synthetized model based on gained knowledge from relevant scientific disciplines.

In the second part of this chapter, practical knowledge from relevant academic fields will be gathered for reasonable vocabulary selection and glossary compilation in accordance with existing standards. These include lexicology, terminology, language for special purposes (LSP) with a special regard to politolinguistics, and computer-assisted language learning (CALL) focused on computer-assisted vocabulary learning (CAVL)

Before diving into various scientific fields in a search for answers to our fundamental questions that would validate our intention to create a multilingual glossary based on a quantitative lexical analysis, it is necessary to define the term glossary.

A glossary is basically a list of terms[2] in a particular field of knowledge with definitions and explanations in one or more languages (see Vanopstal, Vander, Laureys & Buysschaert, 2009; Pearson & Bowke, 2002). It originates from the Latin word *glossarium,* which refers to a collection of glosses. *Gloss* stems from the Greek word *glossa* which denotes the explanation of a specialized expression or a difficult word (Vanopstal et al. 2009). The amount of information contained in glossaries may vary as well as the level of detail which usually depends on the purpose for which a glossary is intended. The glossary compilation process falls under the terminology which sets standards and guidelines (Pearson & Bowke. 2002).

There are number of reasons for building a glossary. These may include identifying equivalents in a foreign language for the terms known in the native language, providing explanation to unknown specialized terms or creating a language resource that can be later used for teaching, learning or speech production in a written/spoken form.

As we have already mentioned, the aim of this thesis is to create a glossary of the most common words used in the political discourse of the European Parliament. The glossary should serve as a useful and relevant learning tool with a potential to boost learners' second language speaking proficiency in the political language. To achieve this many other academic fields must be reviewed and discussed in the next chapters which go above and beyond terminology, lexicology and translation studies. We sincerely hope that our findings will prove relevant, feasible and valid to back our hypothesis mentioned in the introduction.

The Usage of High Frequency Word Lists in Second Language Learning & Teaching

History and usage of high frequency word lists. The review of literature and webpages supports the widely accepted opinion that learning vocabulary in general and high frequency word lists (HFWL) in particular should be an important component of learning a language.

Teaching HFWL has shown as an important aspect in early childhood education by improving reading, writing and spelling proficiency of English native pupils.

[2] See page 58.

The first HFWL list for teaching purposes was compiled by Dolch in 1936 on the assumption of a rough correlation between vocabulary knowledge and text difficulty. It consists of 220 words believed to make up from 50-75% of the reading material encountered by students. The list was later used for the classification of reading level by grades and it is presumably still used in classrooms today (Leibert, 1991).

Fry expanded on the well-established Dolch list in 1996 and produced several HFWL based on author's long-term research. The last Fry list consists of 1000 most frequent words and it is believed to be the most widely used HFWL (Farrell, Osenga & Hunter, 2013).

Despite clear pedagogical implications, it must be noted that these wordlists are used mostly for teaching primary schools pupils or beginning ESL students, hence are far from being useful for advanced ESL speakers.

However, some HFWLs were also produced for more advanced English speakers. We will now briefly describe the history and development of HFWL for academic and ESL purposes.

The General Service List (GSL) created by West in 1953 is considered to be the first HFWL created for ESL learning purposes. This list of 2000 most frequent words was envisioned by the author to be "the selection of English most suitable to set as a first objective for foreign learners" and "is trying to simplify English for the learner" (West, 1953, p. 45, as cited in Graham, 2008, p. 12). This HFWL was of superior quality – despite the fact that it was produced manually, it included semantic differentiation. That was unfortunately not adopted as a standard by other HFWL for long time after the GSL was created even during the advent of corpus linguistics brought by the computer advancement.

The Academic Word List (AWL) created by Coxhead (1998) focused on examining the most frequent occurring words of written academic texts by their range, frequency and uniformity of occurrences. Interestingly, the GSL was used here as a stop list – the AWL was focused on most frequent words outside the first 2000 words in GSL. The list itself contains only 10% of all running words in the used academic corpus and this percentage would be expected even lower in general English, but these words have been identified as strong carriers of meaning and important for vocabulary acquisition.

Gardner and Davies (2013) brought the HFWL development one step further by improving the AWL, created a searchable web database based on the COCA corpus[3] and provided useful information about contemporary HFWLs and their pedagogical implication.

One notable HFWL is the Longman Communication 3000. The above mentioned HFWLs were prevalently made for academic purposes and created from corpora of written academic texts. The Longman Communication 3000, however, gives an important insight into both written and spoken English language through categorizing the list for both written and spoken English. The analysis of the Longman Corpus Network (390 million words) shows that these 3000 most frequent words account for 86% of the language. The authors claim the list to be a very powerful tool for the development of good comprehension and communication skills in English. The list was further developed into an electronic dictionary of contemporary English (LDOCE) with powerful features facilitating the learning process (tests, phrase banks, examples in context etc.) (Longman Communication 3000, 2003).

The latter example of HFWL shows clearly that learning vocabulary can be nowadays enhanced with computer-assisted vocabulary learning (CAVL). This trend is also supported by distinguished scholars in this field. For instance, Notion (2001) sees a great potential in using CAVL for significant effectivity boost in teaching and learning new vocabulary[4].

Vocabulary size in regard to HFWL. In the previous chapter we introduced HFWL consisting of several thousand most frequent words. Although such lists are important for automatization of vocabulary5, here we must point out that an advanced ESL speaker is expected to have a much higher vocabulary size than the one suggested in HFWLs above. It is also natural to expect that these high frequency words only create the backbone of the language, and that the actual meaning of a speech utterance is conveyed by less frequent words. On the other hand, if the learner does possess sufficient

[3] See http://www.wordandphrase.info/academic/

[4] A small account to CAVL will be given in next chapters.

[5] Automatization of language processes will be discussed on page 35

proficiency in the most frequent words, there is little sense in focusing on other vocabulary until these are well learned (Waring & Nation, 1997).

When speaking about the vocabulary size, a clear distinction between the active and passive vocabulary must be made. The size of active and passive vocabulary correlates, but the active vocabulary[6] is by rule in size always smaller than the passive or receptive vocabulary[7]. Additionally, active vocabulary acquisition is reported to be slower and less effective than receptive (Laufer, 1998; Nemali, 2010).

The study by Cervatiuc (2008b) suggests that the average receptive vocabulary size of highly proficient university-educated non-native English speakers ranges between 13,500 and 20,000 word families[8] which is comparable to university-educated English native speakers with an average of 17,200 words. Just for comparison, the most conservative rule of thumb says that a native speaker has a vocabulary size of roughly up to 20,000 words which is expected to grow by about 1000 words per year (Goulden, Nation & Read, 1990). Surprisingly, similar learning speed may also be achieved by non-native speakers and it can be even higher if this learning is done in the second language environment. Some studies (Meara & Jones, 1990; Milton & Meara, 1995) provided interesting findings about vocabulary acquisition that lies in average between 2500-2650 words per year by advanced exchange students participating in these studies. This would imply that an ESL learner would need only 6,5 years to achieve near-native proficiency level (Cervatiuc, 2008a, 2008b), but here we must not forget that students usually do not have the opportunity to study abroad for longer than 6-12 months. Therefore, even though the higher effectivity of long-term immersion compared to learning in the traditional classroom context cannot be denied (Serrander, 2011), it should rather be considered as

[6] Active vocabulary represents all words that a person can actually use at his or her free will by speech production or writing – in other words, the lexical command (Leńko-Szymańska, 2002).

[7] Receptive vocabulary size is understood as "all the words in a person's language repertoire, i.e. words that a person can comprehend and respond to, even if the person cannot produce those words" (Chong, 2011, p. 1231).

[8] A word family consists of a headword, its inflected forms, and its closely related derived forms (Nation, 2001b)

a short-term boosted vocabulary expansion than a steady rate of vocabulary acquisition. For this reason, it is advised to have more modest and realistic views on speed of vocabulary acquisition that vary greatly on learning techniques and time & effort devoted to the vocabulary study.

Vocabulary categorization. As far as the active vocabulary is concerned, it is very difficult to provide statistical data about the vocabulary size as in the case of receptive or passive vocabulary. Numerous reliable statistical tests have been developed for the receptive vocabulary, but the free active vocabulary store lends itself to measurement not easily because it comprises both qualitative and quantitative dimensions, such as lexical density, lexical originality, lexical variations, sophistications, semantic variations etc. (Leńko-Szymańska, 2002). Regarding the count of active vocabulary size itself, one of the proposed quantitative methods worth mentioning is "the beyond 2000 measure" in which researchers analyze a speech sample of ESL learners and count only words not included in the top 2000 most frequent words (Eubank et al. 1995). However, the results can be influenced by many variables (topic, speech corpus size etc.) and we do not take the view that they can deliver reasonably reliable statistical results. We believe it would be very useful to have such information filtering generally known words out from the keywords results, but since there are no reliable data on this matter, we will neither provide any statistical data of an average active vocabulary of an English native/non-native speaker in paper, nor use it as a theoretical basis for manipulating the word frequency analysis by introducing stop lists based on average active vocabulary.

There is another very important aspect by mapping the vocabulary size. Many studies of native speakers' vocabulary growth and size treat all equal in value to the learner. However, frequency-based research has shown very strikingly that this is not the case and some words are indeed much more useful than others.

Words can be divided according to Nation (2001) into following 4 categories based on their usage: a) high frequency words, b) academic words, c) technical words, d) low-frequency words.

The first category includes high frequency words identified for general English. This category mainly consist of *function words* convening grammatical information (Van Gelderen, 2005) and some common content words that convey meaning (e.g. production, government). Corpus research suggests that 80% of the running words in the text are high

frequency words. In order to avoid common and function words by the analysis of the Europarl corpus, we will include 3000 most frequently used words in a stop list which filters them out from our keywords results[9].

The second category includes academic words that are widely represented in academic texts, but in general English, they tend to be represented no more than at 9% of all running words. Academic words are sometimes called *sub-technical vocabulary* which does not contain technical words, but rather formal vocabulary creating the structure of the academic language (Nation, 2001b).

The third category represents technical words. These are related to the topic and subject area of the text. Technical words are reasonable common in technical texts, but not common elsewhere where they cover about 5% of the running words in the text. Yet, technical words typically carry a significant portion of meaning of the text and their active knowledge is necessary for effective communication in language for special purposes (LSP). This category is in the interest of this thesis and we believe that our quantitative lexical analysis identifies a significant portion of them, therefore, we will describe this category in more detail.

Technical words are usually associated with specialized vocabulary. Specialized vocabulary can be prepared by systematically restricting the range of topics or language uses investigated by creating a specialized corpus (e.g. the Europarl corpus) and then counting word frequencies and comparing them with more general corpora. The process is also called the *keyword analysis*, which will be described more in detail in chapter *Main features and methods used in corpus analysis software*. Notion (2001) points out that learning the technical vocabulary is particularly important for general understanding of LSP and significantly more effective for LSP acquisition. For instance, adding the AWL to the GSL's 2000 high frequency words changes the coverage of academic texts from 78% to 86%, which means that one word in every ten will be unknown. This is a very significant change. However, if the learner had moved on to the third 1000 most frequent words instead of learning another 10% of the academic coverage, he or she would have gained only 4% extra coverage. Thus, learning LSP vocabulary helps significantly in

[9] The stop list is explained more in detail in chapter *Main features and methods used in corpus analysis software.*

understanding and production of LSP texts. As a result, teachers and learners should treat specialized vocabulary like high frequency vocabulary during the learning process and pay considerable attention to acquiring competence for using both types of vocabulary (Notion, 2001).

The fourth category includes low-frequency words that cannot be put in the academic or technical words category. They also carry a significant portion of meaning, but since they do not appear frequently, it is more reasonable to infer their meaning from the context by accidental learning than focus on learning them methodically.

Vocabulary acquisition. Our thesis deals with expanding the active vocabulary in language for special purposes (LSP), more particularly the language of EU parliamentary settings. As already mentioned, the active vocabulary is smaller and harder to acquire than the passive one. Since we are aware of these two drawbacks of active vocabulary and at the same time its importance for speech production, from now one we will further focus on explaining active vocabulary acquisition. There would be many topics to concentrate on, such as long term vs. short-term memory, remembering, retrieval process, forgetting, associations, learning strategies, mnemonics etc.), but the aim of this thesis is only to produce a relevant frequency-based English LSP wordlist and to support its usefulness with theoretical knowledge, therefore, above mentioned topics will not be discussed here. We will rather focus on the concepts of accidental learning, and implicit vocabulary learning

To explain accidental learning, we will briefly touch on the subject of vocabulary acquisition. There are two main hypotheses for vocabulary acquisition: an implicit vocabulary learning hypothesis and an explicit vocabulary learning hypothesis.

An implicit vocabulary learning holds that the meaning of new words is acquired at the unconscious level as a result of abstraction from repeated exposure in a range of contexts. This knowledge is also represented subconsciously in the brain, so called "tacit knowledge", as proposed by Chomsky (Krashen, 1989). This so called incidental learning is typical for children acquiring their native language vocabulary, but it applies to adults as well. Research studies directly demonstrate that readers acquire vocabulary from text without paying attention to the learning process. On the other hand, the teaching method that applies reading for vocabulary acquisition and encourages readers to read large amounts of text, yet easy material is called extensive reading (Day & Bamford, 2002). In

general, people who read more know more vocabulary and this relationship appears to be casual in that it holds even when intelligence is controlled (Ellis, 1995).

An explicit vocabulary learning states that the acquisition of new words can be strongly facilitated by the use of a range of metacognitive strategies that include 3 steps: firstly, noticing that the word is unfamiliar, secondly, making attempts to infer the word from context or dictionaries, and thirdly, making attempts to consolidate this new understanding by repetition and associational learning strategies (e.g. mnemonics) (Krashen, 1989). There are great many strategies for optimal vocabulary acquisition that can be found in relevant literature

Accidental learning can also happen in explicit vocabulary learning based on which words a learner chooses to learn. Therefore, we can differentiate between learning HFWL and learning accidentally by encountering an unknown word and getting familiar with its meaning by one way or another (depending on the chosen learning technique) Some scholars claim that learning new words through guessing their meaning in the context is actually the most important strategy of all in learning vocabulary (Nation, 2001b), but there are also many staunch advocates who point out to learning vocabulary lists as the most effective method, for instance:

> The suggestion that learners should directly learn vocabulary from cards, to a large degree out of context, may be seen by some teachers as a step back to outdated methods of learning and not in agreement with a communicative approach to language learning. This may be so, but the research evidence supporting the use of such an approach as one part of a vocabulary learning program is strong (Nation & Waring, 1997, para. 20).

There is a very large number of studies showing the effectiveness of learning vocabulary deliberately from word lists or word cards in terms of amount and speed of learning. This type of learning is often put in comparison with accidental learning or with learning from the context. Here it is important to note that research on learning from context shows that this type of learning does occur but that it requires learners to engage in large amounts of reading and listening for the learning is slow, small, and cumulative. Given the same amount of time, deliberate learning of decontextualized vocabulary has

been found to always surpass vocabulary learning in context even though some studies speak against[10].

Be it as it may, this should not put learning from context in question, just on the contrary. As Nation and Waring further point out, learning from context is by far the most important vocabulary learning strategy. However, for fast vocabulary expansion, it is not sufficient by itself. Therefore, it is strongly advisable to combine both these techniques to achieve the best results (Nations & Waring, 1997; Berns, 2010). Interestingly, direct learning of vocabulary from cards has also proven to have a slightly higher benefit than learning through focused grammar instructions. Nations & Waring conclude that deliberate learning from word cards is a learning tool for use suitable at any level of vocabulary proficiency and that frequency information provides a rational basis for making sure that learners get the best return for their vocabulary learning effort (1997).

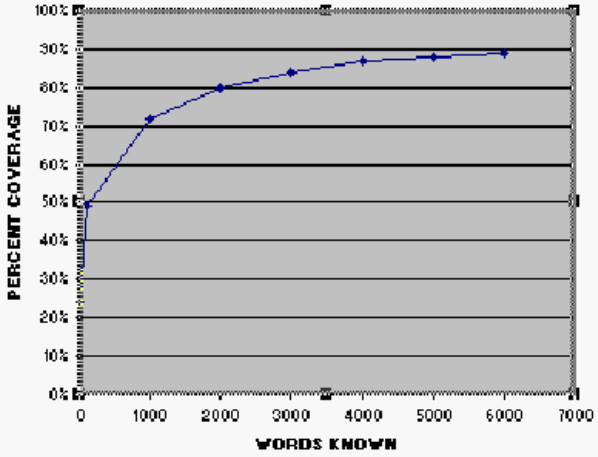
However, there are still some objections worth mentioning: Learning HFWL seems to be an effective systematic approach on the first glance. The learner can learn 2000 most frequent words that account for about 80% of the language, then he or she can move to 3000 most frequent words accounting for 84% etc. The practical problem is that learning vocabulary, similarly as mastering other domains into excellence, follows the exponential curve. In other words, learning 3000 words would not bring the same time/effort vs. effect ratio as learning 2000 words (Cobb, n.d.). Then the question arises how to achieve the ultimate goal of acquiring coverage of 95% that is according to studies reasonable enough for understanding a text without difficulties (Fry & Kress, 2006).

Figure 2. Set of wordlists ranked by their frequency and the percentage of coverage in the total population of BNC Corpus. Cobb. n.d.. 6k.gif. Retrieved from http://www.lextutor.ca/research/6k.gif.

[10] See Ying (2010)

Table 1

Different words	Percent of word tokens in average text
86,741	100 %
43,831	99.0
6,000	89.9
5,000	88.6
4,000	86.7
3,000	84
2,000	**79.7**
1,000	72.0
10	23.7



A good way seems to be learning HFWL and then continuing to expand the vocabulary by accidental learning in which the learner chooses what is relevant for his or her vocabulary. However, there is a risk that one learns words which are rare, irrelevant for usual conversation, or they will not repeat often in the other texts to be remembered effectively.

Cobb (n.d.) proposes a good solution to tackle this problem in combining general HFWL with HFWL obtained from a domain of the learner's particular interest. This is the idea behind academic word lists that are said to greatly improve reading proficiency in academic texts. Nation and Waring note (1997) that there are certain words that are not necessarily frequent in the language at large, but are very frequent in academic texts. These words have been later brought together as the AWL mentioned above. Cobb suggests that by combining the 2000 most frequent words with the AWL a total of 2570 words, it is possible to cover 90% of all words in the special domain of academic texts (Cobb, n.d.).

Taking this assumption to be correct, we have decided to draw an analogy between academic texts and language for special purposes, in our case the parliamentary language. We take the view that combining 3000 most common words in English (used as a filter or stop list in the corpus analysis) together with 2000 keywords identified as unusually frequent in comparison to other texts we can achieve a similar efficiency for acquiring the parliamentary language as in case of combining general HFWL with AWL. We also expect that initial learning of the glossary will lead to more efficient vocabulary retrieval than one would achieve with the same time by accidental learning.

Limitation of HFWL One very important issue with HFWL is the fact that in counting they vastly ignore words with multiple meanings. A general problem with word frequency approaches is that more common words are likely to have multiple meanings. For this reason, where technical vocabulary is also high frequency vocabulary, teachers should help learners to see the connections and differences between the high frequency meanings and the technical uses (Nation, 2001a). It is also important to mention that such uncommon meanings, especially in LSP (e.g. house as dwelling vs. House of Commons) may pose a threat in corpus analysis because they might be filtered out inadvertently with selected stop lists. We are aware of this limitation. The proposed solution is to go through the words in the stop list manually, to select words with expected multiple meaning and to conduct qualitative corpus analysis with concordance search to find collocations suggesting special usage in the political parliamentary language.

Conclusion and suggestion for correct vocabulary acquisition. To sum up the importance of vocabulary acquisition from the teaching perspective, we would also like to emphasize that communicative teaching strategies, as opposed to formal teaching of grammar, agree that the growth of vocabulary regardless the technique can only enhance the natural acquisition of language competence. From the practical point of view, it is generally expected from students to learn lists of vocabulary whether they are directly encouraged to do so or not. Without too much effort, the student can learn well over 30 words per hour by studying lists of vocabulary (Ellis, 1995). Stahl and Fairbanks (1986) performed an analysis of approximately one hundred independent studies comparing the effectiveness of vocabulary instruction methods and found out following:

a) The vocabulary instruction is a useful adjunct to the natural learning from context;
b) The methods which produced highest effects on comprehension and vocabulary measures were those involving both definitional and contextual information about each word;
c) Several exposures were more beneficial to drill and practice methods;
d) Mnemonic techniques produced consistently strong effects;
e) Methods which provided a breadth of knowledge about each to be learned word from more contexts had a particularly good effect on later understanding of texts in which those words were used.

From the gathered knowledge about teaching vocabulary we stay positive that pedagogical implication in general are in favor of creating frequency based vocabulary lists for LSP. It has shown that HFWLs are a very effective method for vocabulary acquisition in LSP. In order to prove its importance for ESL speakers not only from the pedagogical perspective, we decided to look deeper in the area of cognitive sciences. In the next chapter, we will put the case for learning from HFWL that should help in speech production output from the speech science perspective.

Speech Science

In the previous chapter we introduced the importance of frequency-based glossaries from the didactic perspective of second language acquisition. In order to back up authors' claims, we decided to find supportive knowledge in the speech and cognitive science whether improved vocabulary reflects in better lexical competence and consequently in more effective speech production.

In order to explain the benefits of glossaries for speech quality improvement, we should firstly look at the speech production process itself which encompasses both speech production in a mother tongue (hereafter referred as L1) and in a foreign language (referred to as L2)

Speech production theory. A description of the native speakers' language production system suggests that producing or comprehending speech is indeed a complex task involving many sub stages (Crookes, 1991).

All speech production researchers agree that language production has four important components: a) *conceptualization*, i.e., planning what one wants to say, b) *formulation*, which includes the grammatical, lexical, and phonological encoding of the message, c) *articulation*, in other words, the production of speech sounds, and d) *self-monitoring*, which involves checking the correctness and appropriateness of the produced output (Crookes, 1991)..

In L1 production, planning the message requires attention, whereas formulation and articulation are automatic. Therefore processing mechanism can work in parallel, which makes L1 speech generally smooth and fast.

Researchers also agree that one of the basic mechanisms involved in speech production is *activation spreading*, a term adapted from brain research based on findings

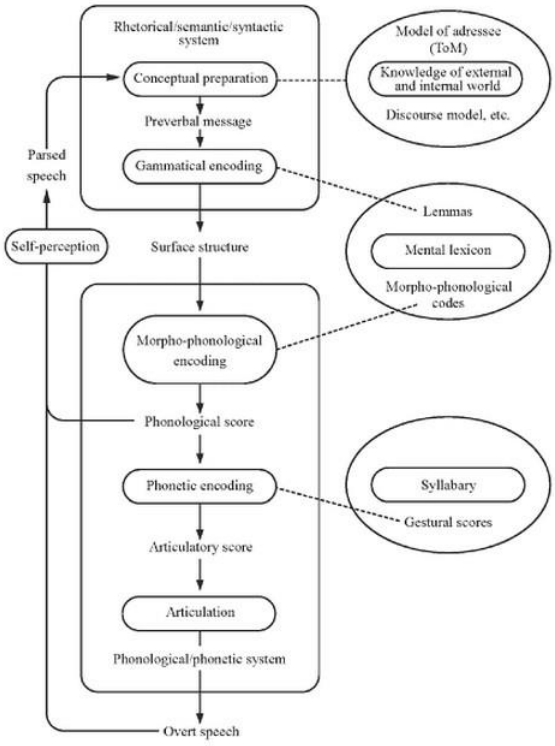
in neurological studies. The speech-processing system is assumed to consist of hierarchical levels (conceptualization, formulation, articulation) among which information is transmitted in terms of activation spreading, exchanges of simple signals among brain cells in the neural network. Information such as the lexicon and conceptual memory store, spread from one item to related items and decisions are made on the basis of the activation level at the so called nodes that represent various units such as concepts, word forms, phonemes etc. The stronger the connection, the lesser the cognitive load and higher degree of automatization (Kormos, 2014).

Kormos gives an exhaustive account to speech production and second language acquisition research in her recently published book (2014), stating that the most widely used theoretical framework in language production research for both L1 and L2 is Levelt's *blueprint of the speaker*. We will use this model for further elaboration on the speech process.

Levelt's blueprint of the speaker incorporates the fourth components of language production (conceptualization, formulation, articulation, self-monitoring). It argues that speech production is modular i.e. it can be described through the functioning of a number of processing components that are relatively autonomous, consecutive (one cannot start before it receives input from the other preceding it), and simultaneous (they are performed more or less parallel). The model distinguishes two principal components: the rhetorical/semantic/syntactic system and the phonological/phonetic system and supposes the existence of three knowledge stores: the mental lexicon, the syllabary (containing gestural scores which are chunks of automatized movements for producing the syllable of a given language), and the store containing the speaker's knowledge of the external and internal world (Levelt, 1999; Kormos, 2014).

The model is illustrated below:

Figure 3. Levelt's (1999a) blueprint of the speaker. Copyright 1999 by Oxford University Press.



Levelt's blueprint of the speaker can be also used for describing L2 speech production, but one must bear in mind that L2 speakers are faced with some additional difficulties compared to L1. We can categorize them into four main areas.

Firstly, on a semantic level, some words in L1 may have no equivalent in L2 and must be described or vice versa. It can be illustrated for instance on the English word *privacy* that has no direct equivalent in Russian.

Secondly, on the lexical level, the majority of studies suggests that the conceptual specification contained in the preverbal plan active both L1 and L2 items in the mental lexicon, making words in both languages compete as candidates for lexical encoding. Additionally, the lexical bank in L2 might be limited or not automatized. The degree of these limitations correlate with the language proficiency. If the speech process is not fully automated, this might result in a much higher cognitive load than in L1. To put it in laic words, in a foreign language we say what we can, but in our mother tongue we can express what we want. The lexical level is particularly important for our thesis because its aim is to introduce a glossary that should reduce the cognitive load of advanced ESL speakers in EU parliamentary language settings by decreasing the cognitive load at the lexical level.

Thirdly, on the syntactic level, it is reasonable to expect that big differences in syntactic structure between L1 and L2 (e.g. transitivity of verbs/gender system) may result in additional cognitive load if the syntactic structures in L2 are not fully automatized during the learning process.

Fourthly, on the phonological level, articulation of words in L2 is also believed to require more attention, but this is rather the case of ESL beginners who rely heavily on L1 syllable programs than of advanced L2 speakers who adopt and master the syllable program from L2 (Kormos, 2014).

Additionally, L2 speakers have to face problems in communication that are mainly a) resource deficits, b) processing time pressure, c) perceived deficiencies in one's own language output, d) issues related to speech comprehension and e) lexical, phonological or semantical interferences from L1.

L2 speakers have to make a conscious effort to overcome these problems. These efforts are called communication strategies and tend to improve with a snowball effect by mastering the second language.

Automatization of language processes in special regard to bilingual lexicon. In general, the bilingual lexicon is built through cross-linguistic activation and encompasses processing differences of phonologic and semantic information and effects of specific word types (cognate status and frequency). Studies have shown that processing is the result of internal factors (such as proficiency), situational factors (L2 status), mode, and recency), linguistic features (such as psychotypology) and learning context (such as immersion) (Serrander, 2011). We would like to focus now on specific areas relevant to the use of glossaries in language learning that mainly concerns the area of automatization of the bilingual lexicon.

An interesting elaboration on automatization in the bilingual lexicons offers Serrander:

> It is frequently emphasized that repetition and practice are essential in L2 learning. Learning L2 words means building mental representations for them in the bilingual mental lexicon. Practicing these words means practicing the process of mapping the L2 segments onto their corresponding mental representations. In this sense, it is the L2 processing skills that improve as a result of practice and

repetition. The improvement referred to here is called automaticity (Serrander, 2011, p. 19).

Speech science and research in second language acquisition provide an interesting insight into the correlation between frequently used words and their effective usage, also claiming that automatization plays a central role in forming the mental lexicon. Numerous studies have proven the so called *word frequency effect* that shows in improved performance in high frequency words compared to low frequency words in almost any language task (Serrander, 2011). This implies that automatizing certain words that are expected to appear in certain contexts (specific vocabulary) can manipulate mental lexicon and "instill" these words in active memory. In other words, the speaker does not need to access his or her long-term declarative memory but can make use of learned automatisms and use the saved cognitive energy for tackling other challenges. This should consequently result in improved speed and accuracy or more generally in a better lexical proficiency (DeKeyser, 1997).

As already mentioned, L2 speakers are at a considerable disadvantage to L1 speakers who perform many subtasks automatically. Therefore, for L2 learners it is important to acquire this automatization in these processes through intentional repetition or long exposure to the right language stimuli. Levelt proposes that the acquisition of skill in speech production, as with any other skill, "consists essentially of automation of low level plans or units of activity" (1978, p. 57, as cited in Crooke, 1991, p. 118). The decision of which section of the language will be practiced depends to a large extent on the learner and his or her form of planning. Pre-planning of an utterance provides a way for less well automatized section of the system to be used, and intentionally improved. This decision-making is important since what is not used will not get automatized. The formal practice is believed to transform explicit knowledge of L2 into implicit knowledge that is important for better language production (Crooke, 1991).

However, the improvement of speech production lays not only in intentional learning, but also in monitoring of produced speech. If the learner monitors his or her own speech output, an utterance produced successfully on one occasion may be noted and reused thereafter with increasing automacity. This has pedagogical implication in using the glossary not only for learning the most frequent words, but also for using them in

context, for instance, in a learning activity during which students try to put them into a more complex sentence by using some word clues or sentence fragments.

To sum up, when some degree of automaticity is obtained, a cognitive activity can be performed without attentional control or at least with less effort and improved effectivity (involvement of attention/awareness) (Young & Stanton 2002; Serrander, 2011). Such automatic skills are consequently said to be unconscious, effortless, smooth and fluent which is especially important when dealing simultaneously with other subtask by L2 speech processing, as mentioned above.

The importance of saving cognitive resources for other subtasks can be implied from the Levelt's model and better illustrated in the next section on the attention model by Kahneman that describes allocation of attention to various cognitive activities.

Concept of shared attention. Originally, it was believed that the human mind cannot execute more than one complex cognitive task simultaneously. Complex tasks are in cognitive sciences described as controlled processes at conscious level limited in capacity, requiring attentional resources and effort and as having capability to be used in changing circumstances at will (e.g. learning), as opposed to unconscious automatic processes with no capacity limitation, no cognitive effort required and certain difficulty with modifying once learned (e.g. walking, breathing) (Quinlan, 2008).

The generally accepted view of a fixed, single channel of limited capacity for controlled processes was challenged by psychologists in the 1970s (e.g. Gerver) who postulated replacing the notion by that of a "fixed-capacity central processor, whose activity could be distributed over several tasks within the limits of the total processing capacity available" (Pöchhacker, 2004, p. 116). This notion gave a birth to attention sharing models that, as we will see later, were also modified and implemented in theories of interpreting studies.

Still, there are some vociferous opponents among neurological scientists to the concept of a multitasking of the human brain. They claim that the human brain only appears to be capable of multitasking. The distraction of more activities performed simultaneously is claimed to thwart the cognitive abilities to perform multiple tasks

effectively[11] (see Medina, 2008). However, we do not need to avert this claim because it still supports the notion that the human brain is indeed capable of multitasking, albeit, with a much reduced effect compared to the situations in which the attention was focused entirely on one complex activity.

Nonetheless, we will assume that the leading position in neuroscience endorsed by Gerver is correct and the human brain is capable of multitasking.

A very suitable model for explaining this notion is Kahneman's model of attention and effort[12]. He assumes that there is a single pool of attentional resources[13] which can be divided intentionally among multiple tasks with great flexibility. Accordingly, attention can be focused on one particular activity, or can be divided between multiple activities.

With increasing difficulty of the tasks, the attention demand rises. If two complex tasks compete for common resources which in sum exceeds the sources, performing multiple tasks at the same time becomes difficult. Attention is here understood rather as a limited supply of resources, whereas effort caused by arousal may vary slightly based on person's motivation. Too much motivation can result in stress causing performance decline (Yerkes-Dodson law), but that is rather the area of stress research which we do not intend to discuss here (Kahneman, 1973; Style, 2006).

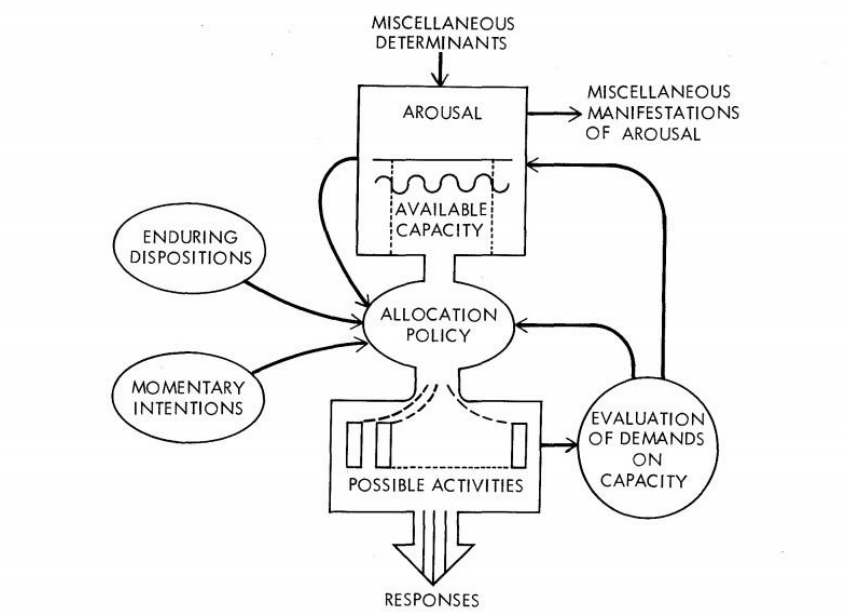
Figure 4. Kahneman's model of attention and effort. Kahneman. 1973. *A capacity model for attention*. Retrieved from

[11] Medina cites among many scientific studies the research findings that the danger of driving and talking on the cellphone cannot be reduced by using hands-free devices because the driver is distracted by speaking and driving, rather than reduced hands movements when operating a cellphone.

[12] Another well-accepted model is Multiple Resource Theory (see Basil, 2012), but we deemed Kahneman's model as more suitable for establishing the link between cognition, speech production and interpreting.

[13] It could be compared to a computer memory (random access memory, RAM) that is distributed for various processes in the operation system.

http://www.princeton.edu/~kahneman/docs/attention_and_effort/attention_lo_quality.pdf.



Kahneman's model is rather broad in generalization and not directly focused on processing language, but it offers a solid theoretical basis for further work in the speech science.

There is another model based on the attention sharing principle worth mentioning, but it includes also other processes above and beyond language production. This model was created to explain attention sharing during simultaneous and consecutive interpreting by Gile. The Gile's effort model was constructed to account for so called "problem triggers" that cause difficulties in interpreting (compound technical terms, proper names, numbers). He hypothesized that interpreters usually work at the limit of their processing capacity and these occasional problem triggers may lead to inability to interpret the whole segment correctly, which results a sharp temporary decline in the output quality of the interpreter (Gile, 1999). The effort model consists of operational constraint and is not intended to describe the architectural process in terms of a particular mental structure and information-processing flow. The model for simultaneous interpreting originally included three main efforts: listening and analysis (L), speech production (P), and short-term memory operations (M). It is important to note that they have non-automatic components, therefore all three require attentional resources. The three efforts were later supplemented by the coordination effort (C) that helps allocating cognitive resources for these complex

tasks at will. The total effort, not necessarily a purely arithmetic sum of described efforts, should be lower than the total cognitive capacity (TotC) if the interpreter is to produce adequate high-quality interpreting. Should the sum of all efforts exceed the TotC, for instance by running into problem triggers by interpreting, the output quality will decline. This competition hypothesis is generally accepted by interpreting studies scholars intuitively and is explicit in many anecdotal accounts of difficulties that interpreters encounters. It has also not generated any criticism when presented to cognitive scientists (Gile, 1999).

Gile's effort model (*c* stands for capacity for the respective effort) (Gile, 1999):

$$TotC = c(L) + c(P) + c(M) + c(C)$$

It is beyond doubt that consecutive and simultaneous interpreting require more subtasks than speech production either in L1 or in L2, but this theory can also give us an important insight into the concept of attention sharing linked to speech production.

Theoretical implications. Proceeding on the assumption that language processing is a difficult task requiring simultaneous attention to several subtasks (Levelt's model) and that the human brain is capable of allocating attention to multiple processes simultaneously (Kahneman's model), we postulate a hypothesis for L2 speech production: If the sum of efforts on cognitive language processes partially exceeds the cognitive capacity, reducing cognitive load at one subtask can help decrease the overall cognitive load and consequently lead to speech quality improvement (fluentness, speech, flawlessness, clarity, structure etc.).

Accordingly, we propose a L2 speech production model based on Levelt's, Kahneman's and Gile's work that integrates both the speech production theory and the theory of shared cognitive resources.

In this model, the sum[14] of total cognitive capacity (TotC), or the maximum cognitive capacity (Mcrc) alternatively, accounts for efforts allocated to individual

[14] Also not purely in the arithmetical sense.

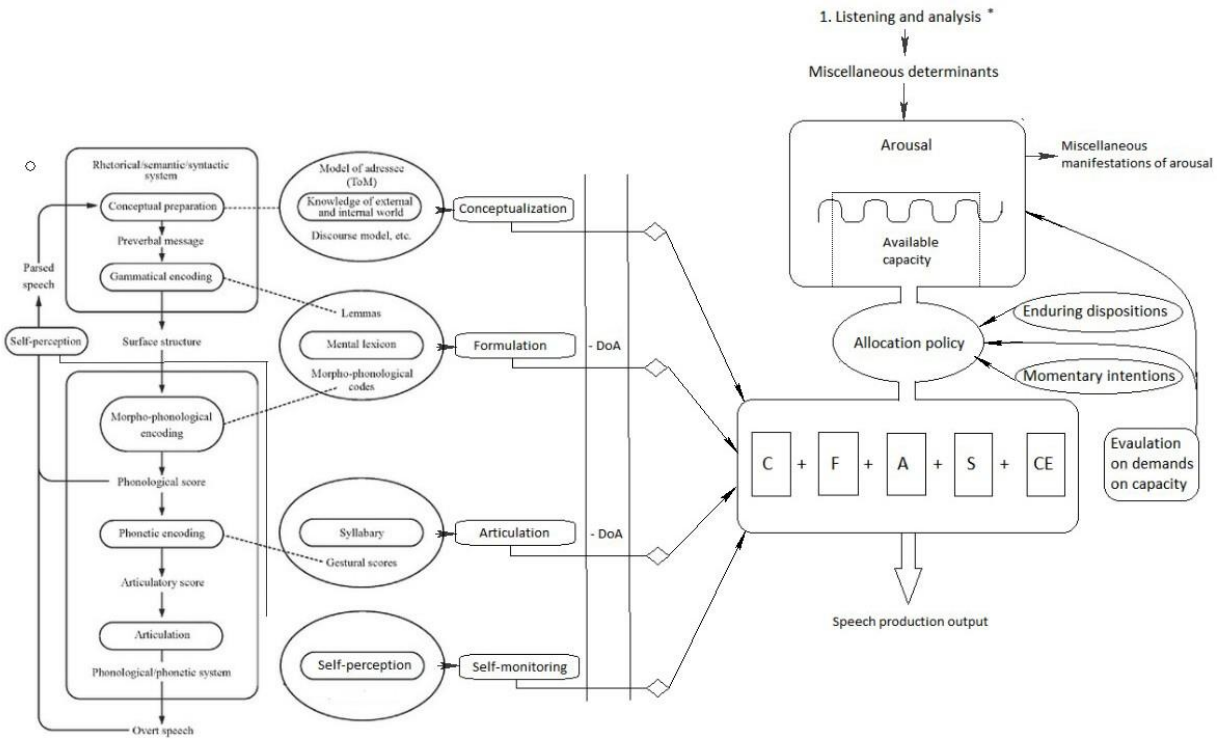
subtasks in L2 speech production (conceptualization /C/, formulation /F/, articulation /A/, self-monitoring /S/) together with coordination effort (CE), and the deduction of the degree of automatization (DoA) for respective subtasks.

The model explained in formula:

$$TotC = C + (F\text{-}DoA) + (A\text{-}DoA) + S + CE$$
$$Mcrc = C + (F\text{-}DoA) + (A\text{-}DoA) + S + CE$$

It is important to distinguish between TotC and Mcrc. Total available cognitive capacity (TotC) represents the total capacity allocated during the arousal which might slightly change based on speaker's intention, motivation, fatigue, and other external factors. On the other hand, the maximum cognitive capacity (Mcrc) stands for cognitive limits of the speakers that cannot be exceeded. The TotC is depicted in the model as the area inside dashed lines and the Mcrc as the outside spectrum limited by outer full lines.

Figure 5. A synthetized model for oral speech production linked to the theory of shared attention.



The degree of automatization is a difficult parameter to measure, but it should be counted on at several levels of speech production.

There can be 3 possible outcomes after reducing cognitive load at one speech production subtask (in our case at the lexical level that is included in the formulation stage).

a) The total cognitive demand (TcD) exceeds the total capacity TotC or the maximum cognitive resource capacity (Mcrc) even after reducing cognitive load at one subtask. Consequently, reducing cognitive load did not help avoid language difficulties, but saved resources could be allocated to other subtasks which resulted in partial improvement of language output.
b) TcD would have been higher than TotC or Mcrc if reducing cognitive load at one subtask had not taken place. Consequently, reducing cognitive load results in successful coping with TcD and improved language output (original hypothesis).
c) TcD is lower than Mcrc and/or lower than TotC. Consequently, reducing cognitive load at one subtask has no result in language output, but might help saving mental energy and avoid tiredness to some extent.

Naturally, one cannot expect that learning and automatizing certain vocabulary will leave him or her free from other unpleasant surprises on the lexical level. This model only proposes that acquiring specific LSP vocabulary might be effective in reducing cognitive load and consequently in improving the language output in the same specific context (discourse). The question which vocabulary to acquire for which topic is a subject for further research.

The aim of this thesis is neither to validate nor test the proposed model or hypothesis, but to prepare a frequency based glossary using knowledge from relevant scientific disciplines. The literature review only resulted in creating this theory. We do not make a claim for its completeness of the proposed theory, but use it as theoretical framework supporting and explaining the usefulness of creating the frequency-based glossary of the Europarl corpus.

Conclusion

In the first part of this chapter, we focused on answering the question *why create a HFWL for SPL acquisition based on a quantitative corpus-based lexical analysis.* We have firstly described the history and purpose of HFWL and how they can benefit vocabulary acquisition, and secondly discussed various speech models that explain the speech production. Combining results of previous interdisciplinary research in speech science and interpreting studies, we have introduced the hypothesis that speech production quality may improve if the cognitive load is decreased at lexical level through vocabulary automatization.

In the next part, we will present practical knowledge that could be applied for compilation of a functional glossary for LSP by consulting other disciplines. These disciplines range from terminology, lexicology, language for special purposes (LSP) with a special emphasis on politolinguistics, and computer assisted language learning (CALL) together with computer-assisted vocabulary learning (CAVL). In other words, whereas the first theoretical part of this paper was dealing with the existential purpose of a glossary based on the Europarl corpus, the second theoretical part should provide knowledge for an effective and purposeful glossary compilation from abovementioned disciplines.

Theoretical Framework for Glossary Compilation: Other Disciplines and Reviewed Knowledge

Introduction

In the previous chapter, we have identified that creating a glossary from a HFWL could be a useful learning tool. In this chapter, we will seek to explain how the selection of words should be organized, performed and presented. To accomplish this goal, we consulted following scientific disciplines: lexicology, terminology, language for special purposes (LSP), politolinguistics and computer assisted vocabulary learning (CAVL).

Before reading into these disciplines, we had to get an overview about freely available glossaries on the EU that could serve as a springboard for our further work. We have found out that there are great many glossaries on topics discussed at the European political level either from the EU institutions or from third parties available on internet. Their quality, quantity and specialization may vary, as well as the offer in various language combination. Since the Web is like a perpetually expansive universe, there would be no point in attempts to map the publicly available glossaries.

After a short review of web content, we have learned that there are great many language resources available online. The only peril might be looking for the most relevant one, like looking for a needle in a haystack. We have come to two conclusions:

Firstly, to our knowledge, there is no EU glossary based on most frequent words. All glossaries tend to compiled in topic words as a result of manual selection.

Secondly, these glossaries mainly include highly specialized vocabulary. This is vastly important for getting versed in the respective topic but they already assume perfect knowledge of semi-technical words typical for the discourse.

It is reasonable not to include semi-technical vocabulary in specialized glossaries since they create only a basis for LSP communication. But on the other hand, the learner without tacit knowledge of the semi-technical vocabulary must get familiar by them through extensive reading – learning the highly technical vocabulary would be useless if the learner could not express the words in a style particular for the political discourse. As we mentioned earlier in chapter about *vocabulary acquisition*, extensive reading takes a long time and it less effective as learning a HFWLs like the AWL.

Therefore, we have tried to look for the way to include the semi-technical vocabulary and to back our hypothesis with knowledge from politolinguistics. This will be more covered in detail in chapter about politolinguistics and the language for special purposes (LSP) which gives a more general overview about working with specialized language.

Next, the chapter gives an exhaustive account of lexicology (homonymy and synonymy in respect to their subtle differences, etymology in regard to internationalisms, neologisms, false friends, lexicometry) and terminology (frameworks for glossary compilation) for its significant contribution to the research subject.

Finally, there was a big question how to present the glossaries in a functional and appealing way. To achieve this, we consulted computer assisted vocabulary learning (CAVL).

Language for special purposes

Language for special purposes, also known as LSP is in simple terms "the language that is used to discuss specialized fields of knowledge" (Bowker & Pearson, 2002, p. 25). This field of knowledge may include everything from professional activities to hobbies provided that they treat a restricted subject. LSP can be also explained as n opposite to LGP or language for general purposes that is used to talk about general things in a variety of common everyday situations. A more sophisticated definition of LSP is offered by Picht & Draskau:

> a formalized and codified variety of language, used for special purposes with the function of communicating information of a specialist nature at any level in the most economic, precise and unambiguous terms possible (Picht & Draskau, 1985, p. 3).

LSP can be also understood as a sublanguage with its specific morphology, syntax and lexis. A sublanguage in respect to LSP is defined as:

> the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles of a technical specialty or science subfield), in which the authors of the documents share a common vocabulary and common habits of word usage" (Hirschman & Sager, 1982, p. 28).

While LGP can be found in almost any native speaker of the respective language at a great proficiency level, this is not the case for LSP that is used for communication in specialized fields. LSP speakers namely do not only use a specialized terminology, but they also have comprehensive knowledge about the subject. Therefore, knowing the terminology alone might not necessarily lead to better understanding.

There is some degree of overlap between LGP and LSP. It is firstly caused by the necessity to use some general words that create the fundamental structure for expressing ideas in a sentence (e.g. function words), and secondly by using LGP words in LSP with another special meaning. Similarly, some specialized words find their way to the general vocabulary as they become the topic in the general discourse (Bowker & Pearson, 2002).

This could lead us to the assumption that LSP is simply a LGP with some specialized vocabulary. However, this would not be exactly correct. Pearson & Bowke point out that LSP may have unique ways of combining terms and/or arranging of information that differ from LGP. They illustrate it on a lexical example with chemistry:

> … you may need to know what verbs are generally used with the noun 'experiment'. Based on your knowledge of LGP, you might assume that the verb 'do' can be used (e.g. 'to do an experiment'), but a search in a specialized corpus demonstrates that in the LSP of chemistry, experiments are typically 'conducted' or 'carried out'. This type of search also reveals another interesting feature of the LSP of chemistry - it shows that the use of passive constructions is very common (e.g. 'the experiment was conducted' rather than 'X conducted the experiment') (Pearson & Bowke, 2002, p. 27).

As we could see, LSP can differ from LGP in a number of ways which should not be overlooked. Specialized vocabulary, stylistic features and collocations contribute to the formation of LSP and speakers must abide these commonalities if they want to communicate in LSP effectively.

Since LSP is mainly used for communication in specialized fields, it is not surprising that research in LSP is more connected to research associated with business schools, translation institutes, informational retrieval, machine translation and natural language processing rather than to pure linguistic departments (Hjřrland & Nicolaisen, 2005).

The communication in LSP can take place in 3 different forms that take into account the expertise of communication partners. Firstly, we can speak about expert to expert communication in which experts share a common background, understand the meaning of specific terms and phrases and do not need to provide their explanations. The second type is the communication between experts and semi-experts who are students or experts from related fields. In this case, experts will probably use specialized terminology, but will accompany it with explanations where necessary. The third type represents the communication between experts and non-experts. In this type, the expert will use fewer terms and try to provide simplified explanations in LGP (Bowker & Pearson, 2002)[15]. In respect to the target audience of the MEPs' speeches at the European Parliament, we suppose that the first and the second type of communication. This will be explained more in detail in the next chapter about politolinguistics.

Politolinguistics. The political discourse used in parliamentary speeches, among other direct or indirect manifestations of human communication, follows certain linguistic patterns. These patterns, mostly distinguishable in lexical and semantic differences, represent a certain language variety, which is further studied in political linguistics or politolinguistics. To understand the context of verbatim records of parliamentary speeches in the European Parliament, we considered as necessary to consult this area of study.

---

[15] I have read an interesting example about the correct usage of language in education. The Czech mathematician Vít Hejny introduced in the 19th century a special approach to teaching elementary school mathematics through experiments and their own experience rather than through repetitive learning and explanations. The fundamental principle of his approach is to present the new information as a part of a familiar pattern that the child easily visualizes. If children are asked to solve the following equation: 7+x=10, they would probably not know what to do without prior knowledge about equations. However, if children were asked "find the number that should replace the asterisk in the arithmetical problem 7+*=10", they would easily come to the correct solution. This demonstrates clearly that the language matters and one should be aware of his or her usage in respect to the needs and knowledge of the communication partner.

Politolinguistics or political linguistics studies is a multidisciplinary research area that draws empirical input from neighboring domains such as linguistic pragmatics, text linguistics, discourse analysis, corpus linguistics, translation and literary studies, politology, social psychology, sociology, anthropology, philosophy and rhetorics (PL, n.d.[1]). The multitude of research perspectives offered by various disciplines is enormous, as well as the complexity of research areas on which politolinguistics can focus. Therefore, a generally accepting definition for political linguistics is lacking, similarly as a clear establishment of boundaries between politolinguistics and other scientific disciplines.

The term politolinguistics was first introduced in 1996 by Armin Burkhardt (in German *Politolinguistik*) as a description for the area of linguistics specializing in critical analysis of political language (Burkhardt, 1996).

In general, the term politolinguistics, political language, political linguistics or political linguistics studies embodies a multitude of research disciplines that deal with overlapping of the research areas of politics and language. These disciplines, each with its own perspective and examining methods, require shared interest in the particular research subject in order to provide a synthetic analysis of subject concerned. This can include for instance the corpus analysis of the word frequency (personal pronouns like I, me and my and other political words) in Obama's speeches compared to his predecessors (RLG, 2014), studying Putin's body language (AFP, 2014) lexicometrical study of parliamentary or unparliamentary language across time and culture (e.g. Thompson, 2011; Smith, 2015), examining the political jargon (Myers, 2015) or lexical/pragmatic analysis of Merkel's speeches (Strauß, 2010). Broadly speaking only from the perspective of linguistics, political discourse can be examined on macro-meso-micro level which implies a considerable number of various approaches to the same subject. To briefly mention a few of these areas, texts can be analyzed on a syntactic level (number of words, number and complexity of nominal phrases, sentences, accent and focus), semantic level (positive/negative word connotations, addressing thorny issues) and pragmatic level (rhetorical expressions, degree of persuasion) (Klein, 2014).

The strengths of politolinguistics lie in the ability to provide more adequate analytical framework than a monodisciplinary approach of political sciences and linguistics (Cedroni, 2010), yet, to our mind, the weakness of political linguistics stems

from the great variety of different disciplines and subject areas concerned. Consequently, it is problematic to find generally accepted definition of its subject, scope, methods, and in the end, its place among other neighboring disciplines.

In regard to its subject and scope, political linguistic studies may indeed include a broad area of study, depending on the definition of what is considered political. This can range from politicians' speeches, to general political discourse in media and political discussions. Political language is namely difficult to define strictly as a language variety, as pointed out by Dieckmann:

> Politische Sprache insgesamt betrachtet, zeigt jedoch weder auf der sprachliehen Ebene noch in den redekonstellativen Merkmalen ausreichende Gemeinsamkeiten. die die Wahlkampfbroschüre, die Beratung in einer Fraktionssitzung oder einer Kabinettssitzung, die Parlamentsdebatte, die Neujahrsansprache des Bundespräsidenten, die politische Gesprächsrunde im Fernsehen oder den Sprechchor auf einer Demonstration, zu schweigen von einer Tarifverhandlung, einem Antrag bei einer Behörde und deren Bescheid oder dem Urteilsspruch eines Richters, überdachen könnten (Dieckmann, 2005, p. 22).

In terms of vocabulary compared to general language, political language does not differ much from general everyday language[16] (Weber, 2011), Political speeches in the parliament share certain textual properties such as politeness formulas used to address other MPs, specific forms of impoliteness and other typical dialogical features (van Dijk, 2004). As suggested above, this area only briefly touches the tip of the iceberg of various research scopes and disciplines that political linguistics encompasses, but it does provide some relevant theoretical background for lexical study of the speeches in the EU Parliament.

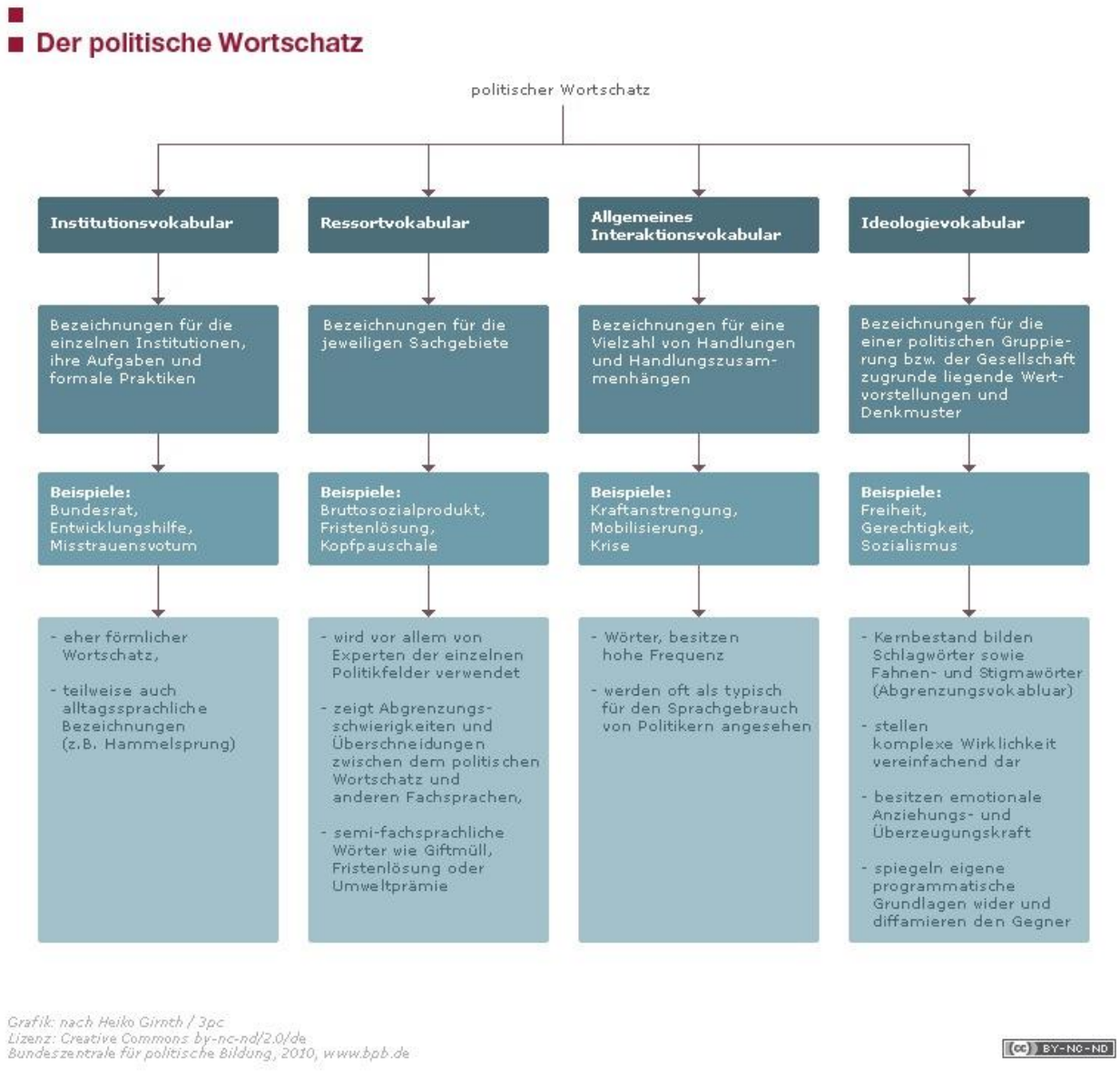
In Klein's view, the language of politics cannot be regarded as a specialist language due to many overlaps with general language, but on the other hand, he notes specialist elements in political language that typical for language for special purposes. To shed some light on this issue, he divided political vocabulary into 4 categories:

[16] It must be noted that connotations of general words in political context may vary from the connotation of these words typical for general language, but by putting semantic differences aside, we can see numerous similarities on the lexical level.

institutional vocabulary, *vocabulary of specialist fields*, *general interactional vocabulary* and *ideological vocabulary.*

A short overview of this categorization can be seen in the following image.

Figure 6. The political vocabulary by Klein. Girnth (graphic design). 2010. Retrieved from http://www.bpb.de/cache/images/9/42689-3x2-galerie.jpg?59DE5



The institutional vocabulary designates various governmental institutions (e.g the parliament), institutional roles (e.g. senator), codified norms of institutional actions (treaty), and institutional process or state (public hearing).

The vocabulary of specialist fields accounts for a certain specialist register used in the political context. These specialization fields may include social policy, finances, environment, culture etc. Here it is difficult to differentiate between political vocabulary and other jargon or everyday language due to significant overlapping.

The general interaction vocabulary includes terms which occur in everyday language interaction not only in parliamentary settings. The vocabulary includes words as *treated*, *demand*, *suggest*. Such words occur frequently in politics and can be considered typical for the political language, but they are used in general language as well.

The ideological vocabulary accounts for expressions defining the values and principles, such as *freedom, justice, nation, terrorism*, etc. This category is rather linked to semantics (flag words, stigma words), because they carry a significant portion of emotion and are often used for influencing the recipients (appellative function), but can be also used for lexical categorization.

We would like to point out that Klein's distinction between the institutional vocabulary, the vocabulary of specialist fields and the general interaction vocabulary is in our view crucial for the glossary compilation for one particular reason. The first two groups represent terminology that is similar to specialized vocabulary in LSP terminological entries. However, general interaction vocabulary, sometimes commonly used in non-specialized settings, but still creating a political jargon or a discourse, can be seen as the necessary prerequisite for the usage of the more specialized words. It serves as a glue to more technical words and enables the speaker to express himself in a language similar to LSP. To put this metaphorically, the general interaction vocabulary in political settings represents the invisible roots of a tree - the various branches of the tree as highly specialized terminology and actually the tree itself could not hold together without basic fundamentals that are provided by general interaction vocabulary.

This is the very reason, why Klein does not consider political language to be a specialized language but at the same time he admits that it uses specialized terminology.

To sum up this chapter, we have explored that political language cannot be fully seen as a specialist language in spite of sharing several characteristics of language for special purposes. On the other hand, it is has been shown that political language represents a language variety and that its specific vocabulary clearly requires special attention by learning, especially for foreign language speakers willing to communicate in the parliamentary discourse. Klein's categorization of political vocabulary is an important reference that we consider fundamental for understanding the commonalities and differences of political language in regard to other specialist fields. By reviewing the literature in the field of politolinguistics, we also came to the conclusion that our quantitative lexical corpus analysis of word frequencies in parliamentary speeches cannot give an exhaustive account of a political jargon due to the fact that the frequency of political vocabulary varies widely depending on the particular term and context (Townson, 1992). Without consulting politolinguistics, categorizing glossary terms would be based only on intuition. Still, we have to keep in mind that the aim of this thesis is to create a glossary based on theoretical knowledge and not to validate the hypothesis that the produced glossaries are an effective learning tool. We can only assume that on the findings from various scientific disciplines that we consulted by creating the glossary.

In order to determine the usefulness of parliamentary glossary based on a quantitative frequency-based analysis of the Europarl corpus, we reached out to renowned experts in political studies to consult whether the glossary would be in their opinion of an asset for students of political sciences.

We have introduced the keyword list and asked: Do you think that learning a vocabulary list of 2000 most frequent words from the Europarl corpus can be of an asset for a person interested in political sciences?

Statement by Mgr. Andrea Tittelová, founder of the civic organization Youth Politics Education, member of JEF Europe, political activist:

> Myslím, že ma zmysel sa naučiť takúto súdobú terminológiu v praxi ak človek pracuje a ide pracovať do danej inštitúcie. Ma zmysel ak ho čakajú EPSO testy alebo na akademickej pôde. Terminológiu som sa učila v inštitúciách, škole vo Washingtone a Euparlamente a veľa na mojich štúdiách v odbore Európske štúdiá,

> kde bolo prirodzené vedieť tieto veci. Na EPSO testy[17] som sa neučila nikdy. Myslím si, že je potrebné sa učiť veci a aj jazyk keď treba intenzívne pracovať v odbore a potom pomôže prax s týmto jazykom (A. Tittelová, personal communication, November 8, 2015).
>
> Translation:
>
> I think that it is reasonable to learn such contemporary terminology in practice if someone is about to start working for an EU institution. It is also advisable if he or she has to go through EPSO exams or it may be useful if this terminology is required for his academic practice. Personally speaking, I learned the political terminology of political institutions, at the University of Washington, in the European Parliament, and initially during my master degree European Studies where we were naturally required to be well versed in political vocabulary. I have never studied for EPSO exams. I take the view that it important to work both on knowledge and the language of this subject. Practice then helps to get the ball rolling.

As we could see, Mgr. Tittelová is in favour of such vocabulary learning tool and considers it useful for people that need to grasp the political language. As we could see, she learned the necessary vocabulary through context during her studies. This implies that the set of glossaries as a proposed learning tool for LSP communication might be more useful for people who are only beginning to get familiar with the specialized topics and the relevant LSP than those who have already acquired both the field knowledge and the language associated with it. We conclude both from the theoretical review and from the practical survey that our tool might be an asset for people learning the basics of the EU parliamentary language.

Lexicological and Terminological Studies

The corpus-based analysis of words would be unthinkable without prior knowledge borrowed from lexical studies including lexicology, lexicography, lexicometry and terminology. We will firstly define the terms and explain their area of focus and then will move to their knowledge and practical solutions relevant for our study.

[17] EPSO exam: tests by the European Personnel Selection Office for the fixed-term contract by the EU institutions.

It is also important to mention that while some authors differentiate between lexicology and lexicography as the practical component of lexicology, and between terminology and terminography, we decided not to make such distinctions. The reasons behind is that we treat them both as a theoretical background that we apply practically in our glossary compilation and therefore there is no point in differentiating which one from them is more theoretical or practical.

Lexicology. Lexicology is a branch of linguistics, the science of language, and is broadly defined as: "the study of the lexicon or lexis (specified as the vocabulary or total stock of words of a language) (Lipka, 1992, p. 1). The term stems from 2 Greek morphemes: lexis, meaning "word, phrase", and logos which refers to learning, a department of knowledge". The literal meaning implies that lexicology is a science of the word but this would be incorrect to think for each linguistic disciplines takes account of words in one way or another approaching them from different perspectives (e.g. phonetics as the science investigating the phonetic structure, grammar, as the study of grammatical structure of words etc.) (Davletbaeva, 2010). Still, despite its specific focus, lexicology should not be strictly separated from phonology and grammar as other constituents building the language system (Singh, 1982).

Similarly, lexicology has its own aims and methods of scientific research. Its basic task consist of a study and systematic description of vocabulary in respect to its origin, development and current use (Davletbaeva, 2010). The lexical level, or in other words, the scope of linguistic analysis in lexicology, is concerned with words, variable word groups, phraseological units, and with morphemes. Linguists proceed on the assumption that the word is the basic unit of a language system, understood as a structural and semantic entity within the language system (Davletbaeva, 2010).

Lexicology can give us practical insights into the theory of homonymy and synonymy that are useful by compiling a glossary. For instance, as we learn in Lipka (1992), synonyms, as words having the same or nearly the same meaning, can pose a threat in language interaction if one is unaware of their subtle differences (e.g. jail as a facility with the duration of incarceration < 1 year vs. prison with the duration > 1 year). Therefore, it is important to look for any specifications in dictionary resources in order to provide clear distinction in the glossary when necessary.

Another interesting finding related to a multilingual glossary is provided by etymology, a sub discipline of lexicology. The etymological research shows that the number of borrowed stock of words is considerably larger than the stock of words with its origin in the respective language. For instance, the English vocabulary consists only of 30% of native words with the rest borrowed from other languages (Davletbaeva, 2010). This high tendency of borrowed words suggests that a large number of words found in the language resources for our glossaries' target languages will in fact be either borrowed words or established internationalisms. Davletbaeva notes that words borrowed a long time ago are practically indistinguishable from a native word without undergoing a thorough etymological analysis (for examples, see Žigo, 2001), but the internationalisms as a neologism can be easily traced as words brought from a foreign language.

As we know, neologisms can be a source of confusion since some of them might not be standardized by codification, thus should not be used in proper language use.

As long as the neologism is codified in the language, there are no practical problems associated with its use. However, if the word has not been codified and is marked as colloquial or non-standard, the speaker might provoke negative reactions with its usage (lack of linguistic competence, low education level, nonchalance etc.) This phenomenon is called lexical interference (as opposed to other types of interferences, e.g. phonetic). Interferences in general can be seen as "transfers of elements of one language into the learning of another (Skiba, 1997). Lexical interferences refer to the practice of borrowing of words from one language and converting them to sound more natural in another language (Skiba, 1997). Interferences are a common topic in language teaching for L2 learners have a propensity towards using lexical and grammatical structures from their native language (see Bhela, 1999), but it should be emphasized that interferences can also happen the other way around, i.e. from L2 in L1. This is mainly the concern of interpreters and translators and is well documented in interpreting studies (e.g. see Lamberger-Felber & Schneider, 2009).

The typical words prone to become the subject of interference are called false friends and already mentioned international words whose standard usage is often unclear.

False friends (in French *faux amis*, coined by Koessler and Derocquigny in 1928) are two words similar or equivalent graphically or phonetically in two or more given languages while having different meanings. They are considered to be an extreme trap for

translators, interpreters, and in general, for any bilingual person. They can be classified in two groups: chance false friends and semantic false friends. Chance false friends represent similar or equivalent words without any semantic or etymological reason, and semantic false friends are those sharing similar or equivalent meaning because they are etymologically related. Such words share the same etymological origin, but over the time have developed different meanings in each language (Chamizo-Domínguez, 2006).

Semantic false friends are further divided into two categories: full false friends (completely different meaning) and partial false friends (several meanings, one of which coincide with the meaning of the word in another language) (Chamizo-Domínguez, 2006). An example of full false friends in French – Slovak is the term *vin rouge* which sounds in Slovak like *ružový* (pink) implying the term *ružové víno* (rosé wine) even though *vin rouge* stands for *red wine*). Here it is important to note that some false friends are more obvious than the others and without proper lexical knowledge one can cause much embarrassment. An interesting example was a misinterpretation by a Slovak host in the Peugeot factory who welcomed a delegation of 50 Frenchmen in a big hall with a warm smile intending to say "and now we would like to initiate you to ….", but actually saying in French: "Et maintenant, nous voudrons vous introduire" ….

Whereas false friends have the same origin and different meanings, international words usually have the same origin and the same meaning. The only peril in internationalisms is the fact whether they are codified or not. International words are the words borrowed by several languages and are significant in the field of communication in different countries. Such words are for instance *philosophy, politics, progress* (Davletbaeva, 2010), or *robot* that actually comes from the Czech language (Robot, n.d).

Here we consider wise to link internationalisms with Zipf's law which is an area related to lexicology. Kingley Zipf, the author of the law, postulated the theory that a language has its tendency to correct itself and strives permanently for an equilibrium state, also called as the *Principle of Least Effort* (Van de Walle & Willems, 2006). Should the there be no equivalent word for an expression of an important phenomenon in a foreign language, the native language has the tendency to borrow it from a foreign language with a higher or lower degree of adaptation. Whether a word also becomes codified depends on time and on the fact how widespread the word is in general population. However,

before the word actually becomes codified, one should refrain from its usage in official communication[18] (Kubišová, 2015).

Speaking about proper lexis in regard to the learning process, it is also important to mention the difference between errors and mistakes. Ellis says that errors reflect gaps in the learner's knowledge and that they occur because the learner is unaware of the fact that the expression is incorrect. Mistakes, on the other hand, Ellis describes as occasional lapses in performance that occur in a particular instance in which a learner is unable to perform what he or she knows (Ellis, 1997). We take the view that by listing potential lexical perils in our glossary, we can both teach the unknown and reinforce the known by language learners which should lead to elimination or at least reduction of both learners' mistakes and errors.

The lexical knowledge mentioned in this chapter is tremendously important in order to avoid unintentional errors and spreading incorrect language usage in glossary compilation. Although we have only scratched the surface of this linguistic discipline, we could already identify some important aspect that must be considered in the process of glossary compilation. The lesson taken from lexicology is to compare words with references of codified language and to highlight language perils for glossary users. This contemplation led to creating Glossary 2 that will specifically deal with correct and incorrect expressions in target languages. Identifying correct sources as lexical reference of standard language will be described more in detail on page 124.

Lexicometry. Looking away from meaning of words and taking them solely as a countable sum of occurrences in a text, we touch the concept of lexicometry. Lexicometry as a methodology rooted in mathematical and statistical linguistics is without preconceived ideas about the language (Bondi & Scott, 2010) which makes it very flexible for creating the overall picture from the statistical point of view. Lexicometry offers some important insight that can be useful by glossary compilation. Its practical

[18] An interesting example is the word *selfie* which is being considered for codification by Slovak linguists as opposed to words *button, post,* and *wall* which are common in internet communication, but not so widespread or without a proper Slovak equivalent (Kubišová, 2015; Sudor, 2015).

results reflect in creating HFWL and other wordlists or keywords list for language for special purposes that facilitate vocabulary acquisition.

From a general point of view, it is important to come back to Zipf's law. The weak version of Zipf's law proposes that words are not evenly distributed across texts. Instead, there is a small number of words that are very common and a very large number of words that are very rare (Sterbenz, 2013). In more exact terms, it states that the frequency of occurrences of a word multiplied by the rank of this word (with rank understood as the place this word has in the list of all the words of a text) is constant. In this sense, the ten most frequent words comprise about 25% of the language, the 100 words 50% and 50,000 words about 95%. To account for the last 5%, one would need a vocabulary of more than a million words. This law has been newly proved in Oxford English Corpus, consisting of text from the 21st century in a size of more than 2 billion instances of a word (Sterbernz, 2013), but Zipf proved its validity already in 1935 on several other languages (Van de Walle & Willems, 2006). Although lexicometry is considered only to be a method within corpus linguistics and statistical linguistics, its results are of a great importance with further application in these disciplines.

Since our quantitative lexicological corpus-based analysis of the EP Corpus is based on statistical results, we cannot oversee the importance of lexicometry for our study. Luckily, there is no need for in-depth knowledge in statistics and lexicometry in this study because the corpus software does all the counting by itself. Nevertheless, the contribution of lexicometry to corpus linguistics in general and to our study in particular should not be left unacknowledged.

Terminology. Terminology can be understood both as "the structured set of concepts and their representations in a specific subject field"[19] (Wright & Budin, 1997, p. 325), and as scientific discipline that studies "the structure, formation, development, usage and management of terminologies in various subject fields" (ISO 1087-1, 2000).

The fundamental principles of terminology science are based on the terminology triangle consisting of a term, a concept and of an object. The object is understood as an entity existing in the world (Schmitz, 2006), the concept as "cognitive representatives" for objects (Felber & Budin, 1989), or "units of knowledge created by a unique

[19] E.g. the terminology of medicine, engineering, sociology etc.

combination of characteristics (ISO 1087-1, 2000), and term as "a verbal designation of a general concept in a specific subject field" (ISO 1087-1, 2000). A term may consist of one or more words (single-word terms vs. multiword terms) which brings us to a clear distinction from lexicology. Terminology deals namely with terms consisting of one or more words referring to a particular object, whereas lexicology is the study of the form, meaning and use of words, and rather interested in meaning of individual words (with the exception of phraseology that studies words forming a certain phrase or an idiom) (Terminology, n.d.).

The importance of terminology in both senses has grown in the recent decades due to the increase in knowledge in almost all technological, economic, political and cultural fields (Schmitz, 2006), intensified through its dissemination over the internet. The spread of knowledge is becoming less limited to a close group of specialists or to one language community[20] and terminology plays a vital role by efficient knowledge transfer. Terminological work takes place at universities, associations of various specialists and at companies in either one or more languages and represents an important part of modern knowledge management. We can only assume that the importance of terminology will gradually increase in the next years to come.

Limitations of the study in regard to terminology. Still, there are two main objections that could question the need for terminology as a linguistic discipline in the theoretical background of our corpus analysis.

Firstly, one might contradict that the high frequency words identified as keywords in our glossary do not belong to a special vocabulary, thus there is no need to consult terminology as one of the disciplines. Indeed, by examining the keyword list, there are not many highly technical words in the selected most frequent keywords. However if we consider the categorization by Trimple and Trimple (1978), who differentiates between

[20] This proliferation of information is a double-edged sword. On the one hand we can name 2 good examples such as medical websites that promote public awareness about diseases and health issues, and Snowden's whistleblowing that shed light on many fishy practices of the US government, but on the other hand, the dark side as the negative effect of this the availability of such information can be seen in breeding ground for cyberchondria and conspiracy theories.

highly technical terms (terms unique to a particular domain), bank of technical terms (from which all disciplines can draw), and sub-technical terms (common words that have taken on special meanings in a certain specific field or discourse), we realize that the number of highly technical words is in a considerable minority, thus not expected to appear frequently.

The second valid objection could be pointing out that the quantitative keyword analysis has identified keywords as the most frequent words of a corpus compared to a reference corpus, but has not identified the most frequent terms, thus this is rather a lexicographic then terminological work. For instance, the corpus analysis identifies the word European, gives an overview how often this word collocates with other words, such as union, project, initiative, but does not provide any information about the terms that are created from these collocations. This objection cannot be refuted, either. The corpus analysis cannot provide identification of the most frequent terms, unless these terms are preprogrammed in a special list.

Regarding the first objection, we must bear in mind the results of other lexical studies which prove that highly technical vocabulary does not tend to be as frequent in a text as other more common words (more in chapter *vocabulary categorization*). Therefore, it is reasonable to expect that our corpus analysis will identify only the tip of the iceberg of vast terminological vocabulary covered in the EU parliamentary language. Still, it is important to note that the aim of this study was to carefully examine the keywords and not to deliver an exhaustive terminological work. To our mind, the challenge on a lexical level of the speech production may be caused more frequently by sub-technical words that are not fully automated than occasional stumbling upon a highly technical term. We expect the compiled glossaries from the keyword analysis to be a useful help for improving communication skills in the EU parliamentary language as an LSP, and not for it to become a terminological resource.

In regard to the second objection, we do not intend to refute it either. We are aware of this great disadvantage of the qualitative analysis, therefore, we will seek to compensate it in the qualitative analysis. In one part of the qualitative analysis, the keywords will be examined if they have any collocations that could together create a term. This analysis will be corpus-driven and dependent on researchers' intuition and careful

consideration. The results will be summarized in Glossary 3 that will focus on explaining the most frequent terms.

Nonetheless, even if our arguments to the two abovementioned were proved false, there is still one important reason for consulting terminology in this corpus analysis: Terminology provides a useful framework for adequate glossary compilation. It offers standards of good practice by terminology management and serves as a theoretical springboard for our glossary compilation. Thanks to ISO 10241-1, which standardizes various glossary properties, we could come up with an optimal solution for the glossary layout and relevant information fields which suits our lexicological needs.

Terminological standards for glossary compilation. Terminology has also served as a great inspiration for Glossary 2 that deals with false friends. Proceeding from ISO 10241-1, we decided to add categories of approved and deprecated terms for each language. To elucidate on the concept of approved and deprecated terms, as well as standard entry according to terminological standards, let us quote Pearson (1998, p. 23):

> Entries for terms will contain as a rule, a term number, the preferred terms, agreed definition, the field or subfield in which the term is to be used, related terms, and deprecated terms. They will not contain any indication of usage in terms of common collocates or grammatical restrictions. Deprecated terms are those terms which have been, or still are, used to refer to the same concept. By stipulating that they are now deprecated, the standardizing authority is attempting to prohibit further use of such terms. Standardized terms are not always new terms in the sense that they do not suddenly come into existence because a standardizing body decrees it. Standardized terms are generally terms that have already been coined by users, of the terminology. What the standardizing body does is give its seal of approval to one term and make recommendations for preferring that particular term over others which may have been used to describe the same concept in the past.

As we could infer from the citation, the glossaries presented in this thesis (more explained in the practical part in chapter *glossary templates*) do not strictly follow the terminological standards. Glossaries 3 and 4 include various lexical information related to the word and the other glossaries may be considered as inspired by the terminological standards, yet not strictly following them. This is the case particularly for Glossary 2 that

was inspired by the concept of standard and deprecated terms. We deemed it particularly important to point out linguistic perils of non-standard language associated and of possible false friends in internationalisms which was found as important to include in our work. Identifying such words and terms will be an arduous tasks requiring not only exhaustive terminological research, but also consultations with language experts. The list of used resources for Glossary 3 will be further described in chapter *corpus linguistics software & language resources*.

Computer-assisted Language Learning (CALL)

We have pondered on the question how to make the presented glossaries easily accessible to the target audience and at the same time how to harness the potential of modern technology in the most effective way. We understood that providing an electronic table document might not be sufficient. Therefore, we looked for an option to provide data in a format accessible to various learning software that is both portable, easy to use, and can facilitate the learning process. In this chapter, we would like to briefly introduce learning programs that we chose both for dissemination of the glossaries and for their most effective teaching capabilities. To reach this objective, we had to consult the field of computer-assisted language learning (CALL).

CALL is standardly defined as the use of computer in the language art classroom for instruction (Fox, 1993). However, many things have changed since the first definitions were coined and nowadays we could call CALL tools basically any program that facilitates the learning process. There is a whole array of various CALL tools, ranging from grammar exercises, virtual classrooms, vocabulary trainers, quizzes and games, to whole websites containing publishers' web materials as enhancement to their textbooks and metasites as collection of resources. Still, CALL tools share some common attributes as proposed by Hubbard:

- Learning efficiency: learners are able to pick up language knowledge or skills faster or with less effort;
- Learning effectiveness: learners retain language knowledge or skills longer, make deeper associations and/or learn more of what they need;
- Access: learners can get materials or experience interactions that would otherwise be difficult or impossible to get or do;

- Convenience: learners can study and practice with equal effectiveness across a wider range of times and places;
- Motivation: learners enjoy the language learning process more and thus engage more fully;
- Institutional efficiency: learners require less teacher time or fewer or less expensive resources (Hubbard, 2009, p. 2).

In recent years, we have witnessed a boom of portable technology. Smartphones and tablets now have the processing capacity of small computers and can be equally used for CALL. We have reviewed the newly published web content (2013-2015), and searched for the most popular and widely used multi-platform vocabulary training programs (aka computer assisted vocabulary learning, CAVL) that also could accommodate our needs for vocabulary import. After careful consideration, we have decided for Anki and InterpretBank which we will shortly describe.

Anki. Anki is a multiplatform program that uses flashcards for remembering information. It is based on two learning methods: recall testing and spaced repetition. The recall testing consists of being asked a question and trying to remember the answer. It is in contrast to passive study in which one reads, watches or listens to material without pausing to consider whether he or she knows the answer. The act of recalling knowledge is said to strengthen the memory and to increase the chances that the subject will be able to remember it again and at the same time points out difficult questions that need further attention.

The spaced effect is a theory based on scientific findings back in the 1885 which proved that subjects tend to remember things more effectively if the reviews are spread over time instead of studying multiple times in one session. The flashcard system (cards with a front page that serves as a question and the back page that provides an instant answer) utilizes this theory for learning purposes. The system was popularized by Sebastian Leitner in the 1970s who developed a method of studying flashcards based on categorization of flashcards according to recall performance of a subject.

Anki is according to our review probably the most favorite and most widely used flashcard learning software. It can be used for learning almost any information for it supports images, audio, videos and scientific markup via LaTeX. Anki uses the SM2 algorithm based on the Leitner system that was programmed for the program

SuperMemo. Anki, unlike Supermemo, is free of charge, open source, running on multiple platforms (Windows, MAC OSX, Linux/FreeBSD, Android and iOS. Flashcards can be easily imported in a CSV format and later search for in an open database[21] (Elmes, n.d.).

InterpretBank. InterpretBank is software designed for professional interpreters to create and manage glossaries. It supports glossary sharing, synchronization on more computers, various import and export functions. InterpretBank consists of three modes or utilities: TermMode, ConferenceMode and MemoryMode. The TermMode is the heart of the program and enables users to create and manage their glossaries in one interface, to access information from the Web (definitions & translations of specific terms on preconfigured trusted websites, simple lookup operations and automatic machine translation) and to import and export glossaries. The ConferenceMode serves for instant looking up glossaries in the booth. The software has powerful search options that can be very useful in the conference booth and provide a great alternative to classical paper glossaries. The MemoryMode is a simple tool for learning the glossary. It shows the glossary terms alternatively in the source and in the target language (similarly as flashcards). The MemoryMode is recommended for rehearsal of small glossaries (50-100 words) before the beginning of an interpreting assignment (Fantinuoli, 2015).

The development of computer tools for interpreters could be seen rather as a novel approach in conference interpreting because interpreters and interpreting students still tend to rely heavily rather on paper glossaries regardless the computer advancement. The future trends of using computer tools for interpreters such as InterpretBank are hard to predict, but some preliminary studies (e.g. Gacek, 2015) have already shown their usefulness and increased effectiveness compared to classical paper glossaries. Therefore, we stay optimistic and think that providing our glossaries in the InterpretBank file will be a good choice in terms of our target audience and of future trends in interpreting, notwithstanding the fact that there is a relatively small end user group of interpreting students compared to other L2 speakers interested in the EU parliamentary language or in LSP language in general. Our glossaries in the InterpretBank format can therefore not only be used as a default database for testing the functionality of the program, but also

[21] See https://ankiweb.net/shared/decks/

could serve for learning the most frequent words of the EP corpus and act as a basis for further building of conference terminology.

While Anki is a popular program used by a wide community of learners on many subjects, InterpretBank is rather a specialized software developed for interpreters with a small user community, but on the other hand, there is a higher chance that students of conference interpreting might get interested in learning the glossary. Since we have seen the necessity to carefully balance the need for making the glossary accessible for potential users and the need to effectively address young interpreting students who might use it the most, we have decided to use these two programs. Even though the size of Anki community is considerably bigger than the community of interpreting students, and only a small fraction of Anki members might be interested in parliamentary language, sharing the glossaries in this community has been evaluated as a good option to reach potential users. Therefore, we deem the combination of both Anki and InterpretBank as the best solution for making glossaries publicly accessible and easy to use.

Conclusion

In this chapter, we consulted other academic disciplines for proper glossary compilation. These included language for special purposes and its special area of politolinguistics for correct selection and categorization of words in the generated keyword list, lexicology to identify and address possible language perils at the lexical level, terminology for creating glossary design and learning the proper procedure for glossary compilation, and finally, computer-assisted language learning (CALL) for reaching out to the target audience and facilitating the learning process. In summary, these disciplines provided important knowledge for proper selection, organization and presentation of keywords in the glossaries.

Corpus Linguistics as the Method for Creating LSP Glossaries

Introduction

This chapter seeks to explain the importance of corpora and corpus linguistics in language studies in general and in creating the intended LSP glossary about the most frequent political terms used in proceedings of the European Parliament in particular.

In the first part of the chapter, the reader will get introduced to corpora, their various aspects, types and properties. Then, a short selection of language resources related to corpora, the EU and our area of focus will be presented.

The second part of the chapter sets out to investigate corpus linguistics as a methodology, and look closer on its history, purpose, importance in a transdisciplinary context. It will also introduce the main features & functions related to the fields of corpus linguistics such as corpus analysis, corpus linguistics software and corpus linguistics tools.

The chapter should provide sufficient background knowledge to understand the process of the quantitative corpus-based lexical analysis described in the next chapter. It must be pointed out that this chapter is not meant to provide a description on corpus design because the corpus used in the study had been compiled according to all corpus linguistics standards, thus can be understood as a reliable source on the whole. Information about proper corpus compilation may be found in relevant literature, e.g. Pearson & Bowke, 2002).

Corpora

Definition. Corpus is in simple terms a body of natural occurring language. In Latin, the word corpus means body, hence it can be used to refer to any body of text (McEnery & Wilson, 2001). There is a considerable amount of various definitions of a corpus, but in this section, we will rather discuss narrowing of the accepted definitions regarding their scope rather than enumerate all the known.

The Oxford dictionary defines corpus in regard to language as “a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject”, or as “a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research“. (Corpus, n.d.). The

latter definition implies the usage of corpora in modern computer-assisted corpus linguistics. A similar definition provides Baker, who defines corpus as "any collection of running texts … held in electronic form and analyzable automatically or semi-automatically, rather than manually" (Baker, 1994, p. 226). As we can see, a corpus is a prerequisite for corpus linguistics which will be described more in detail in the next chapter.

Next, it is important to distinguish between the terms corpus and collection, or archive. Any compilation of text or an archive does not necessarily make a corpus. A corpus is rather to be understood as a collection of texts gathered with a special linguistic objective in mind or from a linguistic perspective, and according to objective criteria set for the respective corpus. For this reason, Sinclair prefers to define corpora in a less flexible way in order to make them more useful to language studies and postulates 4 main criteria that should characterize a corpus: considerable quantity, authentic quality, plain text simplicity and documentation of compilation criteria, annotations etc. These criteria refer to corpora as samples of a language in a broad sense and a discourse in a narrow sense. This reflects in Sinclair's definition of a corpus:

> A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair, 2004, p. 19).

Similar criteria as suggested by Sinclair offer McEnery and Wilson (2001), namely sampling and representativeness (quality), finite size (quantity), machine-readable form (plain text simplicity) and a standard reference (documentation). These will be elucidated more in details in the chapter to follow.

Qualitative aspects of a corpus. The quality of a corpus encompasses its sampling and representativeness, balance and authenticity.

Sampling and representativeness of a corpus stands for the criterion that seeks to make a corpus maximally representative of the variety under examination. The aim is to provide a reasonably accurate picture of the tendencies of the language variety examined, as well as their proportion. This variety can be understood and specified in many regards, for instance from the analysis of a particular author to the analysis of the whole subgenre, genre or even the whole written discourse. The exact aim of the analysis lies in hand of a

researchers and the representativeness of the corpus or design criteria for its compilation must be accustomed to envisaged expectations and needs accordingly.

There are two options for data collection: first, researchers could analyze every single utterance in that variety; or second, they could construct a smaller sample of the variety. The first options is clearly problematic due to the fact that a corpus is finite, but language is not. Language is in fact living and evolving, therefore the number of occurrences is constantly increasing and is theoretically infinite. Thus, analyzing every utterance would be an unending and impossible task. Chomsky criticizes corpus linguistics stating that corpora, however large they become with modern technology, will always be skewed, some utterances would be excluded either by rarity or chance (McEnery & Wilson, 2001). Still, as he further elaborates, "Standard methods of the science is not to accumulate huge masses of unanalyzed data and to try draw some generalizations from them" (Chomsky, 2004, p. 97), therefore there is no point in doing that. Documenting a full coverage, as the first mentioned option, is reasonable and achievable only in specialized areas or in case of dead languages with a finite number of texts[22].

We should consider the fact that Chomsky's criticism does not mean that corpus linguistics is faulty. He speaks in absolute terms and challenges this area of research rightfully by pointing out the perils that corpus researchers might encounter. Instead of arguing, corpus linguists should rather seek to establish ways to construct much less biased and more representative corpora. From the perspective of corpus analysis, it is extremely important to have clear research objectives and to use a maximally representative and relevant corpus which provides the researcher with an accurate picture of the tendencies of that variety, as well as their proportions (McEnery & Wilson, 2001). According to Leech (2014, p. 27), a corpus is representative when "the findings based on its contents can be generalized to a larger hypothetical corpus".

22 E.g. the Indus Valley civilization around Harappa and Mohenjo-daro flourished between 2500 – 1900 BC, and total amount of found written material (700 inscribed objects) remains to represent the language used by the civilization. There might be some new archeological findings in the future, but it is unlikely that they would alter extent of this stock of text (McEnery & Wilson, 2001).

The second mention qualitative criterion is corpus balance. The balance of the corpus goes hand in hand with its representativeness, but while representativeness accounts for the selection of texts for building a corpus chosen to represent the selection in respect to wider population, corpus balance refers to the distribution of various text categories within the corpus. The purpose is to build a manageably small scale model of the linguistic material that is to be studied (McEnery & Gabrialatos, 2006). By corpora compilation (selecting texts for corpora), both representativeness and balance should be considered and specific criteria derived from research objectives should be put in place (e.g random sampling methods, demographic sampling etc.).

The last mentioned qualitative factor of a corpus is its authenticity. Since corpus linguistics (more in chapters to follow) is the study of real language or a language in use, the corpus must include real instances of the language or the language variety under study (Lang, 2010). Therefore it is expected that corpora are built from genuine communication of people, rather than artificially invented sentences. Authentic corpora can then found application in language teaching, dictionary building, translation and other fields, further elaborated in chapter *Corpus linguistics application*.

To conclude, it is important to mention that the degree of authenticity, balance and representativeness are not precisely definable and attainable goals and hence difficult to measure quantitatively. Nevertheless, they must be used to guide the corpus design and the selection of its components. These qualitative criteria are namely the very core prerequisite for building a corpus that will be used for making claims about a larger population.

Quantitative aspects of a corpus. Under quantitative criteria for corpora falls the size of the corpus. In recent years, we could witness a dramatic progress in available corpora who have grown in size exponentially. Corpus linguistics has indeed thrived under technological advancement and thanks to bigger corpora generating more reliable results in corpus analysis, it has found its application in various other fields (see chapter *corpus linguistics application*). In general, the commonly shared belief, especially in corpus-driven linguistics, is that bigger corpora can offer more reliable results and a more accurate picture of a language under scrutiny. Sinclair points out the perils of small corpora used for making claims about broader population:

> "There is no virtue in being small. Small is not beautiful; it is simply a limitation. If within the dimensions of a small corpus, using corpus techniques, you can get the results that you wish to get, then your methodology is above reproach - but the results will be extremely limited..." (Sinclair, 2004, p. 189).

Although the size of a corpus matters and significantly improves its representativeness, it cannot be seen as sole guarantee for it. Again, the optimal size of a corpus depends vastly on research goals and the range of investigated linguistic phenomenon (grammar, lexis) in a particular language portion (language variety, specific text group, discourse etc.) (Blecha, 2012). This consideration applies both to the investigated and the reference corpus, the kind of query that is anticipated from users and the methodology considered for the study.

Regarding making general language claims, small corpora tend to be representative only for certain high frequency linguistic features (McEnery & Wilson, 2001). It implies that Chomsky's criticism of the great limitedness of corpora compared to the unending language was correct. Yet, nowadays this criticism is correct only in absolute terms since modern corpora of hundreds of millions of words, or the web as a corpora with rough estimation of 100 trillion words (Gatto, 2004) can fairly represent the language as it is., Next, small-sized specialized corpora (explained in the next chapter) do not need to be as vast in volume as general corpora to provide accurate and overall picture about the specific text group that they represent. In fact, a well-designed small-sized corpora might provide more useful information than large corpora that are not customized to researchers' needs (Bowker & Pearson, 2002). There are several studies that utilized small-sized corpora and provided accurate findings about the broader population that they represented (see Anthony, 2013).

This implies that the size of a specialized corpora also goes hand in hand with the size of the whole population. There are some statistical tools used for determining the optimal size of a corpus in regard to its representativeness (see Pastor & Seghiri, 2007), but in general, the optimal size of a corpus is a vexing question and there is no panacea solution (Blecha, 2012).

To conclude the discussion about corpus size, we should consider that the value of a corpus is dependent not on its size, but on the kind of information that we can extract from it (Anthony, 2013). The only characteristic regarding the corpus size that all corpora

have in common is their finite size. The exceptions is web that is sometimes advocated as corpus (e.g. Gatto, 2014), and monitor corpora that will be briefly described in the next chapter.

Modern corpora in the machine readable form. As mentioned in the previous chapter, the modern understanding of corpora explicitly expect them to be stored in an electronic form and to be easily searched, processed or augmented (e.g. with corpus annotation). The other options of storing a corpus are in a written or spoken form. Since corpora are meant to document both written and spoken language, they can be recorded in both written and spoken form. Still, for a reasonable corpus analysis, it is necessary to transfer corpora in a machine readable form. This clearly poses obstacles (e.g. OCR recognition of written text and speech transcription of spoken language into an electronic form), but the rapid technology improvement of IT helps overcome these problems.

This resulted in the situation that nowadays, only a few corpora exist in a written or spoken form (McEnery & Wilson, 2001). Keeping large-sized corpora has become rather an act of curiosity than a systematic approach. For instance, a crowdfunding initiative plans to print out 1000-volume edition of Wikipedia (Flood, 2014) to create an absurdly-large collection of Wikipedia texts in written form. Considering the price for printing, handling and storing large corpora in a written form, not mentioning the inability to search through them, there are very few good reasons not to follow the trend of digitalizing texts for corpora in spoken or written form.

In sum, machine readable corpora possess three advantages over corpora in written or spoken formats.

Firstly, they can be searched and manipulated at high speed. As we will show in the chapter *corpus linguistics application*, modern computers revolutionized corpus linguistics and opened new avenues of research. Computer-assisted corpus linguistics tools enable to search for a particular word in a corpus, to calculate the number of occurrences, sort linguistic data, or to display searched words in context. Those are only basic function of corpus tools that would be impossible for large corpora without the use of computers (McEnery & Wilson, 2001).

Secondly, machine readable corpora can be easily enriched with additional information that further improves the search options and increase the informative value

of the corpus (McEnery & Wilson, 2001). This process is called annotation and will be further discussed on page 73.

Thirdly electronic corpora can be gathered more quickly than those in written or spoken form (Bowker & Pearson, 2002). There are plenty of software tools capable of compiling corpora either from web (web spiders, used for instance in Europarl corpus by Koehn) or from a collection of digital texts (e.g. Corpus).

Corpus documentation. The standard practice is to provide documentation alongside a corpus, explaining the design and encoding, information about the purpose, size, date, format, distribution of genres, character set used, annotation, markup system and various tags used in the corpus and other useful information (Baker, Hardie & McEnery, 2006). This information gives a useful insight about the corpus and represent something like a "title page of a corpus"

This short elaboration has shown how important corpus characteristics are for correct corpus compilation. It goes without saying that a poorly designed corpus will inevitably lead to poor results. However, as will be shown in the chapters to follow, a well-built corpus is a prerequisite for a corpus study, but it is not the only essential component. The other two fundamental components necessary for a successful corpus study must also be included, namely the right selection of corpus software and human intuition to interpret data derived from the corpus (Baker, 2012).

Corpus types. The following categorization represents a short overview of various corpus types (see Pearson, 1998; Bowker & Pearson, 2002; Blecha, 2012)[23]

Scope

General (reference corpora): are designed to be a broadly homogenous representation of all relevant language varieties and the characteristic vocabulary (e.g. BNC).

Monitor corpora increase its size with a predetermined rate of flow and allow scholars to monitor and record changes in the language and its use.

Specialized corpora: Many corpora do not fit any known category in corpus classification, therefore are simply classified as specialized corpora. They contain a high

[23] Examples of corpora in mentioned categories can be found here: http://cl.indiana.edu/~md7/13/615/slides/05-corpora/05-corpora-2x3.pdf

proportion of unusual features that differ from general language and are restricted to a particular field, text type or demographic group.

Language

Monolingual corpora: Consist of textual information in a single language.

Bi- / multilingual corpora: Corpora made up of texts in two or more languages. They can be further divided into comparable corpora (the same context, but not same texts) and parallel corpora (exact translation, usually simultaneously produced texts in multiple languages).

Time frame

Diachronic corpora: Represent a particular language in a long period of time that can show the evolution of the language (e.g. Helsinki Corpus of English Texts containing archaic texts from 700 - 1700).

Synchronic corpora: Show a language use within a limited time frame.

Size

Open corpora: Are constantly expanded and used in projects where information about the latest development and a record thereof is necessary. Tend to be unbalanced both in the genre and in the input texts and usually larger than closed corpora.

Closed corpora: Most of modern corpora are closed which implies that the do not get augmented after their compilation. The aim of closed corpora is not to keep track of latest development in a language, but rather to seek balance of texts within the corpus.

Form

Written corpora: Contain written textual information

Spoken corpora: Contain transcripts of originally spoken language (conversations, broadcast, lectures etc.).

Author

Learner corpora: Learner corpora are all those that are compiled from written or spoken utterances of non-native speakers of a language (foreign language learners).

Corpus annotation. Corpus annotation can be defined as the process of "adding such interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Leech, 1997, p. 2). In a broader sense it might be understood as both textual information and interpretative linguistic analysis, and in a more common narrow sense as encoding of linguistic analyses in a corpus text. It is important to distinguish

between corpus annotation that is interpretative, subjective linguistic information to reflect that human language is to some degree the product of human mind's understanding of a text (e.g. annotation of a word that might be ambiguous), and corpus markup that summarizes objective information (e.g. sex of a speaker, discourse details etc.) (Xiao, 2010).

There are various reasons for corpus annotation. In general, we can summarize them by saying that the purpose of annotation is to enrich a corpus with linguistic information that enable researchers to compare similarities and look for patterns. In particular, there are three main reasons for corpus annotation: better information extraction (in order to extract information, some patterns must firstly be built in), re-usability and higher value (proper annotation is an expensive and time-consuming business and an annotated corpus is much more valuable as the original corpus), and multifunctionality (corpus can be used for a multitude for purposes – the final product of annotation can be used as "raw material" e.g. in speech recognition software, SMT, language teaching, development of natural language processing software etc.) (Rayson, 2003; Garside et al., 2013).

The multitude of purposes for which annotated corpora can be used reflects in a great variety of annotation types.

Corpus annotation should adhere to proposed standards which are: recoverability (reversion to the raw corpus must be possible), extractible annotation, and available documentation (the annotation scheme, how/where/by whom annotation were applied, some account of the quality of annotation). Corpus annotation is encoded in tags which are codes assigned to various language features of the respective word (Dash, 2010). This area comprises various encoding styles which are not be discussed further.

In this thesis, we describe the POS and lemmatization as the most common ones and briefly outline other annotation types.

Lemmas. Lemmatization is one of the most basic types of annotation and can be categorized under lexical annotation. Lemma is "a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and / or spelling" (Franzis and Kučera, 1982, p. 1). Thus, lemmatization can be understood the process in which words in various inflected forms (e.g. declension of nouns, conjugation of verbs, different spelling, inflection etc.) are grouped together under their basic form

(lemma). The purpose of lemmatization is to enable corpus users to make generalizations about the behavior of groups of words in cases when their individual differences are irrelevant (Knowles & Zaraidah Mohd Don, 2004). Lemmatization makes possible to treat all the forms of a word together, produce frequency and distribution information about the lemma, retrieve results in various forms by searching the headword etc. For instance, as discussed in chapter *wordlist generation on page 128*, the basic word "be" in the Europarl corpus was found under various forms such as *am, are, been, is, was* etc. The corpus analysis software counted both the frequency for the basic form in respect of other words, and individual frequency of all forms within the basic form. Lemmatization can be performed by using an existing form-lemma database, a (semi-)automatic approach called stemming (cutting of characters in a predefined mode to achieve basic word forms) or by combination both strategies which is useful for text with special terminology that might not have been included in available lemma lists (Gries & Berez, to appear).

It is important to note that lemmatization groups words together on a lexical, but not a semantical or syntactical level which might lead to ambiguity of some words.

Point of speech tagging: syntactic and morphological annotation. The second most common annotation type is POS (point-of-speech) tagging. POS is "the identification of the morphosyntactic class of each word form using lexical and contextual information." (Paroubek, 2008, p. 99). POS tagging is important in various areas because some words can take more than one word class which changes their meaning and pronunciation. Leech (2004) illustrates POS tagging on the word present with simple POS tagging consisting of an underscore symbol and a code for word categories:

Present_NN1 (singular common noun)

Present_VVB (base form of a lexical verb)

Present_JJ (general adjective)

There are various tagsets and tagging methods (parsers) with different accuracy undertaking constant improvement by the linguistics community. The precision of POS is highly dependent on many factors such as the language represented in the corpus and its complexity, the corpora texts, the kind of a tagger and its training for the relevant corpora etc. The precision for English texts varies between 90-95%. (Gries & Berez, to appear). Still, POS only scratches the surface of the complex field of syntactic annotation.

There are many others, more sophisticated corpus linguistic annotation methods that reveal an important insight into the grammatical characteristics of a text based on specific needs of linguistic researchers (Gaside, Leech & McEnery, 2013).

Other types of corpus annotation.

- *Phonetic annotation*: can be generated automatically from orthographic transcription via a pronunciation lexicon and/or rule based algorithms, but manual transcription is vastly time consuming (1 minute: 40-60 minutes) (Gries & Berez, to appear). It is widespread in speech linguistics and uses mainly corpora produced in laboratory situations (Garside et al., 2013).
- *Prosodic annotation*: also wide in scope (from paralinguistic features (stress, intonation, pauses, mispronounced words etc.) in EPIC interpreting corpus to diacritic notations for capturing tonal phrase accents. This annotation type is not very common and still in its early stages (Gries & Berez, to appear).
- *Semantic annotation*: adds a label to every word in a text which indicates its semantic field. This type of annotation enables finding different words with the same meaning, and conversely, distinguishing between different meanings of the same word (Wilson, 2013).
- *Pragmatic annotation*: adds information about the kinds of speech act that occur in a spoken dialogue. For instance, the word *okay* may be on different occasions an acknowledgement, a request for feedback or acceptance (Leech, 2004).
- *Stylistic annotation*: adds information about speech and thoughts presentation such as direct speech, indirect speech etc.
- *Discoursal annotation*: adds information about causality, contrast and temporality. There are two types of discourse relations: the first refers to relations that are signalized explicitly via discourse connectives and the second refers to relations that can be inferred without explicit signaling.
- *Gestures/sign language annotation*: used for annotation of nonverbal aspects of spoken language (Gries & Berez, to appear).
- *Alignment of multilingual corpora*: parallel corpora may contain translation of texts from a source language into one or more other languages separated in units consisting of words, phrases, or sentences that can be used in SMT (Gries & Berez, to appear).

- *Annotation of learner corpora*: learner corpora can be annotated for errors made by English language students. Annotation of a learner corpus is challenging also for other types of annotations (lemmatization, POS, etc.) because non-native English language is more likely to contain non-standard grammar, spelling or syntax which poses a problem for automated lemmatization tools that are trained on the edited and mostly correct English texts (Gries & Berez, to appear)

Certain corpus annotations are useful or even needed for some disciplines whereas useless for the others. For instance, in language teaching, lemmatization helps students to identify the total number of surface forms of a lemma and POS to explore which words are frequently used in combination with a certain word (e.g. looking for common all adjectives in the front of the word *trend*). Similarly, semantic annotation might be useful for elaborating on word differences, whereas a prosodic annotation would be more or less useless in this context. Another example would be using annotation in SMT that can similarly derive a great advantage from semantic and syntactic annotations (the above mentioned example of the word *present* is to have different translations in respective word categories), but could have little use for paralingual annotations which, on the contrary, is of a great importance in interpreting studies.

To summarize, corpus annotation is a very useful way for enhancing the value of a corpus and create new possibilities for working with the corpus. On the other hand, corpus annotation has also some drawbacks. Sinclair offers an interesting critique:

> One cozy consequence of using tagged text is that the description which produces the tags in the first place is not challenged – it is protected. The corpus data can only be observed through the tags: that is to say, anything the tags are not sensitive to will be missed (Sinclair, 2004, p. 191).

Corpora limitations. In order to conduct proper research, one must be also aware of specific limitation of corpora. As already mentioned, Chomsky criticized corpus linguistics fervently in its early stages for zealot interpretation of corpus data instead of using language introspection, and for limitations of corpora compared to infinite possibilities of language. Another criticism of corpora is put by Widdowson for their inability to describe member categories in ethnomethodological terms, to provide contextually appropriate insight into the encoded and to showing decontextualized language (Widdowson, 2000). The criticism of corpus linguistics was legitimate and it

was a source for good for it has led corpus linguistics to adjust and improve methods of corpus building and corpus analysis. Nowadays, corpus linguists are mostly aware of following limitation of corpora, according to Bianchi (2012):

Firstly, corpora present language out of its context, considering other than textual context, such as social and pragmatic context, visual context etc. Despite the fact that there are possibilities to include information about contextual data into the corpus, such annotations are consuming and consequently very rarely used.

Secondly, any corpus must be considered as limited sample of language that shows only its own content. This content can easily interpreted wrongly outside the context and without thorough examination[24]. Therefore, it is very important to treat quantitative data carefully and not to jump into quick conclusions and generalizations

Thirdly, a corpus can provide information about whether some occurrences are used or frequent, but not whether they are correct (e.g. from the perspective of standard grammar). Similarly, corpus linguists cannot say that something is not possible simply because it was not found in a corpus (Bianchi, 2012).

Fourthly, corpora provide linguistic evidence, but not linguistic information. They do not automatically provide answers to linguistic question, thus analysis and intuition are always needed to make sense of the data (Bianchi, 2012).

In sum, corpora are a useful tool for many areas of research as will be shown in chapter *corpus linguistics application on page 89*, but one should also be aware of their limitations.

[24] "For instance, finding two times more occurrences in the Bank of English corpus for left-handed than right handed can lead to wrong conclusion like left-handed people enjoy higher social status and public presence than right-handed people, or imply that there are more left-handed people than right-handed. Careful qualitative analysis reveals that right-handedness is considered to be the norm and left-handedness is a deviation from the norm that is likely to be mentioned, but still, it is important to recognize that this is an interpretation of evidence rather than a fact" (Hunston, 2002, p. 66 , as cited in Bianchi, 2012, p.52).

The Europarl Corpus and Another Corpora & Language Resources in the Context of EU Institutions

The European Unions (EU) is a final result of nearly 50 years' effort for peace, cooperation, stability and prosperity among nowadays 28 member states. It origin can be traced back to the European Coal and Steel Community and the European Economic community that were formed by six countries in 1951 and 1958 respectively. The EU can be described as an economic and political union. With its 500 million citizens, it has a strong position in the world trade (2nd supreme position surpassing the US). The EU has a unique institutional set-up in which member states relinquished part of their sovereignty to EU institutions that represent the EU on a national and international level. The 5 most important institutional bodies are the European Parliament, the European Council, the European Commission, the Court of justice, and the European Central Bank (European Commission, 2014).

Parallel Language Resources Issued by the EU. With its multilingual policy and all 24 official languages enjoying equal status, the EU institutions require translation and interpreting services which they mostly cover by their own language departments. This requires substantial workforce and resources. To improve the efficiency of interlingual communication in the EU community, the EU institutions have supported various projects aimed at language resources development. Their motivation covers mainly the development of more business potential, the improvement of democracy through transparency of information, the maintenance of the EU's linguistic diversity, and the preservation of the EU's cultural diversity (Steinberger, Ebrahim, Poulis, Carrasco-Benitez, Schlüter, Przybyszewski & Gilbro, 2014). Some of those freely available related to corpus linguistics will be mentioned in the next chapter.

According to Steinberger et al. (2014), the EU organizations have released a number of large multilingual linguistic parallel resources in cooperation with the Joint Research Centre (JRC), the in-house science service of the European Commission. The next paragraphs will serve as their brief introduction.

JRC-Acquis: the first sentence-aligned and pre-processed corpora issued by the European Commission. The last version comprises of 22 languages and built on documents from the Eur-Lex website which were made public, but not customized for specific linguistic needs.

DGT-Acquis: a family of four multilingual parallel corpora with 23 languages produced by the European Commission's *Directorate General for Translation (DGT)*. DGT-Acquis is also built on Eur-Lex data that stem from the *Official Journal of the European Union (OJ)*.

DCEP (Digital Corpus of the European Parliament): the DCEP is the latest EU corpus. It is accompanied by tools that allow to produce separate sentence-aligned corpora for each of the 276 language pairs, and contains the majority of the documents published on the European's Parliament website. The corpus includes various categories such as reports, adopted texts, written answers to questions, written questions, national or EU-wide press releases, motions and minutes of plenary meetings with a total of 103 million English words. To avoid overlapping with the Europarl corpus, the DCEP does not contain the verbatim reports of the EP's plenary sessions. The corpus is further subdivided by languages and document type.

DGT-TM: is a translation memory (TM) covering 24 languages and 552 language pairs out of which 58 directly. It provides reasonable quality of sentence alignment data in a considerable quantity. Its further improvement and customization for direct import into statistical machine translation (SMT) accounts for the fact that DGT-TM is the most used language resource by human translators.

ECDC-TM: a TM provided by the *European Centre for Disease Prevention and Control*. The TM was produced from translation of its website into all 24 EU's official languages.

EAC-TM: a TM provided by the EC Directorate General for Education and Culture (EAC). It was created from translations of project & funding applications and reports from the EU programs *Life-long Learning Programme (LLP)* and the *Youth in Action Programme.* Although smaller in size, it covers a wide range of domains, namely education, training, culture, youth and sports. In addition to the usual 22 languages, it also includes documents in Icelandic, Norwegian and Turkish for these countries have participated in the respective EU programs. EAC-TM is expected to be enriched every year with new data (Steinberger et al. 2014).

Additionally, we consider wise to also include to this list the vocabulary thesaurus *EuroVoc* which may serve as a powerful terminological database for people interested in the standardized EU terminology. EuroVoc is a multilingual thesaurus that was originally

built up specifically for processing the documentary information of the EU institutions. It covers a great variety of fields related both to the EU community and national points of EU member states, with a certain emphasis on parliamentary activities. EuroVoc is a controlled set of vocabulary with the aim to provide the information management and dissemination services. It adheres to the newest ISO standards[25], includes 23 EU languages plus Albanian and Serbian and provides users with the preferred term in each language and its preferred equivalents in other languages (Europa.eu, n.d.).

Other parallel language resources not issued directly by the EU. There are also some parallel language resources as outcomes of private initiatives or of EU-funded projects. Steinberger et al. mention in their last report (2014) following resources:

The Europarl Corpus: more in the next chapter (Steinberger 2014).

The Mulltext project: Aimed at developing standards, tools, corpora and linguistic resources for multiple languages. Currently covering 18 languages
Mulltext-East – a spin-off of Mulltext, developed morpho-syntactic specifications and language resources of 18 central and eastern European languages with English as a hub languageas.

OPUS open parallel corpus collection: a collection of translated texts obtained from the web. Its coverage reached over 150 languages in altogether 5 billion aligned translation units

European Parliament Interpreting Corpus (EPIC): EPIC is a trilingual parallel (English, Spanish, Italian in all combinations) corpus of European Parliament speeches and their corresponding interpretations. It contains 357 speeches with roughly 177 thousand words in video format and their corresponding interpretations in audio format. The corpus has been orthographically transcribed and annotated for scholar purposes in interpreting studies (annotation of paralinguistic features, truncated and mispronounced words) (Russo, Bendazzoli, Monti, Sandrelli, Baroni, Bernardini & Mead, 2011; SSLMIT, 2015).

The Europarl corpus. A special attention in this master thesis should be given to the Europarl corpus which is the subject of our research. The Europarl corpus (EP Corpus), compiled between 1996 and 2011 by Koehn, contains the verbatim minutes of

[25] ISO 25964 – Thesauri and Interoperability with other Vocabularies

debates at the European Parliament. It initially covered 11 languages and had been later extended to 21 official languages with a total up to about 60 million words in one language. The EP corpus is one of the biggest multilingual corpora that are freely available (Zahurul & Mehler, 2012; Steinberger et al. 2014).

The EP corpus was extracted from the website of the European Parliament by using web crawling techniques. The goal of the extraction and processing was to generate sentence aligned text that could be used for training statistical machine translation systems (SMT). The EP corpus has found wide application both in SMT and natural language processing (NLP), whose progress is driven by the availability of natural data The final product, partially supported by the EuroMatrixPlus project funded by the European Commission, comes with the source release of English files, preprocessing tools and sentence aligner, and parallel corpus of 21 European languages paired with the English language (Koehn, 2005). The detailed statistical data about the EP corpus can be found in chapter *Keyword list categorization on page 138*.

For our glossary compilation it is important to note that in the minutes of debates of the European Parliament, all members of the European Parliament (MEPs) usually speak their own native language. These statements are afterwards transcribed, edited and translated into the other official language of the EU. Therefore, the register of the Europarl corpus is either *transcribed spoken language* (speech edited and homogenized by language specialists), or *written to be spoken*, when MEPs read aloud their written discourse (Cartoni, Zufferey & Meyer, 2013). This very nature was the best fit for intended mapping of the spoken parliamentary language with the quantitative corpus-based lexical analysis.

Despite the wide application of Europarl corpus in SMT and LNP, little attention to the corpus has been paid in academic disciplines such as lexicology, politolinguistics and translation studies. Still, there can be found a few other studies demonstrating the usefulness of the EP corpus for linguistics. We will briefly mention some examples of them.

Firstly, Cartoni et al. (2013) customized the Europarl corpus by identifying and adding missing language tags which enabled them to create other language combinations than those offered in the original EP corpus. This customized corpus was used for investigating casual connectives in English and French in both original and translated

texts (Zufferey & Cartoni, 2012), and comparing the influence of various source languages on 2 target languages (Zufferey & Cartoni, 2014).

Secondly, Zahurulof the & Mehler (2012) also customized the EP corpus to use it for investigation of the simplification of translation. They concluded that the generally accepted Type-Token-Ratio is useful for considering the German-English or German-French language pair, but it fails as a universally valid indicator of the tendency of translators to simplify their target texts in other languages.

Thirdly, Stoyokova, Simkova, Majchrakova & Gajdosova (2015) used EP Corpus for detecting and evaluating time expressions in Bulgarian and Slovak. They concluded that the corpus is a reliable source for comparative linguistic research for the respective languages.

These are just examples, not an exhaustive account of possible application of the EuroParl corpus.

To finalize, the EP Corpus represents a rich, reliable and useful source of data for linguistic research. It is only up to the linguistic scholars to harness the full potential of the corpus for their specific needs and research objectives.

Corpus Linguistics

Definition and history. Linguistics has various branches that focus on a certain aspect of language, such as syntax, semantics, phonology or grammar. Traditionally, corpus linguistics was defined as "the branch of linguistics that studies language on the basis of corpora, i.e. bodies of texts assembled in a principled way" (Johansson 1995, p. 19). However, this conception is rather misleading – corpus linguistics is not directly about the study of any particular language aspect as opposed to branches such as phonology, syntax or grammar. Instead, it is rather a set of procedures and methods for studying a language. Corpus linguistics should be thus understood as a broader concept applicable to many aspects of language enquiry that serves to obtain and analyze data qualitatively and quantitatively, rather than a theory of language or even a separate branch of linguistics. Despite not being a linguistic field itself, it does allow us to differentiate between approaches taken to the study of language which creates a series of linguistic areas (e.g. corpus-based syntax as opposed to non-corpus-based syntax) (McEnery-Wilson, 2001).

As in case of many scientific areas, defining the term of corpus linguistic is a rather challenging task. An interesting elaboration on this is offered by Taylor:

> There is no shortage of overt discussion among theorists of what corpus linguistics is or should be. However, at the same time, to the casual observer or new arrival there might also appear to be a bewildering variety of definitions and descriptions (Taylor, 2008, p. 179).

The term Corpus Linguistics is described by McEnery and Wilson simply as "the study of language based on examples from real life language use" (McEnery & Wilson, 2001, p. 1).

A more sophisticated definition of corpus linguistics is provided by Müller and Waibel who define it as "the study of language by means of naturally occurring language sample" (Müller & Waibel, n.d.).

It might be a contentious issue to define corpus linguistics strictly as a methodology or as a linguistic field. Mukherjee (2005) argues that the terms discipline and methodology are not mutually exclusive and draws an analogy with the introduction of microscopes leading to the creation of the discipline of microbiology:

> I would contend that that corpus linguistics represents both a new method (in terms of computer-aided descriptive linguistics) and a new research discipline (in terms of a new approach to language description) (2005, p. 86).

One point upon which all writers agree when defining corpus linguistics is that corpus linguistics is empirical for it examines and draws conclusions from attested language use, rather than from intuitions. This does not mean that intuitions do not play some role in corpus linguistics, but they do not provide the data for analysis, nor supersede the empirical evidence.

Another central aspect of modern corpus linguistics is the use of computers. The pivotal role of computers is corpus linguistics led to a strong association with computer linguistics and even to creating synonyms such as *computer corpus linguistics* or *computer-assisted corpus linguistics.* Computers have enabled researchers to collect, store and manage vast amounts of data quickly and inexpensively and to analyze large amount of language data using specialized computer software with an ease:

> The use of computers "gives us the ability to comprehend and to account for, the contents of corpora in a way which was not dreamed of in the pre-computational era of corpus linguistics" (Leech, 1992, p. 106, as cited in McEnery, 2006, p. 4)

This enabled wide application of corpus linguistics in various other scientific disciplines and in everyday use.

From the historical point of view, corpus linguistics in academia can be divided into three stages: early corpus linguistics (before 1950s), Chomsky's criticism (1950), and modern corpus linguistics (1950-now) (McEnery & Wilson, 2001). Chomsky rightfully criticized corpus linguistics that it cannot make absolute claims that a large corpus can truly represent the real language and thus serve as a source of evidence in linguistic inquire by introducing his concept of language competence and language performance. He understood the corpus as a collection of external utterances that are a result of language performance. In his view, language competence may be affected by various external factors and thus the language performance only poorly mirrors the language competence. There has been a great deal discussion written on this thorny issue, for more information, please see Čermák, 2003.

The way for modern corpus linguistics was paved by the work of Henry Kucera and W. Nelson Francis who performed a computational analysis of present-day American English in 1967 that was used for creating the first corpus-based dictionary. This innovative step set modern corpus linguistics in motion and it is until today considered as one of the most influential studies in the field (Meyer, 2002).

Strengths and weaknesses of corpus analysis. Corpus analysis has certain notable advantages and disadvantages, as suggested by Vander et al. (2011). The advantages are authentic data sources, easily comparable data and replicable results from available corpora. The disadvantages can be seen in methodical risks also applicable to other empirical disciplines, such as false conclusions derived from insufficient representativeness of given data and methodological errors and/or misinterpretation of statistical data.

The authenticity of data sources is an advantage due to the fact that the texts were produced in a largely natural context. Although corpus data might make corpora messy and noisy compared to experimental data, they surpass them in sheer size which would be difficult, if not impossible, to produce in artificial settings.

The second advantage lies in the possibility to directly test, validate or replicate linguistic studies thanks to straightforward statistical data. This leaves little room for vague statements such as “Type X is rather untypical”, or “Y is marginally acceptable”.

The disadvantages, or perils in corpus linguistics arise from incorrect methodology and human error.

Firstly, making claims about representativeness of a corpus might be somehow tricky. One must bear in mind that findings can only be generalized to a larger population only to the extent that the corpus is representative with the targeted population. Unfortunately, some authors are quite happy to generalize more liberally.

Secondly, corpus data as samples from naturally-produced texts are not always suitable for particular research needs. For instance, large available corpora of journalistic data might be unsuited as a general corpus due to a very peculiar register in the discourse (journalistic rules and restrictions about sentence lengths, high information density, omitting unimportant words, changes by editors and typewriters etc.). Such characteristics can undermine the research purpose and result in false statements and observations.

Thirdly, some analysts tend to rely heavily on brief examination of the top most frequent results which can lead to the complete lack of statistical significance testing and to the issue that multifactorial phenomena are studied monofactorially disregarding the nature of complex interaction from other factors (Vander et al. 2011).

Qualitative vs. quantitative corpus analysis. As the terms imply, the qualitative and quantitative corpus analyses represent two different, yet not necessarily incompatible perspectives on corpus data. Quantitative analysis usually precedes and serves as a basis for the qualitative analysis which examines data more carefully based on decisions also derived from the quantitative results. Another viewpoints offered by Schmied (1993) is to see the qualitative analysis as the precursor for the quantitative analysis. This is because the categories for classification must first be identified before linguistic phenomena can be classified and counted.

Quantitative research of a sampled corpus allows for comparison of findings to a larger population based on statistical data. It provides statistically reliable and generalizable results by focuses on classifying linguistic features, counting occurrences

and their frequencies, and deriving complex statistical models in an attempt to explain what is observed (McEnery, 2001).

Qualitative research, on the other hand, describes aspects of usage in the language and provides real-life examples of particular language phenomena. It enables very fine distinctions to be drawn with regard to great ambiguity so much inherent in human language. In qualitative analyses, rare phenomena should receive the same attention as more frequent ones. It can provide greater richness and precision, but its main disadvantage is that the findings cannot be extended to wider populations with the same degree of certainty as in the case of quantitative analysis (McEnery, 2001).

Both approaches are different in scopes, but complementary in results. Corpus linguists can benefit as much as any field from a multi method research, combining both qualitative and quantitative perspectives on the same phenomena (McEnery, 2001). In fact, most scientific research derives the advantage of both approaches and combines them to a certain degree. Corpus linguistics software tools also built to provide both qualitative and quantitative analysis of a given corpora (word frequency lists, keyword lists vs. concordance search, N-grams etc. – explained more in detail in the next chapters).

The presented study also encompasses both qualitative and quantitative analysis. The quantitative analysis as the dominant approach comprises of generating statistical data (wordlist, keyword list, number of occurrences, keyword frequency etc.) that serves as a basis for word selection included in the glossary. The qualitative analysis, on the other hand, looks closely at selected words and investigates their usage in the corpus (collocations, n-grams, concordances etc.).

Corpus-based and corpus driven linguistics. Another notable area where differences emerge between corpus linguists concerns forming and testing linguistic hypothesis. Here we differentiate between corpus-based and corpus-driven language study, as suggested by Tognini-Bonelli (2001).

Corpus-based studies use corpus data in order to "expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" Tognini-Bonelli (2001, p. 65). Corpus-based approach has an existing theory as a starting point and later corrects/revises this theory in the light of corpus evidence. In extreme cases, however, strict adherence to language theory validation may intuitively lead to lower commitment to corpus data as a whole or even to

discarding inconvenient corpus evidence that do not support the formulated pre-corpus theory. Corpus-based linguistics uses corpus annotation and various classifications as a standard procedure by corpus analysis (McEnery & Gabrialatos, 2006).

Corpus-driven approach, on the other hand, rejects the characterization of corpus linguistics as a method and claims instead that the corpus itself should be the sole source of language hypotheses or in other words, it should serve an empirical basis from which lexicographers extract their data and detect linguistic phenomena without prior assumptions and expectations (Tognini-Bonelli (2001, p. 84-5). Corpus driven approach advocates for very large corpora and does not make any serious effort to achieve representativeness, claiming that the representativeness of a large corpus corrects itself when it becomes big enough. This may be true to a certain degree especially in small corpora, but the claim of self-balancing is rather overstated and therefore has been sharply criticized and even disproved in some studies. Next, as opposed to corpus-based studies, corpus annotation is deemed as undesirable in corpus-driven linguistics because it already implies some traces of language theories. This is also a rather rigid point of view because if we took corpus-driven linguistics too far, it would require someone with no educational background related to language use that would make him or her free from preconceived theories. The next contrast to corpus-based linguistics is that corpus-driven linguistics makes no distinction between lexis, syntax, pragmatics, semantics and discourse and uses instead the holistic approach with only one level of language description, namely the functionally complete unit of meaning. In general, corpus-driven approach claims to be a new paradigm within which a whole language can be described (McEnery et al. 2006).
The distinction between corpus-based and corpus driven approach is corpus studies is rather fuzzy. In general, corpus-based approach is often used in a broad sense to describe both approaches (McEnery et al. 2006).

In the presented study, we also combine both the corpus-based and the corpus driven approach. Corpus-based approach can be identified on the first hand in the main hypothesis which we are trying to back with literature review -- namely that glossaries based on word frequency from a specific corpus may be an asset to ESL speakers interested in the particular language area (in this case the political variety of the EU parliamentary language). Next, we make use of other linguistic research outputs such as lemmatization, word frequency lists and stop list with a reasonable belief about their

correctness for conducting the corpus analysis. This makes this corpus study to a high degree dependent on and influenced by other research findings and already postulated language theories. On the other hand, we also apply the corpus-driven approach, mainly in the qualitative analysis. We attempted to look at qualitative data with genuine interest and no bias firstly to identify language patterns typical for selected words and secondly to document them in the glossary (concordance, n-grams, collocations etc). We also did not use annotation tools for we considered that of no use for the corpus analysis. To conclude, the presented corpus study entails both corpus-based and corpus-driven approach, with the first one predominant in the quantitative and the second one predominant in the qualitative analysis.

The next chapter will provide an overview of corpus analysis possible application in various language disciplines. The aim of this chapter is to demonstrate the synergy inter- and transdisciplinarity that arises from harnessing corpus linguistics in a broad spectrum of other language studies.

Corpus linguistics application.

Corpus analysis in everyday use. If we understand corpus analysis in a broad sense as an automated text analysis method, we could presume that it is the cornerstone on which modern digital technology is built. Search algorithms used in programs from web search engines, websites, databanks to simple text processors have to be able to wade through data and deliver search results that may vary in their complexity (various data, not only texts). This process is fully automated (e.g. search bots as automated programs searching the web for predefined queries) and omnipresent in its results. Nowadays we can fully enjoy all benefits of computer advancement without being aware of the complicated processes behind the results. It is difficult to grasp the stunning complexity of digital technology making up our modern life and to realize how much of the modern world is based on statistical analysis. Similarly, understanding corpus analysis and corpus linguistics in a narrow sense in its contribution to linguistics still opens so many possible applications that it is nearly impossible to give an exhaustive account to all of them. In the next chapter, we will try to give a concise overview of possible utilization of corpus linguistics in various linguistics disciplines.

Corpus linguistics and translation studies. It is not an overstatement to say that corpus linguistics has revolutionized translation studies because it played the key role in

creating fundamental principles underpinning statistical machine translation (SMT). SMT utilizes statistical methods borrowed from corpus linguistics to analyze text samples (mainly source texts with respective translations) and makes conclusions based on the results. The first attempts for automating the translation process date back to 1950s, but they were deemed unfit for practical use[26].The first breakthrough, after a quiet period, took place in 1990 when IBM introduced the first SMT model called Candide. The idea behind SMT is that the whole sequence of words can be translated instead word-for-word translation. Candide utilized a word-aligned parallel corpus, but modern SMT tools are also trained to use monolingual data. Thanks to corpus linguistics, SMT is able to harness linguistic data with great efficiency and produced better and more accurate results. The quality has improved exponentially in recent years, also thanks to advancement of large corpora on which SMT is vastly dependent (Hutschins, 2010). Nowadays, we all may use SMT on a fingertip by automatic translation of e-mails, webpages both on PCs and mobile devices or even make use of speech translations. Although SMT is still far from being perfect, the rapid development in SMT sparked by corpus linguistics and statistical methods bear fruit and are fit for everyday use (for instance Linguee, Google Translate for laics on the one hand and CAT[27] tools like Trados, Across, MemQ on the other).

Apart from practical application, corpus linguistics proved useful also in the theoretical translation studies as corpus-based translation studies. Saldanha (2009) points out that corpus linguistics and descriptive translation studies focus on 'attested' language. Corpus linguistics uses and can provide translation studies research with authentic, naturally occurring, texts instead of intuitive, invented, isolated sentences. Both disciplines are interested in performance rather than competence and they study the full range of varieties of language production, including spontaneous, non-edited language use as well as edited, usually written, language. Neither grammaticality, nor translation quality are necessarily prerequisites in this regard. This does not necessarily mean, however, that one of the criteria for compiling a specialized corpus cannot be translation

[26] The US government decided to cancel its funding for machine translation in 1966 based on the ALPAC report (Automated Language Processing Advisory Committee), arguing that machine translation will never produce high quality translations.

[27] CAT – computer assisted translation

quality (for instance if it is to be used as a resource for assessing translations or for translation training), but the selected translation should be ‘attested and real, rather than translations created for the purpose of translation training or translation assessment (Saldanha, 2009).

Saldanha also explains the synergy of combining descriptive translation studies and modern computer-assisted corpus linguistics. She suggests that descriptive translation studies encourage moving away from the traditional comparison of translations against source texts that originally focuses on evaluating degrees of equivalence and faithfulness. The object of a descriptive approach is instead to explain translated texts in their own terms and not as mere reproductions of other works. This requires identifying linguistic patterns that are repeated across large numbers of translations, for which electronic corpora are particularly suitable. Texts and text types are studied comparatively across text corpora both in descriptive translation studies and in corpus linguistics, hence corpus linguistics is a particularly useful in translation studies (Saldanha, 2009). Corpus-based translation studies have also elaborated on other various key aspects such as investigating of representativeness of a corpus, advantages and disadvantages of parallel corpora, the potential benefits of using corpora in translation training and the need to correlate the results to the context (Bendazzoli & Sandrelli, 2009).

Corpus linguistics and interpreting studies. Interpreting, defined as “a form of translation in which first and final rendering in another language is produced on the basis of a one-time presentation of an utterance in a source language” (Pöchhacker, 2004, p. 11), can also utilize corpus linguistics as a method in numerous ways, following the footstep of translation studies. Corpora of interpreted speeches and their transcripts have been used in interpreting studies since their beginning for both research (analysis, observation, documentation, experimentation etc.) and teaching purposes. This area of interpreting research is called corpus-based interpreting studies and it was firstly proposed by Baker (1993) as “the way to analyse features of translated texts as texts in their own right” (Bendazzoli, 2010). Originally, only small and manual corpora were predominantly used, with large interpreting corpora starting to emerge only few years ago (EPIC, ComInDat, Footie, Dirsi etc.). The process of building interpreting corpora remains the same: systematical and purposeful recording of authentic source language and interpreted speeches and their transcription or eventually annotation. However, there

are very few studies which actually make us of corpus linguistics methods and there is much to be explored (for some examples, see Shlesinger, 2009). It is up to IT technicians, computational linguists and interpreting scholars to use their expertise and work together to compile a machine-readable corpus (Bendazzoli & Sandrelli, 2009; Bendazzoli, 2010).

For instance, at the lexical level, Pöchhacker (2004) suggests that research might begin by considering such aspects as word frequency and lexical variability, move on to non-redundant items such as proper names and numbers and pay special attention to semantic phenomena like false cognates in a given language pair, non-standard and culture-bound usage, and even 'creative' or humorous language use. Pöchhacker also points out that corpus-linguistic procedures were lately also applied in text-statistical analysis of verbal features in large quantities of machine-readable text which created a fresh impetus to inquiry into the relative orality of texts in interpreting (e.g. quantitative analysis of the ICSB Corpus for voiced hesitations (Pöchhacker, 1994, 1995).

There is also a number of difficulties arising from compiling a corpus for interpreting studies. Firstly, there is limited access to authentic data because it is difficult to obtain collaboration from conference speakers and organizers for confidentiality reasons on the one hand, and the consent from professional interpreters, who tend to perceive scientific research as attempts to evaluate the quality of their work on the other. Secondly, there is are too many variables that must be controlled to obtain reliable quantitative results. Those variables vary from the type of interpreter-mediated event that determines the role of all participants (conference interpreting vs. liaison interpreting), interpreting mode (simultaneous vs consecutive), variability of speakers (native vs. foreign language used, accent, expertise, expectations etc.) to the written-to-spoken continuum that includes the use of visual information, interpreters' performance, educational background etc. Thirdly, when researchers manage to collect sufficiently homogenous and representative data, they are faced the problem how to transcribe the data – what to transcribe, which conventions to use and how to encode the data to make semi-automated or automated analysis possible. Since there ae no internationally accepted standards, transcripts are usually produced according to the conventions that are consistent with research project objectives (Bendazzoli, 2010). These and other problems associated with corpus-based linguistic studies are illustrated in Pöchhacker:

> Segmenting the flow of speech into punctuation-delimited clauses and sentences, paralinguistic and nonverbal discourse features can be incorporated in machine-readable transcriptions only with great difficulty. Thus, until advances in speech signal detection and electronic text encoding (see Cencini and Aston, 2002) make it easier to overcome the written-language bias of corpus linguistics, studies of the paralinguistic features of orality in interpreting will have to rely on the intensive 'manual' analysis of limited scale corpora, albeit with ever more advanced technological support (Pöchhacker, 2004, p. 140).

The lack of homogenous and representative data resulting from these obstacles does not cause a big hurdle in interpreting teaching or evaluation, but pose a serious problem in testing interpreting theories. There is clearly a call for more corpus-based research in this regard (Pöchhacker, 2004; Bendazzoli & Sandrelli, 2009).

These pessimistic outlooks come from the very nature of specific needs for using corpus linguistics in interpreting studies. Although development of corpus-based interpreting studies (CIS) has been clearly less advanced than the corpus-based studies on written translation (CTS) both in terms of corpora size and, number of studies and pedagogical application (Bendazzoli, 2009), we should bear in mind that the very reason for this dwells in specific obstacles of using corpora in interpreted studies as discussed above. This does not mean, however, that corpus linguistics would hardly find its application in interpreting studies. Just on the contrary, its need and contribution to interpreting studies is eminent and it is reasonable to expect that after meeting specific requirements of interpreting studies with computer advancement, corpus linguistics is about to unveil its full potential also in this discipline.

Corpus linguistics and lexical/terminological studies. In lexicography, empirical data have also been used long before the discipline of corpus linguistics emerged. Some of first corpus methods applications could be seen in work by Samuel Johnson, who illustrated his dictionary with examples from literature, or in the Oxford Dictionary from the 19th Century that comes with citation slips to study and illustrate word usage. Corpus linguistics provided lexicography with powerful tools to analyze and identify words and phrases from many millions of words of text in a snippet of time that would be needed for searching the terms manually. Even though modern lexicography might rely heavily on corpus linguistics, but it must not be forgotten that careful human consideration and

manual analysis cannot be replaced by statistical methods (Bowker & Pearson, 2002). Instead, they can both work in synergy and enhance one another in a number of ways.

Firstly, through combining computer-assisted corpus linguistics and human revision, dictionaries and glossaries can be produced and revised much more quickly and accurately than ever before, thus providing up-to-date information about language. Evidence shows that pre-corpus dictionaries tended to take rare word senses into account, "but missed important, common ones" (Klosa, 2007, p. 111). Nowadays, the evidence found in a large general corpus can help to identify the most common modern meaning of a word, but this must be still treated with caution. Frequency alone is not enough and lexicographers also need to look at the distribution of the word.

Secondly, definitions can be also more complete and precise since a large number of natural examples are examined.

Thirdly, corpus data contains a rich amount of textual information which makes possible to narrow down usages of particular words as being typical of particular region varieties, genres etc. This is especially useful in discourse studies and dialectology.

Fourthly, large constantly-growing corpora that are today automatically compiled from the internet websites enable lexicographers to track new words entering the language, or existing words slightly changing their meanings or to investigate the balance of their use across genres. More and more dictionary websites change from the static "editorial" dictionary to a dynamical platform enhanced by generated corpus-driven content (Lindemann, 2013). Next, finite corpora, on the other hand, also have an important role in lexical studies in the area of quantification. They make possible to rapidly produce reliable frequency counts and to subdivide areas across various dimension according to language varieties in which the words are used. And lastly, corpus linguistics enables to call up word combinations rather than individual words together with the existence of mutual information tools establishing relationships between co-occurring words make threating phrases and collocations more systematical than was previously possible.

Lexicography has to tackle many challenges arising from using corpus linguistics (e.g. failure-to-find fallacy, identifying the core meaning of a word (McEnery, 2001; Mitkov, 2005). Corpus linguistics can be equally beneficial to terminology and terminography as it is to lexicography. Although terminology is different in scope from

lexicography, in essence they both study words forming language's wordstock. Budin sets a clear link between terminography and corpus linguistics:

> Today we could hardly find any terminological activity without using computational methods. Corpus linguistics is a paradigm using computational methods, that enable to do empirical discourse studies and leveraging their quantitative coverage of discourse by many orders of magnitude. Corpus linguistics has become the common denominator and joint methodological basis for a sociolinguistic discourse analysis as for terminological investigation, in particular of a socio-terminological investigation (Budin, 2010, p. 21).

Corpus linguistics and grammar. Grammatical (or syntactic) studies have been along with lexical studies the most frequent types of research that applied corpus linguistics ant it is even believed that corpus linguistics has radically changed grammar research (Conrad, 2000). Corpus analysis is used in syntactical research for both testing grammatical theory on real-life language and for empirical description/inductive generation of theory as 2 main approaches in syntactics. It offers the potential for the representative quantification of a whole language variety and a methodology to process empirical data for the testing of hypotheses. Research using corpus analysis focuses on grammatical frequencies (e.g. various English clause types). (McEnery, 2001). Computer-assisted corpus analysis with parsing tools helps considerably to investigate large amount of quantitative data that would otherwise be too complicated. Similarly, syntactic studies also enhanced corpus linguistics and contributed greatly to other disciplines (e.g. language teaching) by compiling parsing patterns for computer-assisted corpus analysis (Clark, Fox & Lappin, 2010).

Corpus linguistics and lexicogrammar. Lexicogrammar is a level of linguistic structure where lexis and grammar combine into one and are not regarded strictly independent as in the case of syntactic and lexical studies. It incorporates both the formal and functional features of syntactic structure, focusing on the wording of a text (Morley, 2000; Berber Sardinha, 2000). Corpus linguistics is said to have had a clear and major contribution to English corpus linguistics in regard to lexicogrammar. Since English corpus linguists tended view language lexically in order to get lexical answers to what have been traditionally treated as grammatical questions, lexicogrammar was created to offer a unifying perspective on these issues. Recently, there is an increasing number of

studies investigating a grammatical phenomenon on the lexis involved or studies focusing on the collocational behavior with regard to semantic and grammatical properties. However, this brings along higher requirements on annotation tools (category-based methods) which are time consuming to develop as opposed to a raw corpus analyzed through word-based methods (McEnery, Xiao & Yukio, 2006).

Corpus linguistics and speech research. Corpus linguistics has also proven particularly useful in certain areas of speech research that were until recently solely dependent on speeches elicited under artificial conditions. Speech research is also encountered with similar difficulties resulting from orality as in the case of interpreting studies. Although recordings in language labs are of a better audio quality, spoken corpora provide a broad sample of speech extending over a wide selection of variables (gender, age, class, genre etc.), naturalistic speech outside controlled laboratory conditions (enhanced sample authenticity) and prosodic annotations that enable speech researchers to investigate closely at the phonological level (McEnery, 2001; Cole & Hasegawa-Johnson, 2012).

Corpus linguistics and language teaching. Corpus-based linguistics research has provided increasingly clear and accurate description of native and learner languages and has furnished language teaching with new insights into language structure and use. Exploring a corpus facilitates comparing native intuitions with actual use, and shifts learning from prescription to description methods. Large corpora not only make it possible to identify frequent patterns, but also affords enough data to examine rare ones. Apart from downloadable corpora, there is a wide selection of free general or specialized online corpora and corpora tools, not mentioning that the web itself can be used as a rich corpus by language learning (Fagan, 2005; Römer, 2010; Boulton, 2010; Diemer, 2011; Krieger, 2003). From the practical point of view, corpus linguistics has been used widely in calculating the readability of books used by teachers at primary schools (Nagy & Stahl, 2007).

It is believed that introducing corpus linguistics to language teaching could be an asset because it would change the role of students from passive recipients of didactic material to active participants exploring the language, and similarly, the role of teachers from authoritative sources of knowledge to the roles of guides and facilitators. There is a growing notion of using corpus linguistics in language teaching, (Gebrielatos, 2005;

Thomas, 2009). This encompasses the usage of so called experimental learning that is believed to be more effective than classical didactic methods.

Experimental learning involves 4 stages of a so called learning circle: concrete experience, reflective observation, abstract conceptualization and active experimentation (Kolb & Boyatzis, 2001), and it lead to more engagement by students and consequently to better memory retention, as some studies suggest (e.g. Pérez-Sabater et al. 2011). Clearly, it requires proper methodology[28] and commitment both from students and from teachers who need to get familiar with using corpus analysis software. Corpus analysis can provide learners with useful insights into the language and has several advantages that printed sources, dictionaries or learners' intuition lack. Its main advantages lays in the possibility to search easily and quickly for language utterances in context, improving research skills and validating one's intuition on real-life language examples. These advantages apply both to the general and the specialized language (Pearson & Bowker, 2002).

Corpus linguistics in language teaching is regarded as a novel approach, but some teaching scholars still call for some degree of caution because language teaching had allegedly a propensity for marketing and uncritical acceptance of new methods (Gabrielatos, 2005).

Corpus linguistics and discourse studies. For corpus linguistics, a discourse is a totality of texts produced by a community of language users who identify themselves as members of a social group on the basis of the commonality of their world views (Teubert, 2005). Discourse analysts usually examine small numbers of very specialized texts within the investigated discourse in order identify particular social practices. Next to small corpora, discourse analysts tend to make use of large corpora as reference data to discover how far certain features are distinctive in the examined discourse and how far they occur elsewhere in the language as a whole. There is a notable history for usage of corpus analysis methods in the field of discourse analysis, but the amount of corpus-based research is rather small, due to the fact that these fields focus on context that is poorly represented in available large corpora. The amount of conversational data available has

[28] "Not just what to learn, but how to learn it." (Johns, 2002, p. 110 as cited in Boulton 2010, p. 18).

increased in recent years and more research in this field is expected (McEnery, 2001; Weninger, 2010).

Corpus linguistics and cultural studies. Using corpus linguistics in cultural studies is helpful for revealing cultural differences in to countries with the same language by investigating word usage, their frequency and collocations in two comparable corpora (e.g. US vs. UK culture). It can also provide an interesting look into cultural generalities or particularities within one language, for instance by investigating collocations of new words (e.g. identifying collocations for the word metrosexual in gender studies) (Hyland et al. 2012). This application of corpus linguistics started only few decades ago (e.g. Hofland and Johannson, 1982; Leech and Fallon, 1992) as large electronic language corpora became available. It seems to be a promising field of study that may also bring a contribution to language teaching given the fact that effective second language acquisition is to a great extent intertwined with understanding its culture (McEnery, 2001; Cooke, 2007).

Corpus linguistics and social psychology. Social psychologists are required to use naturalistic data that cannot be reproduced in laboratory conditions, while at the same they need to quantify and test their theories on authentic speech utterances rather than rely on qualitative data. Researches have relied on naturally occurring texts such as newspapers, diaries, reports etc. However, the main problem lies in the fact that such texts are in the written mode and social psychologists prefer to analyze spoken language because the most human interaction takes place in the spoken mode. With the advancement of electronic spoken corpora, this was no an obstacle anymore and social psychologist are now able to analyze vast quantitative data with an ease by using various corpus analysis tools (McEnery, 2001).

Corpus linguistics and dialectology. Dialectology as an empirical field of linguistics comparing language varieties tends to concentrate on experiments and less controlled sampling than corpora and is usually focused on vocabulary and pronunciation. However, with the increase of dialect corpora, new research possibilities emerge in which corpus linguistics can assist in and expand the knowledge of dialectology. It has already proven effective in testing two theories of language variation: Quirk et al' common core hypothesis and Braj Kachru's "the Circles of English" that postulates many unique world ´Englishes´ as varieties. (McEnery, 2001; Anderwald & Szmrecsanyi, 2008).

Corpus linguistics and sociolinguistics. Sociolinguistics is also an empirical discipline and it studies the relationship between language and society, such as social functions of language, speaking differently in different social contexts. (Holmes, 2013). It primarily uses the collection of research-specific data rather than more general corpora. Yet, corpus can provide sociolinguistics with a representative sample of naturalistic data that can be quantified and thus open new areas for research. Some examples of corpus analysis usage in sociolinguistics are examination of masculine bias In American and British English by Kjellmer (1986) or Granham's et al. (1981) study of speech errors occurrence in natural conversational English. These studies could fully harness the potential of corpus analysis within the field of psycholinguistics. Despite the fact that this field comprises various areas for experimentation (from measuring the length of time for positioning a semantic boundary in reading to eye movements change), it is believed that corpus linguistics will prove its value to sociolinguistics in the near future (McEnery, 2001).

Corpus linguistics application: conclusion. Corpus linguistics has also been used to some degree in semantics, stylistics, and historical linguistics.

As we could see, corpus linguistics proves to be useful in various linguistic fields. The particular need for corpus linguistics explains Léon (2005):

> …what is called 'Corpus Linguistics' covers various heterogeneous fields ranging from lexicography, descriptive linguistics, applied linguistics – language teaching or Natural Language Processing – to domains where corpora are needed because introspection cannot be used, such as studies of language variation, dialect, register and style, or diachronic studies (2005, p. 36).

Through enlisting a broad, yet not exhaustive application of corpus linguistics, the authors aim to put the case that the results of corpus analysis may be in the same way of an asset for various end users with different fields of expertise. In this case, the output of the corpus analysis is a set of glossaries of mostly used terms of the Europarl corpus that may be useful for students of interpreting and translation studies, political sciences, language enthusiasts and employees of national or international institutions alike. In the next chapter, the authors will attempt to identify and discuss possible usage of the EU parliamentary glossary based on the Europarl corpus analysis.

Corpus analysis software. As before mentioned on page 84, it was the computer advancement that opened a new avenue in research in corpus linguistics. This computer advancement was reflected mainly in compiling new corpora and development of new software. Corpus analysis software is nowadays an indispensable tool for corpus analysis. In our thesis we can barely touch the tip of the iceberg dealing with this topic. However, the purpose of this chapter is not to give an exhaustive account of available corpus analysis software, but rather to provide an overview that the reader can use as a springboard for further investigation into the subject.

Corpus linguistics software is mainly used by three types of users: lexicographers, linguistics researchers and students, and language teachers and learners. Translators are also said to use corpus analysis software, but they tend to be often neglected as a user category and their specific needs are rarely kept in mind by further development (Jääskeläinen & Mauranen, 2000). Nevertheless, in general, the available corpus analysis tools generally fit the specific needs of translators, even though there is still much room for improvement. It is reasonable to expect that first tailor-made translators' corpus analysis software could lead to a large adaptation of corpus analysis software in the translation industry (Wilkinson, 2011).

History of corpus linguistics software. Computer-assisted corpus linguistics has gone through different stages that can be reflected in 4 generations of corpus linguistics software.

The first generation appeared in the 1960´s and 1970´s. They ran on mainframe computers and were able to perform only simple tasks such as word counts and concordance search. The processing speed was very slow to today's standards (e.g. 1000 lines of poetry took 4 minutes to process).

The second generation consisted (1980´s / 1990´s) was also limited in functions and speed, but could be run on personal computers.

The third generation includes corpus analysis software further described in the next chapter. It began appearing in 1990´s and is still being improved and developed today. Software of this generation is able to perform complex tasks, offers a user-friendly interface, and unlike the previous generation, it also supports other languages than English and different encoding outside of the ANCI character set.

The fourth generation represents web-based corpus analysis interfaces that enable users to search through vast corpora that are automatically compiled from web. This compensates the limitation of the 3rd generation that was technically appropriate to process not more than 100 million words. While the 3rd generation suits the needs for specialized corpora, the 4th generation can fully harness the technological advancement for a considerably bigger quantitative corpus analysis of automatically compiled general corpora. Some examples of the 4th generation software are Wmatrix, corpus.byu.edu, SketchEngine and Google Ngramm (McEnery & Hardie, 2000).

Typology of corpus linguistics software. Corpus linguistics software represents the tool that can work with electronic corpora. They essentially consist of two main functions: a feature for generating wordlists and a feature for generating concordances (Bowker & Pearson 2002), but the majority of nowadays corpus linguistic software offer a variety of other functions. Corpus analysis software can be categorized according to its usage into several categories.

Firstly, we distinguish between *computer-based* (stand-alone) *tools* and *online tools.* Computer-based tools require that both software and the corpora are installed and stored on the user's PC. The most popular computer-based tools are WordSmith, MonoConc, and AntConc as a freeware alternative. Online corpus tools, on the other hand, allow users to access the corpus/corpora from any computer using online interface. In this case, both corpora and software are stored on server. Some examples of online software tools are The Sketch engine, Intellitext, Corpus-Eye or the Corpora at BYU[29] (Granger & Paquot, 2012; Dillon, 2015). It is important to note that with using advanced search function of search engines, or search sites especially customized for corpus searches, the web can be also treated as a one big corpus. One must only know how to search it effectively.

Secondly, software linguistic tools can be categorized whether they can be used only with a particular corpus (corpus-related tools), or independently (corpus-independent tools). Some corpus tools were designed as a part of a specific corpus project

[29] Detailed information about corpus linguistics resources can be found here:
http://courses.washington.edu/englhtml/engl560/corplingresources.htm
http://www.corpora4learning.net/resources/materials.html

or for a specific purpose. Examples of corpus-related tools are XAIRA and BNCWeb, two highly-specification interfaces designed to access the British National Corpus (BNC). Corpus-independent tools can ben on the other hand used for analyzing any corpus within the range of supported formats (MonoConc, AntConc, WordSmith etc.) (Granger & Paquot, 2012).

Thirdly, a differentiation can be made between corpus software for using prepared corpora and software that search web as a corpus. The majority of corpus software are used to access a corpus that was compiled with specific linguistic research objective in mind. However, web also offers a plethora of linguistic data, thus can be treated as a vast corpus with very large quantities of texts in many languages. Consequently, search engines can be viewed as corpus tools that also response to a query. Although they are not designed specifically for linguistic purposes, they can prove to be useful by applying advanced search options. Some specific tools have been developed which sit between the search engine and the user. They are especially customized for linguistic searches and offer options similar to standard corpus linguistic tools. Those are called web concordancers. The leading system is WebCorp (Barlow, n.d.; Granger & Paquot, 2012).

Fourthly, with increasing diversity of corpus software users, it is important to differentiate between simple and advanced tools. Simple corpus tools offer only basic features such as concordancing, collocations, wordlist and keyword analysis or n-grams that serve users who intend to perform less complex tasks and analyses. Examples of simple corpus tools are AntConc and Conc Easy. Advanced corpus tools (e.g. Sketch Engine, Xaira etc.) are rather designed for users performing complex analyses like lexicographers who deploy advanced functions such as CQL searchers or word sketches (Granger & Paquot, 2012).

Corpus analysis software

Due to the fact that the topic of corpus analysis software is a very broad topic encompassing a great variety of specialized corpus analysis tools, this chapter will focus only on corpus analysis software used in this thesis. This includes WordSmith, AntConc and ParaConc which are only examples of many software packages available to carry out corpus linguistic research.

WordSmith. WordSmith is a software tool developed by Mike Scott since 1996. It has become one of the most popular corpus linguistics software and presumably the

most widely used (Nesselhauf, 2005; Römmer & Wulff, 2010; Granger & Paquot, 2012; Müller & Waibel, n.d.). It offers a very comprehensive package of various corpus analysis tools, such us wordlist and keyword list generator, concord search with KWIC results (Key Word in Context), and utilities for sorting tags and file management. The new version of WordSmith 6 comes with other powerful tools like word clouds and chargrams among others (Scott, 2015).

AntConc. AntConc is a very popular and reliable multi-platform alternative to commercial programs. It was released in 2002 by Laurence Anthony and since then further developed for specific use in the classroom. The freeware includes the most common functions such as word and keyword frequency generators, concordance with KWIC, tools for cluster and lexical bundle analysis, and the plot tool analyzing the word distribution throughout the text. Unlike WordSmith containing three separate programs for word list and keyword list generator and a concordance, in AntConc all tools are integrated within the program itself and is installation free. AntConc can be extremely effective in working with small specialized corpora, but the author states its limitations in fast processing of large corpora and handling annotated data in HTML/XML format (Anthony, 2005).

ParaConc. ParaConc is a commercial bilingual or multilingual concordance program released in 1999 by Michael Barlow. It is particularly useful for contrastive analyses, language learning and translation studies/training because it can display aligned translations to searches in bilingual and multilingual corpora in a KWIC layout. The successful search depends on the presence of alignment segments in corpora. ParaConc can also identify so called *hot words* that display the rank list of most common translations of a particular word found in the corpora. It also offers frequency statistics, e.g. corpus and collocates frequency and more advanced search options.

Discussion. Although corpus linguistics as an empirical method should provide great replicability and verifiability of research results, practical experience shows that this is not always the case. Different software can yield different results even by using the same settings. The reason behind these discrepancies are complex, but the most common one worth mentioning is that different software deals with words differently. For instance, Anthony points out (2013) that AntConc counts shortened words with apostrophes (e.g we´re) as two words in contrast to MonoConc and WordSmith that treat it mistakenly as

one single word. Corpus linguists should be aware that the replicability and verifiability of their studies depends not only on the standardized settings, but also on the software tools used. It is important to choose the right software tool based on research scope and the characteristics of investigated corpus. Anthony draws an interesting analogy between corpus linguistics and astronomy that has a wide range of research scope from individual planets and their trajectories to observing solar systems and whole galaxies which applies the most appropriate tool for each research scope (Anthony, 2013).

Main features and methods used in corpus analysis software. In this chapter, we will briefly introduce main corpus linguistic tools used in corpus analysis. Corpus linguists have access to a range of techniques that can be implemented in the analysis of a text. The quantitative corpus analysis of a raw corpus includes mainly computing word frequency and keyword frequency lists which will be given more attention due to the scope of this thesis. These are usually compared to other results obtained from other corpora, usually larger and representing a certain norm. The qualitative corpus analysis, on the other hand, includes more specific searches of a concrete word in context. This includes application of corpus linguistics tools such as concordance and collocation search, n-grams or plot search. The world list and the keyword list, as the main tools used in our corpus analysis, will be described more in detail than the other corpus analysis tools.

Word frequency lists. Generating a word frequency list (also referred to as a wordlist or frequency-sorted word lists) is one of the easiest and fastest ways to analyze a corpus by using corpus linguistics software and usually the first thing the linguists tend to examine. However, there is much to know behind a seemingly easy task consisting of few clicks in a program.

A wordlist is the basic foundation for corpus analysis and has long been part of the standard methodology for exploiting corpora. Sinclair (1991, p. 30) outlined that "Anyone studying a text is likely to need to know how often each different word form occurs in it".

Archer defines frequency word analysis as "construction of word lists, using automatic computational techniques, which can then be analyzed in a number of ways, depending on one's interest(s). (Archer, 2009, p. 2).

In order to understand the purpose of wordlist, we should firstly elaborate on the meaning of word frequency. Word frequency is a) placing of occurrence statistical data on language, b) instantiation of the claim that "linguistics is the scientific study of language" (all theoretical claims can be demonstrated on statistical results) c) a promise for precision and objectivity in linguistics although the outcome might also be imprecision and relativity[30], d) statistical information that needs interpretation through contextualization whence the relativity and comparison, e) is not a science but a methodology, which lends itself to replicability. (Archer, 2009).

Frequency word lists can be used for a variety of purposes, for example for determining whether a collection of texts are written by the same author, whether the most frequent words suggest a potential meaningful pattern, or for language teaching (Archer, 2009; Smith, Kilgarriff & Sommers, 2008). Scott, the creator of the WordSmith corpus analysis tools also mentions other some corpus linguistics reasons such as studying the type of vocabulary used, identifying common word clusters, comparing the frequency of a word in a different text files or across genres, comparing the frequencies of a cognate words or translation equivalents between different languages and getting a concordance for words in the list (Scott, 2015). Many dictionaries also directly or indirectly make use of statistical information about word frequency (Rayson, 2003).

Though the field of corpus linguistic seems to be relatively new, the review of literature suggests that the first large word frequency count was published by Keading in 1897/98, in which he conducted a manual lexicometrical research of a corpus with approximately 11 million running words. It was an ambitious project that took several years and required considerable workforce. Similar research was reported to continue throughout the 20th century1 before computer-assisted corpus linguistics was introduced (Graham, 2008).

[30] This does not happen because words would be miscounted, but because of how the frequencies relate to use in the English language as a whole. A word might be identified with a high frequency not because of its wide usage in the language as a whole, but because it appears frequently in a much smaller number of texts within the corpora. To correct this, the range of dispersion statistics should be applied, but it still leaves room for false interpretation of statistical data (Rayson, 2003).

The early research into word frequency demonstrates particular effort not to merely count words, but to categorize them semantically and to give some practical value to the created word lists (e.g. West's General Service List in 1953 for ESL learners). After the advancement of computer technology and following increasing availability of corpora, corpus linguistics research has become incomparably faster and easier. This development has spurred a great interest in the venerable topic of word frequency. However, the quality of research was little affected by the quantity caused by the vast increase in popularity in the early lexicometry (Graham, 2008). Preller (1967) called for more valid, up-to-date lists compiled by more objective procedures which implies the lack of standardized methodology which might have resulted in inconsistent methodology and consequently inaccurate study results. Bearing this peril in mind, we strived for the highest achievable accuracy in generating word lists of the Europarl corpus and for this reason we applied the most standardized methods, as mentioned in the practical part of this thesis.

A wordlist can be arranged alphabetically, in order of the first occurrence or in frequency order and it can contain information on the word rank, the word frequency, the total number of running words (tokens) and total number of word types. If necessary, a wordlist can be abridged by applying a *stop list* which is a control list that excludes from the wordlist all words to be generated included in the stop list. A stop list is used for filtering out undesired words such as function words (Bianchi, 2012).

In computer assisted corpus linguistics it is usually stored electronically in some of the various wordlists formats (e.g. *.lst in WordSmith software). Alternatively, it can also be stored in a plain *.txt file or in an excel table. The majority of corpus linguistics software supports both import and export of above mentioned format types. There is a great availability of various downloadable wordlists, many of which are freely available (e.g. the BNC wordlists).

The wordlist may include following information for each word, but its main function is to give statistical information about words used in the corpus. For illustration a wordlist generated in WordSmith 6 consists of the following categories for each word: word rank, number of all occurrences, its frequency as a percent of all the running word in the corpus, the number of texts in which the word appears and the percentage of texts in which the word appears and all lemmas found for the respective word. Wordlists can

be also compiled from texts within a particular language variety, for instance computer science or psychiatry, or further classified and categorized by POS. A wordlist without lemmatization and POS-tagging provides rough information that do not take into account issues such as polysemy, homonymy and different word classes. Consequently, untagged entries must be manually checked for differentiation in context (e.g. a word *bank* as a *bank of the river* or as *a financial institution*) (Francis & Kučera, 1982; Leech, 2001; Rayson, 2003; Bianchi, 2012). Generating a wordlist is a prerequisite for creating a keyword list.

Keyword list. A keyword list is a very popular statistical procedure and mainly used in corpus linguistics. The keyword list is a result of comparing two word frequency lists using computational statistical techniques (Archer, 2009).

The comparison meant above is called keywords analysis. It compares the frequencies of each word in the study and the reference corpus and it shows an unusual high or low word frequency of an investigated corpus compared to a certain norm (Berber-Sardinha, 2000). It gives the researcher an insight into similarities and differences of both corpora and answer questions such as what the topics discussed in the investigated corpus are.

To understand the purpose of a keyword list, we should firstly define the term keyword as it is used in corpus linguistics. A keyword in this sense is something different from a *topic word* or an *important word*, which are generally used as synonyms for keywords, because in corpus linguistics, the keyness of a keyword is defined by frequency. A word is a keyword "if it occurs in a text at least as many times as a user has specified as a minimum frequency, and its frequency in the text when compared with its frequency in a reference corpus is such that its statistical probability as computed by an appropriate procedure" (Baker, 2004, p. 346). Another widely accepted definition was introduced by Scott: "A key word may be defined as a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency by comparison with a reference corpus of some kind" (Scott, 1997, p. 236). This also implies that a word can have an unusual low frequency compared to a reference corpus. Such words are called *negative keywords*. Paquot (2009) proposes a modest amendment to the definition of a keyword:

> Keywords of a specific corpus are lexical items that are evenly distributed across its component parts (corpus sections or texts) and display a higher frequency and a wider range than in a reference corpus of some kind (Paquot, 2009, p. 19).

Earlier writing outside the field of corpus linguistics referring to keywords focused intuitively on words that they believed embodied important concepts reflecting societal or cultural concerns, also referred to as distinctive words. This traces back to Firth who in 1935 introduced the concept of focal and pivotal words or to Williams book on keywords. It is important to note that those were not always single words, but also phrases or collocations, together bearing a meaning (Rayson, 2003). But even in the computer linguistics, keywords must not always be a result of computational statistical analysis. For instance, Stubbs introduces the term cultural keywords as words which capture important social and political facts about a community" (1996: 172). He puts the case for manual identification through intuition with relying on a systematic method for searching. As far as cultural keywords are concerned, there is also a broader definition of a keyword that is generally shared outside corpus linguistics. The definition comes from Williams, who paved the way to a research tradition in cultural analysis. He defines keywords as "significant, binding words in certain activities and their interpretation; they are significant, indicative words in certain forms of thought" (Williams, 1964, p. 13).

This definition could lead us to a broad generalization that all most frequent words excluding functional words are keywords for a certain text. Still, as mentioned above, keywords are in corpus linguistics generally understood as words identified by the keyword analysis in a corpus linguistics software. There is no claim that statistically identified keywords would match those selected by human readers. This does not mean, however, either manual or automated approach would be inferior to the other and that both could not complement each other. Rayson points out the usefulness of the keyword analysis in corpus linguistics, stating that the linguistic features worthy of microscopic analysis requiring a careful human consideration can be suggested by a macroscopic study of a keyword analysis, rather than by intuition or previous research studies (Rayson, 2003). On the premise of such data-driven investigation is in the end built the hypothesis of this master thesis.

There are three types of words usually identified as keywords in statistical keyword analysis. Firstly, proper nouns, secondly keywords that humans would also

identify as keywords because they are indicators of *aboutness* of a particular text, and finally, high frequency words such as prepositions, functional words etc. which might be an indicator of style, rather than aboutness (Baker, 2004).

Problems with generating keywords. First, one practical problem arises from the number of statistically identified keywords which is usually higher than the amount that a researcher can analyze and process. This can be solved by reducing the number of keywords to a certain limit (e.g. ½+1) or by obtaining a significant subset of keywords, as suggested by Berber Sardinha (1999).

The second problem raises from the selection of the investigation and reference corpus. It must be taken into account that all corpora are different in a multitude of ways. For instance, one might be interested in examining a keyword list because of one particular dimension of difference, such as a difference of genre, domain or region. However, the keyword list might be dominated by other differences in which the researcher is not interested. Keyword lists tend to work best if the corpora are very well matched in all regards but the one in question. This leads us to an important issue of preparing and/or selecting the investigated corpus and the reference corpus. The general rule is that the more heterogeneous the corpus, the less evident it is to identify exactly what a keyword reveals about the research corpus. Another problem might arise when the reference corpus is not big enough. Berber-Sardinha (1999) concluded from his research that a reference corpus five times bigger than the investigated corpus is sufficient for proper keyword analysis. Reference corpus of a bigger size makes no difference, but a smaller reference corpus could cause that some keywords would be left out from the keyword analysis (Kilgarriff, 2009; Paquot, 2009).

Thirdly, if a word is the topic for one text in the corpus, it may be used in that text with high frequency and consequently identified as a keyword, but happens to appear only in one single text or just a few texts. Such a word is called a local keyword as opposed to a global keyword. Statistical measures are computed on the basis of absolute frequencies and cannot account for the fact that corpora inherently vary internally and consequently cannot distinguish between global and local keywords (Paquot, 2009). This burstiness should be treated by applying the range of dispersion or other statistical techniques (Kilgarriff, 2009; Gries, 2008).

The fourth problem is purely statistical. It is not possible to divide by zero, therefore it is not clear what to do about words which are present in the investigated corpus, but absent in the general reference corpus. There is a vivid debate on this matter and corpus linguistics have to borrow some solutions from statistics. The simplest solution is to add number 1 to all frequencies to avoid counting from zero, but there are more sophisticated methods to tackle this problem (Kilgarriff, 2009).

The fifth problem also implies mathematics. Any ratio in which keywords occur more frequently in the investigated corpus than in the reference corpus can be relative and tricky. For instance, if we find 10 times more occurrences of a keyword in the investigated corpus than in the reference corpus, e.g. 100 vs. 10 occurrences, we can assume that it is normal. However, if we happen to find words with frequency per million of 10 thousands occurrences in the investigated corpus compared to only 1000 in the reference corpus, the same ratio would result in a striking difference worth investigating. Therefore, the word ratio should be interpreted individually on many factors involved (Kilgarriff, 2009).

Keywordlist analysis. A key word analysis by default usually involves at least two wordlist files. Typically, one will be the generated word list from the examined corpus (the one under consideration), and the other will be generated from a reference corpora. Modern corpus linguistic tools can also conduct multiple comparisons of batch keywords and wordlists files. Optionally, the researcher may choose a stop list in case he or she wants to weed out some specific words, for instance the commonest words such as ‘the’ and ‘of’. The researcher may also set a minimum frequency threshold, maximal calculating frequency, max. wanted words, max. p-value and the statistical test (log-likelihood or chi-square, will be explained more in detail in the next chapter).

The final output of a keyword analysis is the keyword list identifying 2 types of keywords: positive and negative keywords. Positive keywords are those which are unusually frequent compared to a reference corpus or a wordlist. Negative keywords, on the other hand, are those which are unusually infrequent in the target corpus or a wordlist (Rayson, 2003).

The keyword list is displayed in a table with words, their rank, the keyness rank and additional statistic information (e.g. in WordSmith the frequency in the source text, percentabe, representing the frequency, the frequency in the reference corpus, the

reference corpus frequency as percentage. From that time on, it is up to the researcher and how much he or she can infer from the statistical data provided depending on the examination scope, experience and knowledge, and particularities of the investigated language (Gabrialatos, 2013).

Statistical tests in the keyword analysis. As already mentioned, the keyword list is a result of statistical comparison of the words' keyness. "The keyness of a keyword represents the value of log-likelihood or Chi-square statistics; in other words it provides an indicator of a keyword's importance as a content descriptor for the appeal. The significance (p value) represents the probability that this keyness is accidental" (Biber, Connor, Upton, Anthondy & Gladkov, 2007, p. 138). There is a vivid debate about which statistical test is the most appropriate for a keyword analysis. There is no exact answer to this question because it depends on many factors, e.g. the corpus size or the examination scope. Paquot (2009) also points out that the results of a keyword analysis are largely dependent on a number of arbitrary cut-off points such as the probability threshold, a minimum frequency, a minimum number of texts in which a keyword appears or on the minimum coefficient of dispersion. An interesting discussion about comparing various statistical techniques offer for instance Rayson 2003, Paquot 2009, and Gabrielatos 2013. For the purpose of this master thesis, we will not dwell on details of statistical computing, but only briefly mention the most common statistical tests.

Chi-squared: introduced by Pearson in 1904 to test the independence of two variables. It is applicable to a general two-dimensional contingency table and has been used broadly in corpus linguistics before introduction of the Dunning's log-likelihood test (Rayson, 2003; Paquot, 2009). The main problem of using the chi-squared test in corpus linguistics lays in the fact that it computes on samples (two texts, two sets of text, etc.) that are randomly drawn from the same population and a random and even distribution of words does not apply to languages which are vastly influenced by many factors (topic, register, style etc.) (Sardinha, 1999).

Dunning's log-likelihood: introduced by the author as the solution to the problem of chi-squared test that takes a normal distribution of words throughout texts as granted. This test uses a parametric analysis based on the binominal or multinominal distribution as a better alternative for smaller texts and is based on a contingency table. It compares the observed frequencies with expected frequencies by the null hypothesis. Log likelihood

test is preferable to Pearson's chi-squared test in general" and is the most widely used test in corpus linguistics (Williams, 1976, as cited in Rayson, 2003, p. 49).

T-test: The T-test looks at the difference between the means from 2 different groups and evaluates the significance of such differences. This however, similarly as in the chi-square test, can deliver valid results only if the data is normally distributed which is not the case for word count in general (Paquot, 2009).

Wilcoxon-Mann-Whitney test: the WMW test is regarded as the non-parametric equivalent of the t-test for two independent samples. It differs slightly in the null hypothesis and is computed on ranked scores rather than word frequencies. The main criticism of the WMW test in corpus linguistics accounts for ignoring the actual frequency of occurrences. This means discarding most of the evidence researchers have about the distribution of words (Rayson, 2003; Paquot, 2009).

The latter two tests include the T-test and Wilcoxon-Mann-Witney test, but these have been only experimentally used for keywords extraction. Paquot (2009) compared these two with the log-likelihood test that is mostly preferred by corpus linguists. He concluded that they are viable options for the keyword analysis and calls for implementing these tests as alternatives to log-likelihood and chi-squared tests in corpus linguistics tools.

Matrix: is a rather a software tool than a statistical test, but its novel approach makes it noteworthy in this discussion. The Matrix was developed by Rayson (2003) and it is based on a macroscopic analysis to inform about the microscopic level. The Matrix integrates the log-likelihood test and the comparison of corpora with word-class and lexical semantic tagging which, is according to the author, vital in defining a practical data-driven approach. None of the corpus linguistic software has until present combined both annotation-awareness capability with the comparison of frequency lists like the one offered by the Matrix method. The author claims its main limitation is the language and size of corpus that can be processed (Rayson, 2003). For this very reason we decided not to use it in our Europarl corpus analysis for which a considerably large corpus was used. This leaves ample room for further research that can yield interesting comparisons. Matrix can be used for identifying the differences between the corpora and is shown to have applications in studying social differentiation in the use of English vocabulary, profiling of learner English and document analysis (Rayson, 2003).

It is important to note that most corpus linguistics software offers only two types of statistical tests: the log-likelihood and chi-squared. It is questionable whether log-likelihood or chi-square tests are always the best statistical measure to identify the words that are particularly characteristic of a corpus (Paquot, 2009). One must also bear in mind that there can be no reliability on statistical tests in absolute terms and no test is exempt from drawbacks (Blanchi, 2012).

Keyword analysis has been used in a large number of corpus linguistics research papers. We can only give a brief mention to some of them for illustration of application for keyword analysis.

To gain descriptive accounts of particular genres, Tribble derived a keyword list from comparing a corpus of romantic fiction with a general corpus and finds evidence to suggest that spoken language features in romantic fiction such as more frequent first and second person pronouns and fewer complex noun phrases (Tribble, 2000).

In discourse studies, Fairclough compares a corpus of „New Labour“ with a corpus of older Labour texts to analyze how Labour's ideological stance had changed over time to stress business interests and competition. This could be inferred from emerging keywords in the new corpus such as *partnership, deliver, deal, new* and *business* (Faairclough, 2000).

In vocabulary studies, Rayson, Leech & Hodges (1997) conducted selective quantitative analyses of the demographically sampled spoken English component of the BNC and compared two Labour corpuses of different age to demonstrate the discourse shift in the political language over the time by using his own statistical tool called Matrix (Rayson, 2003).

Keywords have been used in many other fields than in linguistics. They proved extremely useful in information retrieval and automatic abstracting. From the internet search it can be also inferred that the use of keywords is also a basis for web searches and the CEO ranking of a webpage in search engines.

In conclusion, it would be also appropriate to illustrate the difference between a word frequency list and a keyword list. Baker draws an interesting comparison, saying that a keyword list "gives a measure of saliency, whereas a simple word list only provides frequency" (2006, p. 123). Keywords are an extremely rapid and useful way of directing researchers to text elements that are unusually frequent or infrequent, helping to remove

researcher bias and paving the way for more complex analyses of linguistic phenomena. However, one must bear in mind that a keyword list provides the researcher only with language patterns which require interpretation in order to answer specific research questions. A keyword analysis may lead the research to overplay differences rather than similarities, and it draws the attention to differences at the lexical level, whereas the semantic, grammatical or pragmatic differences remain unnoticed (Baker, 2004).

Concordance search. A concordance is in the Collins Cobuild English Language Dictionary (1987) defines as: "an alphabetical list of the words in a book or a set of books which also says where each word can be found and often how it is used". Sinclair (1991, p. 32) claims that "a concordance is a collection of the occurrences of a word-form, each in its own textual environment". Concordance programs or concordancers are then understood as programs capable of such searches. Concordancers produce a list of linguistic phenomena (usually words) located in a corpus, displayed in the center and shown within the context. If the corpus is lemmatized and tagged for grammatical categories, concordance search can utilize this information to provide more advance search option (Bianchi, 2012) This display mode is also known as key words in context or KWIC (Rayson, 2003). Although concordancers build the core of computer-assisted corpus linguistics, they can be traced back to the pre-computer era. This required immense manual effort and man labor. Some examples are concordance efforts of the Holy Bible in the middle age performed by hundreds of monks (Kezhen, 2015), and a complete concordance to Shakespeare in 1881 by Cowden-Clarke which took 16 years to finish (Rayson, 2003). Interestingly, modern concordancing was born from the vision of Robert Bosa, a Jesuit priest and Thomas J. Watson Sr, the pioneer of the modern computing31.

[31] Bosa had a scholarly interest in linguistic examination of the medieval philosopher St Tomas Aquinas. He is said to be the first person to produce a machine-readable corpus (10 000 sentences). Bosa wrote one sentence per card and indexed those containing the word *in*. However, he found out soon that he needed to improve the means he used for searching if he was ever to study anything else but the word *in.* Therefore, he went to see the IBM CEO Thomas Watson who suggested that he transfers the sentences to punch cards that he allowed early IBM computers to search through and retrieve searches from the corpus on a word-by-word basis (McEnery & Wilson, 2001).

Concordancing is very useful in both monolingual and multilingual corpora because it gives an important insight into the context and it is useful for lexicologist, language teachers and translators alike (Rayson, 2003; Wilkinson, 2011; Lamy & Klarskov Mortensen, 2012; Kezhen, 2015).

Collocations / n-grams. Linguists argue that not a single word, but word phrases constitute the true vocabulary of a language (Teubert, 2004). The basic and famous maxim provided by Firth (1957), one of the most prominent British linguists, is "You shall know a word by the company it keeps!" With this axiom, he points out that some words are more likely to be found together and to form common meaning than the others. Firth coined for such habitual word combinations the term collocations. It is expected that phrases as a whole carry more information than the sum of its individual components (Wang, McCallum & Wei, 2007).

Collocations can be seen as a result of language evolution. They share 3 main characteristics: Firstly, their meaning cannot be directly derived from the meaning of its parts (non-compositionality, e.g. *kick the bucket*), secondly, components of a collocation cannot be freely substituted with words bearing similar meaning (non-substitutability, e.g. *yellow wine* instead of *white wine*), and thirdly, they cannot be freely modified with additional words or through grammatical transformations (non-modifiability, e.g. *apple of your eye* cannot be changed to *apple of your beautiful eye*). Additionally, they exhibit some commonalities, such as arbitrariness (it is not clear why *pick up* means learn quickly, domain dependency (e.g. *interest rate*) and they tend to have recurrent and cohesive lexical clusters which means that the presence of one component strongly suggest the rest of the collocation (United could imply Kingdom or Nations, but less likely united tribes despite the fact that in some context it would be quite correct) (Anagnostou & Weir, 2007).

Simply speaking, collocations represent a way of expressing ideas in a language. This poses a great difficulty for second language speakers because their language output can never sound natural unless they use the right collocations.

The main characteristics of collocations lead us to imply their definition. Collocations are in a broader sense understood and named with various terms, such as

arbitrariness (multi-word expressions, multi-word units, n-grams[32], idioms and co-occurrences even though some distinctions between those are to be made. A satisfactory general definition for collocations was given by Choueka (1988):

> A collocation is a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components.

Semantics offers classification of collocations according to their meaning, in which two types are distinguished: lexical collocations and empirical collocations. Lexical collocations represent a series of more or less transparently fixed expressions, ranging from idiomatic expressions/set phrases (e.g. a school of fish) to multiword expressions (e.g. credit card). The term empirical collocations refers to the general fact that some words tends to appear more frequently than others in the same linguistic environment (Bianchi, 2012).

Bianchi also summarized different perspectives for collocations that helps in determining the importance of collocation in corpus linguistics:

> ... From a textual point of view, "collocation is the occurrence of two or more words within a short space of each other in a text" (Sinclair 1991, p. 170). From a statistical point of view, it is 'the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, pp 6). Finally, from a psychological or associative perspective, collocation is the expectations (or 'expectancies', in Firthian terms) that native speakers have of encountering a given word in the same environment as another one (Leech, 1974) (Bianchi, 2012, p. 49).

In corpus linguistics, we understand collocations as words which occur in the neighborhood of a search word (Scott, 2015). Statistical measurement of collocations are very reliable (see Anagnostou & Weir, 2007), and therefore corpus linguistics is very helpful by computational lexicography as an important part of dictionary compilation,

[32] *N* stands for the searched word that appear in the word cluster of an n-gram.

training machine translation[33] and natural language generation, language teaching and translation training alike, as already mentioned in previous chapters. Collocations is a term prevalently used in lexicology due to its meaning, but from the statistical point of view it is also advisable to use the term *n-grams.* N-grams are in computer linguistics words that are expected to appear in the string with a certain probability (Brouwer & de Kok, 2010).

It is important to note that the main limitation of our quantitative corpus-based lexicological analysis is the inability to treat proper nouns and other complex phrases (e.g. the European Union) as one single phrase in the word and keyword list. This goes against the view described above. Nevertheless, we try to compensate this limitation by manual identification of 100 words that will be examined in depth on a qualitative basis. These words will be provided with collocations for the political discourse not only from those identified in the Europarl corpus, but also with collocations listed in dictionaries (for detailed information, see page 152).

Conclusion. The goal of this chapter was to establish the context, background and importance of corpus linguistics in our research project.

For a better introduction to corpus linguistics, the first part of the chapter dealt with corpora as the source material used in corpus linguistics. It discusses their usage, features and characteristics.

Next, a small chapter on EU related corpus-like linguistics resources followed, including the Europarl corpus, the source corpus used in the presented study.

The third part of the chapter covered a lot of ground about corpus linguistics, its history, methodological approaches, transdisciplinary application, and it also gave an insight into corpus linguistics software and main corpus linguistics tools.

---

[33] SMT tend to treat unlearned phrases as individual words which leads to funny translations. For instance, the awkward translation *Where she married herself, here she married herself* is to mean *she appeared out of thin air* (Czech idiom: *Kde se vzala, tu se vzala*), or *guilty basements* (correct translation vineyards) represent a verbatim translation of a weak collocation *vinné sklípky.*

The knowledge presented in this last theoretical chapter should help the reader better understand the quantitative corpus-based lexical analysis of the Europarl corpus. The procedure used in our study will be described in detail in the next chapter.

Europarl Corpus Analysis

Introduction

This chapter describes in detail the lexical-based corpus analysis conducted on the Europarl corpus with its primary aim to identify the most frequent words used in proceedings of the European Parliament. It is believed that this keyword frequency list based on the whole Europarl corpus will reflect the specific vocabulary used by Members of the European Parliament (MEPs) and could give an important insight into this language variety.

This chapter will firstly discuss glossary templates intended for this study, secondly, it will present various language and corpus linguistics resources used in the corpus analysis & the glossary compilation process, and thirdly, it will provide a detailed description of the corpus analysis & glossary compilation procedure for the verifiability of the study.

The corpus analysis can be also divided into 3 phases. The first phase represents the quantitative corpus analysis of the Europarl corpus. This encompasses generating the wordlist and the keyword list. The main output of the first phase is the keyword list which has served as a basis for vocabulary selection by compiling the terminological glossary. The second phase included manual examination of the keyword list, dividing words into various categories for better identification of suitable words for respective glossaries, and obtaining their language equivalents from reliable language sources. The third phase of the corpus analysis involved adding relevant information to glossary entries by using corpus linguistic tools (concordance and collocations search, n-grams).

Glossary Templates

As described in more detail in previous chapters, the main goal of this master thesis is to compile a glossary of the most frequent words used in the spoken EU parliamentary language. After preliminary research in terminology and lexical studies, we have already had in mind some glossary templates that could be used after the keyword analysis. These were later refined after consultations and careful consideration to suit the optimally suit the learner's needs. In this chapter, we will present 4 glossary types, and explain their purpose.

It goes without saying that different words require different approach: some information might be redundant in some word categories but very important in the others. Therefore, we decided to create 4 individual glossaries that will be freely available through various sources summarized later. After a careful study of ISO/FDIS-102421-2 that sets standards for terminological entries, we decided to accustom our glossaries for specific purposes for which they should serve.

Glossary 1. The first glossary (hereafter referred to as Glossary 1) includes all 2000 words identified in the keyword list both in the alphabetical and frequency order with their respective translations in all target languages (German, Czech, and Slovak). The purpose of Glossary 1 was solely to list all 2000 keywords without any additional statistical or lexical information. Accepted synonyms identified from lexicographic resources were inserted after a backslash (e.g. translation /synonym/, /synonym/). Glossary 1 was saved in 3 various formats (*.xls file, InterpretBank glossary file and the flashcard application Anki which allows for learning the vocabulary on the go. That should enhance the didactic value of Glossary 1 for language learners. The translation sources were identified in separated columns only in the *.xls file that enables the reader to check the lexicographic source. The source was in an abbreviated form for saving space and more clarity. The list of all abbreviations was provided at the end of the glossary. Identifying the source of target language equivalents was evaluated as redundant in other glossary formats and omitted.

Figure 7. Example of Glossary 1 in *.xls table.

English (source)	German	Czech	Slovak		SRC GE	SRC CZ	SRC SVK
word	word & synonym	word	word & synonym		bca	cba	bac
word	word & synonym	word	word & synonym		bca	cba	bac

Glossary 2. The second glossary (Glossary 2) consists of selected internationalisms that may or may not be translated in a similar form across languages. It includes English keywords and various equivalents in target languages. Each equivalent was assigned to a category of a preferred, admitted or deprecated term based on the gathered linguistic information. The glossary also includes lexicographical sources (indicated similarly as in Glossary 1) and optional notes for each language. The purpose

of this glossary was to highlight internationalisms in our keyword list used in target languages and to identify the most common false friends that are clearly a source of confusion for a many language students. The didactic value of this glossary can be particularly appreciated by students striving for correct language usage in political settings. The inspiration for creating this glossary stems from terminological ISO standards and the particular importance put on correct target language usage in interpreting (see Makarová, 2004).

Figure 8. Glossary 2 template

_	*Categorization*	*Translation*	*Notes*	*Source*
EN	***source word***		KW no.:	
GE	preferred			
CZ	preferred			
	admitted			
	deprecated			
SK	preferred			
	admitted			
	deprecated			

Glossary 3. The third glossary is the most comprehensive one of the presented glossaries. It comprises of statistical data (Wordlist rank, KW rank, no. of occurrences), equivalents in target languages, synonyms and collocations in all languages, n-grams in English, the English definition and 2 examples of context use in the source language. In addition, lexicographic sources will be provided for target language equivalents, synonyms, collocations and the English definition (similarly as in Glossary 1). The authors carefully selected 200 terms out of 2000 keywords for this glossary. The main criterion for selection was the complexity of the term with regard to various collocations and meaning and its significance as a keyword. The purpose of Glossary 3 is to offer a detailed insight into the usage of the discussed terms in all languages. Pedagogical implications of this glossary lie in observing the selected keywords in context with their natural collocations. This should lead language students to a deeper understanding of different usage for selected terms.

Figure 9. Glossary 3 template

TERM	word, (n)		Sources:		
GE:,		**CZ:**	**SK:**		
syn EN					
syn DE			**Worldlist rank**		
syn CZ			**Keyword rank**		
syn SK			**Occurance**		
Definition					
Collocations EN					
Collocations GE					
Collocations CZ					
Collocations SK					
Further English Collocations					
N:grams	sample collocates /N/		/N/ sample collocates		
Context					

Glossary 4. The last glossary, Glossary 4, takes the form of a typical glossary used for special terminology. In the context of the EU parliamentary language, it includes special terms that require certain background knowledge not only for better understanding of the individual terms, but also for grasping the whole context of a speech segment in which they occur. Therefore, we selected following attributes as relevant for this glossary: original term, equivalents in 3 target languages, statistical data (occurrence and KW rank), definition in English and notes. This should allow language students not only to understand the particular term from the English definition provided and compare it to equivalents in other languages, but also to get a valuable insight into the statistical data about its usage in the Europarl corpus. The pedagogical implications of this glossary could be seen in pointing out specialized terms for learning purposes that would otherwise be learnt through incidental reading at a much slower rate (Jenkins, Stein & Wysocki, 1984).

Figure 10. Glossary 3 template.

TERM	sample word, (n)	**Occurrence:** 10 000	**KW rank:** 0000		
GE: sample translation	*CZ: sample translation*	*SK: sample translation*	src	src	src
Definition: (src)		**Notes:** (src)			

Glossaries overlapping. Since we compiled 4 individual glossaries with different scopes, there may be instances of some terms overlapping. Glossary 1 includes all 2000 most frequent keywords from the keyword list, some of which will be discussed in glossaries 2, 3 and 4. Whereas Glossary 1 includes all terms used in other glossaries, minimal overlapping is expected between glossaries 2, 3 and 4. It might occur; however, that certain terms discussed in glossaries will be identified during the corpus analysis as frequent collocations to other terms.

Corpus Linguistics Software & Language Resources

Our endeavor to obtain the most accurate, reliable and replicable results led us to use the WordSmith tools as a standard procedure in computer corpus linguistics for conducting the overall statistical analysis of the Europarl corpus. The reason for this decision was explained in more detail in discussion about corpus linguistics software on page 101, stating that the results of quantitative corpus analysis may vary slightly depending on the software tools and the statistical test used. Keeping possible deviations at minimum was particularly important because the output of the quantitative corpus analysis, the keyword list, served as a basis for the terms selection during the glossary compilation.

Computing the statistical analysis was a rather easy task that did not require much time or effort. Conducting an in-depth analysis of Europarl for selected terms; however, is a vastly time consuming procedure. Since we could see no reason that would restrain us from using AntConc as an alternative free software computer-assisted corpus analysis tool for the in-depth lexical analysis, we decided to proceed with AntConc despite the fact that WordSmith 5.0 and 6.0 seemed to work faster in concordance and collocation search. This enabled us to work independently from the university computer lab and freely without any time restrictions. We used AntConc n-grams, collocation and concordance search for providing additional terminological information in Glossary 3 and occasionally WordSmith 5.0 and 6.0 to verify several doubtful results. Additionally, we also used the demo version of ParaConc 1.0 for searching through the parallel Europarl corpora.

Other lexicographic sources can be divided into two categories. The first category includes parallel corpora and the second category consists of dictionaries. However, with

the advent of information technology, many online tools could be considered as hybrids of corpora and dictionaries, which might blur this attempted classification. The lexicographic sources used by the glossary compilation and introduced in the next chapter are Europarl Parallel corpora, Linguee and Glosbe for the first category and IATE and Lingea, for the second.

Parallel corpora.

Europarl Parallel corpora. Obviously, the main parallel corpora that ought to be used in this study are Europarl parallel corpora for our target languages. We downloaded them from the Europarl corpus official website (latest release: v7 from 2011) in language combinations GE-EN, CZ-EN and SK-EN.

The statistical data for these corpora are following:

Figure 11. Statistical data of parallel corpora for GE, CZ, and SK. Koehn. 2011. *Sizes of parallel corpora after sentence aligning, tokenizing, and removing XML.* Retrieved from http://www.statmt.org/europarl/.

Sizes for parallel corpora after sentence aligning, tokenizing, and removing XML

Parallel Corpus (L1-L2)	Sentences	L1 Words	English Words	Size	Time Period
German-English	1,920,200	44,548,491	47,818,827	189 MB	04/1996-11/2011
Czech-English	646,605	12,999,455	15,625,264	60 MB	01/2007-11/2011
Slovak-English	640,490	12,942,434	15,442,233	59 MB	01/2007-11/2011

Europarl parallel corpora were used mainly for context search and identifying correct target language translations & synonyms. As already mentioned above, the software used for this purpose was ParaConc 1.0, a parallel concordancer for contrastive corpus analysis. We also experimented with the freeware AntPConc 1.1, but due to difficulties experienced in loading the large text file in this freeware alternative, we had to use ParaConc 1.0 instead.

The demo version of the software works very well and fits all needs for initial parallel corpora screening. Firstly, it is fully customizable and secondly, the limitations of 150 results and blocked saving & copying results did not prove as an obstacle for the analysis. It also supports UTF-8 character encoding for Slovak and Czech. The Europarl parallel corpora were already prealigned and cleared from XML tags which did not require any further processing before their loading in ParaConc. A very useful function

worth mentioning is the *hot words* tool that sorts all translations of a specific word in the target language by their frequency. The other frequency tools in ParaConc were not used because the analyzed parallel corpora makes up only a part of the whole Europarl corpus for which we had already generated all statistical data needed.

Other corpora sources used for identifying adequate translations were Linguee.com and Glosbe. All of them were accessed through their online web interface.

Liguee. Linguee is a unique online translation tool combining an editorial dictionary and a wide corpus with a powerful search engine. The unique process of building the corpus is explained below:

> (Linguee is) A specialized computer program, a web crawler, automatically searches the internet for multiple language webpages, particularly from professionally translated websites of companies, organizations, and universities, EU documents, patent specifications etc. These pages are detected automatically, and the translated sentences and words are extracted. The texts are then evaluated by a machine-learning algorithm which filters out the high quality translations for display. This system is capable of autonomously learning new quality criteria to tell apart good translations and bad ones…. Through this training process, our algorithm is continuously learning to find thousands of such correlations and reliably extract the best translations autonomously. Our computers have already compared and evaluated more than 10 trillion sentences. However, at the end of the day, just the top 0.01 per cent, i.e. 100 million of the translated sentences, were retained (About Linguee, 2015, "Where does the text content come from?" para. 1).

Linguee was founded by the former Google post-doctorate Dr. Gereon Frahling and Leonard Fink in 2008. Since then the project grew rapidly both in quality and quantity (due to a constantly improved search algorithm and user interface, developing an offline application for mobile devices, adding 218 new language pairs including Slovak and Czech in 2014 etc.). Apart from precisely tuned automatic corpus building, numerous editors are constantly working on enhancing the dictionary by adding new entries and correcting existing ones in order to provide reliable and comprehensive editorial dictionary (Linguee, 2015).

Glosbe. Glosbe is also a hybrid between a corpus and a dictionary. It was created in 2011 and uses data from both open source free databases and from users contributions (corpora, translation memories etc.). It aspires to become the biggest online dictionary and its greatest advantage is an exhaustive amount of available languages. It boasts around 7000 languages for various language combinations that cover the vast majority of existing languages. However, unlike Linguee, Glosbe has been developed and improved voluntarily by the online community much like Wikipedia, hence, the quality of dictionaries without paid professional editors is vastly dependent on provided corpora and translation memories (Glosbe, 2014)[34].

Dictionaries. There is a plethora of comprehensive dictionaries freely available online for the most popular languages (e.g. for English and German Duden Wörterbuch, Cambridge and Oxford dictionaries etc.) that could be used in our lexicographical search by looking for correct translation equivalents. Unfortunately, this is not the case for less common ones such as Slovak and Czech. Although the freely available online resources for these languages have experienced a boom in the last few years, the great expansion of multilingual corpora did not reflect in quality advancement of free online dictionaries. A possible explanation could be seen in the fact that offering dictionaries online for free is not profitable for Czech and Slovak companies due to the high development costs and low revenues from online advertisement. Clearly, they have a much lower online target audience resulting in lower revenue from online marketing than companies developing dictionaries for the more frequent language combinations with millions of visitors daily. Instead, the companies offer free online dictionaries limited in size, and use their websites for promoting the full versions of their products[35].

Lingea. The most notable and quality dictionaries in Slovak nd Czech are developed by the Czech company *Lingea s. r. o.*. The company, founded in 1997, states that it had developed the most comprehensive dictionaries on the market available in both

[34] https://glosbe.com/

[35] Free online web dictionaries on www.slovnik.cz / www.slovnik.sk contain about 50% of the database from the full version of PC Translator 2010 developed by Teos.
Free online web dictionaries on www.slovniky.lingea.cz and www.slovniky.lingea.sk contain 30% of the database from Platinum Lingea dictionaries.

printed and electronic version. The Platinum dictionary contains 110,000 English keywords, over 100,000 Slovak/Czech keywords, and 155,000 idioms and phrases (Lingea, 2015a, 2015b). Upon discussion with the sales department and introducing our glossary project, we were granted 1-month free web access to the full version of the dictionary. In our master thesis, we used their German-Czech, German-Slovak and English-Czech, English-Slovak Platinum dictionaries as the preferred dictionary source for Czech and Slovak next to the IATE.

IATE. IATE (= "Inter-Active Terminology for Europe") is the EU's inter-institutional terminology database. IATE has been used in the EU institutions and agencies since summer 2004 for the collection, dissemination and shared management of EU-specific terminology. The project was launched in 1999 with the objective of providing a web-based infrastructure for all EU terminology resources, enhancing the availability and standardization of the information. IATE incorporates all of the existing terminology databases of the EU's translation services into a single, interactive and accessible interinstitutional database (IATE, 2015)[36]. IATE enables searching a particular term in two or more languages simultaneously.

In this introductory chapter, we introduced the style and functions of individual glossaries and discussed main lexicographic sources used by the glossary compilation. In the next chapters, we will provide the detailed description of the quantitative corpus-based lexical analysis of the Europarl corpus and discuss individual contextual findings.

Regarding identification of lexicographic sources in the glossary, it should be noted that all sources are identified under assigned numbers and listed at the end of each glossary.

Phase 1: Quantitative Corpus Analysis of the Europarl Corpus

Files Preparation & Processing. For the purpose of the analysis, we downloaded the latest version of the Europarl corpus (source release v7 from May 15th, 2012) and proceeded with the English monolingual corpus which consists of 9672 text files (*.txt) in total size of 346 100 kB. The length of individual files varies broadly due to special requirements for sentence alignment used in multilingual Europarl subcorpora and due to

[36] http://iate.europa.eu/

the data mining procedure used by creating the Europarl corpus (see Koehn, 2005). The corpus can be merged into one single text file by TXTcollector (Bluefive software), but since the majority of corpus software can analyze numerous text files simultaneously, this procedure is not necessary.

In order to keep results of the KeyWord list analysis as standardized as possible, we decided to use WordSmith, presumably the most widely used software for corpus analysis (Nesselhauf, 2005; Römmer & Wulff, 2010; Granger & Paquot, 2012; Müller & Waibel, n.d.) with the statistical log-likelihood test, also preferably used in corpus linguistics (Paquot & Bestgen, 2009).

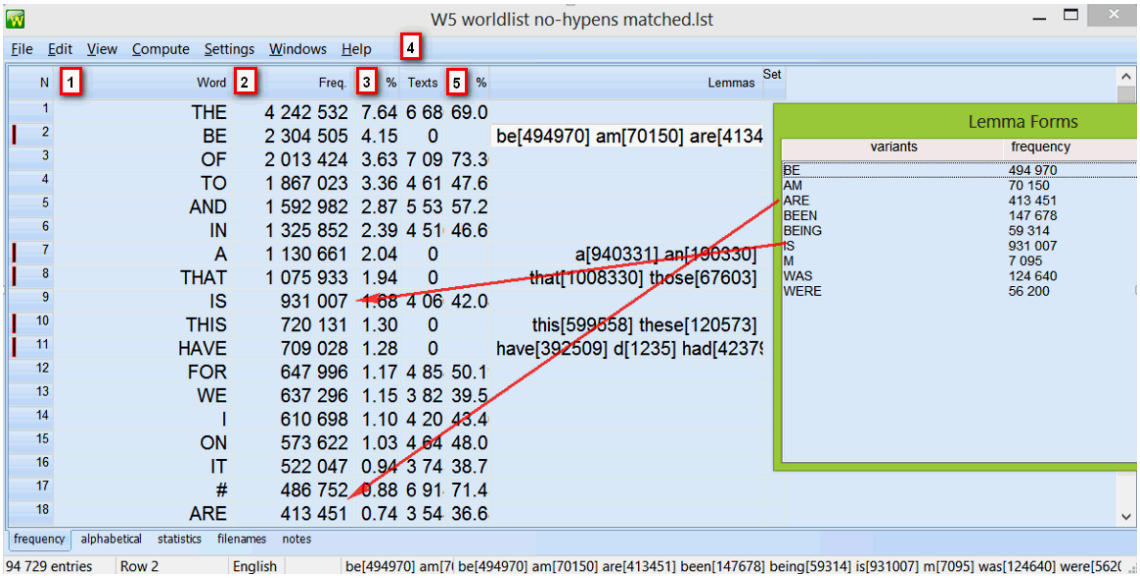
Due to the fact that WordSmith is a commercial product and the demo version of the latest release (WordSmith 6.0) lists only first 100 queries in its wordlist and keyword list generator, we had to rely on the 5.0 older version that is preinstalled in a computer lab at the Centre for Translation Studies at the University of Vienna. To our knowledge, the latest version, 6.0, has not been purchased by the University for scholar purposes at the time of conducting the analysis. WordSmith 6.0 might offer some new functions, but in essence, the older version can perform the statistical corpus analysis equally well. Furthermore, WordSmith 5.0 proved to be faster at processing and searching through the vast amount of data of the Europarl corpus compared to WordSmith 6.0 and AntConc 3.4.3. Nevertheless, this small observation might have its limitation as the hardware differences (university PC vs. high-tech notebook) were not considered.

Wordlist generation. After feeding WordSmith 5.0 with all files from the Europarl English monolingual corpus, we generated a wordlist using the lemma list by Someya available on the WordSmith tools webpage. It currently contains 40,569 words (tokens) in 14,762 lemma groups. The author compiled the lemma list in 1998 and states that “It is still far from being complete”. This is a limitation that must be considered, given the fact that creating an exhaustive customized lemma list for the Europarl corpus would be vastly time-consuming. As we will have found out later, some specific keywords in the generated keyword list appeared mistakenly several times in various lexical forms because they were not included in the lemma list.

The generated wordlist contained 65,519 unique words. The lemmatized words appeared in the general count as the sum of all inflected occurrences identified for the respective lexeme and in its infliction form in the word order. For instance, as indicated

in figure 12 below, the function word “be” with its inflected variations *be* [494970] *am* [70150] *are* [413451] *been* [147678] *being* [59314] *is* [931007] *´m* [7095] *was* [124640] and *were* [56200] appears on position 2, but we also find the inflection of *be* under *is* and *are* on positions 9 and 18. It appears that lemmatization helps identify the basic form of a word (lemma), but still leaves all inflections of the respected word in the wordlist.

Figure 12. Generated wordlist in WordSmith 5.0, opened and viewed through the demo version of WordSmith 6.0.



Legend from the WordSmith Tools Manual:

1. the word
2. its frequency
3. its frequency as a percent of the running words in the text(s) the word list was made from
4. the number of texts each word appeared in
5. that number as a percentage of the whole corpus of texts (WordSmith Tools Manual, 2015, “WordList display”, para. 1)

Keyword list generation. As mentioned in the theoretical part about keyword lists on page 107, keyword lists are used to identify unusually high frequencies of words in comparison with some norm. The standard procedure in all corpus analysis programs is firstly to generate a frequency wordlist from corpus files and secondly, to compare the

wordlist with another frequency wordlist(s), generated usually from a much larger corpus or corpora that should represent a certain norm. However, the vexing question here would be to ask what this norm really should be since it may vary in its scale and focus – from typical words used in essays written by English second language speakers to a more general wordlist comprising texts from written or oral sources of different styles and nature, all balanced in a careful linguistic fashion.

The main objective of our corpus analysis was to create the most accurate keyword list that would reflect the specific register used in the European Parliament. We intended to use the generated keyword list for creating a glossary of spoken EU parliamentary language that could be used as didactic tool in various fields of specialization by non-native English speakers. Therefore, we deemed this as the best option for a frequency list based on a corpus representing general British English with a carefully balanced selection of texts from various styles and registers. We identified the British National Corpus (BNC) as the most suitable source for this comparison.

The BNC contains up to 100 million orthographic words and is one of the most important corpora used in computer linguistics. It represents a wide cross-section of British English from the latter part of the 20th century in written (82%) and spoken (18%) form. The written part includes, for example, extracts from regional and national newsletters, academic books and essays, popular fiction, memoranda etc, whereas the spoken part comprises of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different ages, region and social class in a demographically balanced way and collected in different contexts ranging from formal business or government meetings to radio shows). However, the ratio of written and spoken text sources stays unclear because the website states that the BNC World comprises texts from 90% of written and 10% of spoken sources whereas the table below suggests that the percentage is approximately 82% to 18% (BNC, 2015).

Figure 13. Composition of the BNC World Edition. BNC. 2009. *The BNC in numbers.* Retrieved from http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=numbers.

Text type	Texts	Kbytes	W-units	S-units	percent
Spoken demographic	153	4206058	4.30	610563	10.08
Spoken context-governed	757	6135671	6.28	428558	7.07
All Spoken	910	10341729	10.58	1039121	17.78
Written books and periodicals	2688	78580018	80.49	4403803	72.75
Written-to-be-spoken	35	1324480	1.35	120153	1.98
Written miscellaneous	421	7373707	7.55	490016	8.09
All Written	3144	87278205	89.39	5013972	82.82

Despite the ratio findings, the overall disproportion of spoken language in the BNC might seem to be a significant limitation at first glance, considering that its wordlist should be used as a reference for keywords identification in the Europarl corpus consisting solely of spoken language. On the other hand, we must bear in mind that for the most appropriate keywords identification for didactic purposes, the quantity and variety of general English words is more important than the mode in which these words were uttered.

A considerable advantage of the BNC corpus can be seen in the fact that scholars do not need to analyze the whole corpus (size: 4,5Gb) to obtain the word frequency list. This has already been done by Adam Kilgarriff. Kilgarriff has made a significant contribution to computer-assisted corpus linguistics as he made an exhaustive part of his linguistic research freely available. This includes the word frequency list of BNC World that is used as a springboard by many corpus linguists over the world. Various functions of this word frequency list such as lemmatization and POS tags can be used for specific purposes in an in-depth linguistic analysis, but these will not be described in detail since this topic is outside the scope of this master thesis.

One issue worth mentioning is the fact that we did not use the latest version of the BNC. The British National Corpus was compiled between 1991-1994. Although no text has been added after the completion of the project, the corpus was slightly revised in its second (BNC World, 2001) and third (BNC XML Edition, 2007) edition. To our knowledge gathered from internet research in 2015, the only freely available BNC word frequency list was created from the second edition, BNC World, by Adam Kilgarriff. This wordlist seems to have been distributed for free on many webpages and in manuals devoted to corpus linguistics. The official BNC website supports our findings stating:

> When we last checked (2009), none of the following had been updated to use the BNC XML Edition. The counts they provide all relate to the earlier BNC World Edition. Adam Kilgarriff has produced word frequency list for the BNC World Edition, available from his webpage. (BNC, 2009).

We also got confused by the release and update dates stated at the bottom of Kilgarriff's webpage. Strangely enough, the last update as stated on the webpage comes from 1996, several years before the second edition of the BNC corpus (BNC World, 2001) was released. The only logical assumption here would be to conclude that the text of the webpage was updated, yet the update information at the bottom of the webpage remained obsolete.

Figure 14. Dates inconsistency issue. Kilgarriff. 1996. *BNC Database and Word Frequency Lists*. Retrieved from https://www.kilgarriff.co.uk/bnc-readme.html.


The last is useful because, for distributions like the normal, poisson, high variability, and prepositions, low (cf. Kucera and Francis 1982)

Uncompressed, available here.
Compressed, available here.

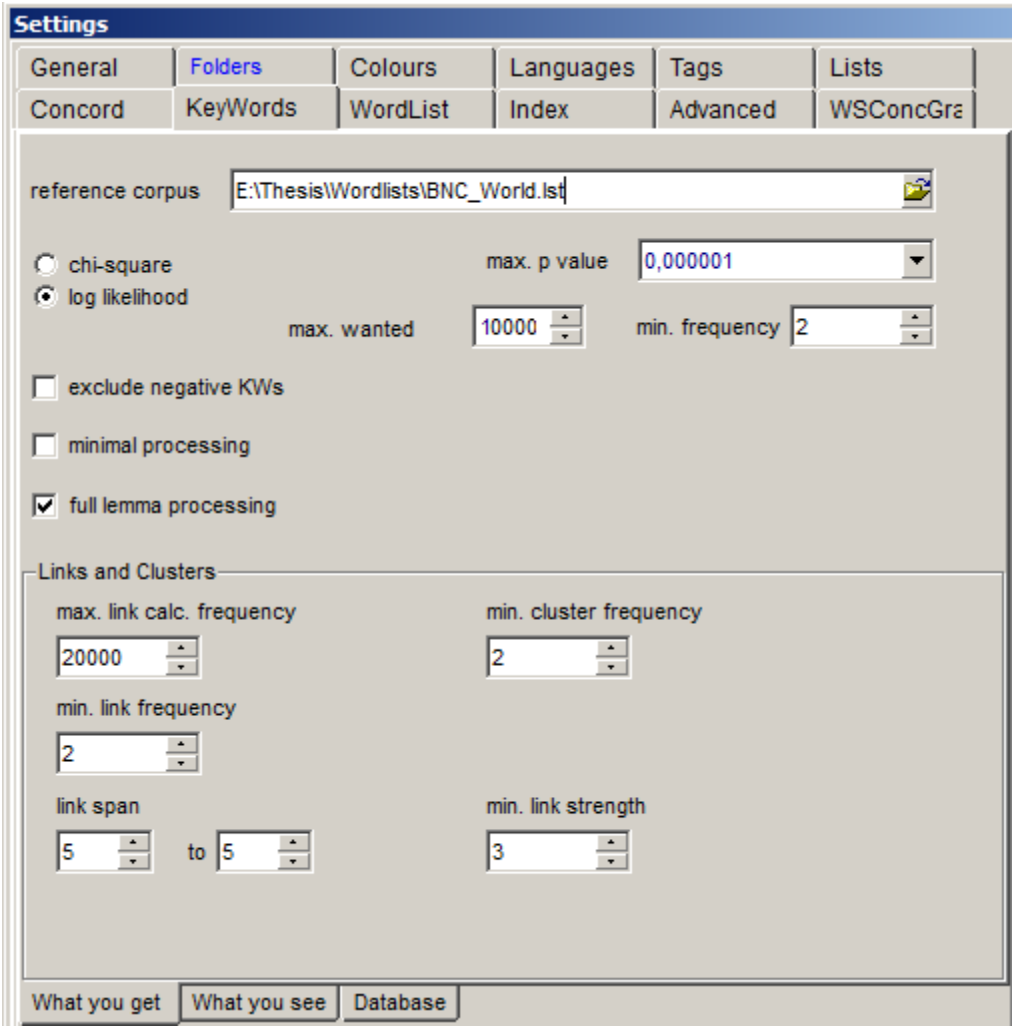
This page: http://www.kilgarriff.co.uk/bnc-readme.html
Written: 20 Nov 1995
Updated: 15 March 1996
HTML version: 3 Nov 1998
Moved to http://www.kilgarriff.co.uk website: 1 Mar 2006

The process of creating the keyword list included downloading the BNC World frequency wordlist (4MB zipped, 14MB unzipped), adding the BNC wordlist to WordSmith 5.0 as a reference wordlist, loading a non-hyphens lemma list by Someya[37] and preparing so called stop list, whose main function was to eliminate certain words that researchers do not want to include in the generated keyword list.

For the KW generation, default settings of WordSmith 5.0 were used as illustrated in the screen capture below.

[37] The lemma list with hyphens by Someya generated obscure and faulty results.

Figure 15. Default KW settings used in WordSmith 5.0 during the keyword analysis.



We used the stop list to exclude the most frequent words such as a basic nouns, adjectives, functional words etc., The reason behind this decision is the fact ESL learners already have active knowledge of these words and their inclusion in the keyword list would not bring any additional didactic value, but on the contrary, adding them would be intrusive by learning the keywords. The underlying thorny issue, however, is to define which and how many words belong to the active vocabulary of an advanced ESL speaker to speak the language fluently. As discussed more in detail in the theoretical part about vocabulary acquisition, linguists and educationalists are particularly sensitive about this question. The general claim is that there is no exact answer due to great many factors involved (Libermann, 2008). Still, by rule of a thumb, we can proceed on the assumption that vocabulary of 3000 most frequent words provides coverage for around 95% of

common texts (Fry & Kress, 2006). Therefore, these 3000 words should be a part of the active vocabulary among advanced ESL speakers.

We decided to create a stop list from the word frequency list of a BNC subcorpora for spoken English compiled by Leech, Rayson and Wilson (Rank frequency order: spoken English; not lemmatized, 4841 words in total[38]). We believe that the most frequent words will be better reflected in a spoken corpus rather than in a written one. After saving the *.txt file from the website, we opened the file, copied the data into an excel table, cleared all columns except the plain words and cleared the list from inconsequential wildcards[39], letters (36), and interjections (56). The vast majority of proper nouns (193) and numbers (47) were left in the wordlist in order not to manipulate the frequency word list extensively. The only deleted proper nouns (12) were manually identified as potential keywords in EU politics[40]. Leaving them in the stop list could lead to their omission in the generated keyword list, which was not desirable for the Europarl corpus analysis. Additionally, we had to delete duplicates resulting from POS tagging in the BNC spoken English word frequency list. For instance, the lemma "about" appeared in the list twice – the first time as preposition (position 69) and the second time as an adverb (position 172). Removing duplicate words (308 out of 3308) could be easily done with the tool "remove duplicates" in MS Excel 2013. Afterwards, we saved the only existing column with 3000 words (from originally 3410) as a plain text format (*.txt) and imported it as a stop list in WordSmith 5.0. For the reference, please see the file *BNC Spoken word frequency list.xlsx*.

It might be important to note that the BNC subcorpus of spoken English also contained some slang words and interjections (e.g. *eh, aye, aha, er* etc.) that were not expected to occur in the Europarl corpus and therefore were also deleted from the stop list to free up space for more important words in general spoken English. The reason is

[38] http://ucrel.lancs.ac.uk/bncfreq/flists.html

[39] WordSmith stop list does not support wildcards.

[40] Deleted proper nouns: England, Wales, America, Scotland, Germany, United, States, Ireland, West, Thatcher, Australia and Western. The other proper nouns either did not make it to the top 3000 most frequent words (e.g. Russia, Egypt), or referred to domestic topics (e.g. Brighton, Glasgow).

the following: The Europarl corpus consists of verbatim transcripts of the EU-proceedings that were mined from the EU webpage and it is evident that these transcripts had to be professionally proofread prior to their online release. Consequently, corpus linguists are not to find slang and interjections in the Europarl corpus due to the official status and register prescribed for released proceedings, as opposed to corpora focused also on paralinguistic features such as ELRA or EPIC. Thus, leaving interjections in the stop list would be of no use for the Europarl corpus analysis.

Computing the keyword list with all mentioned resources was an easy and quick task. Finally, the generated KW list was saved both in an *.xls file and in WordSmith KW list file.

Figure 16. Generated KW list exported into an Excel sheet.

	A	B 1	C 2	D 3	E 4	F 5	G 6	H 7	I 8
17	N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P	Lemmas
18	1	CONJURER	183969	0.3313	13		377783.8438	1E-26	
19	2	Â	70343	0.1267	0		144458.7031	3E-25	
20	3	THE	4242532	7.6393	6055105	6.0877	135852.2656	3E-25	
21	4	PROPOSAL	77463	0.1395	4202		129691.8047	4E-25	
22	5	COMMISSIONER	56107	0.101	1386		103377.9766	7E-25	ommissioners
23	6	AMENDMENT	56713	0.1021	1797		101995.3906	8E-25	amendments
24	7	CITIZEN	48111	0.0866	1149		88901.3125	1E-24	
25	8	DIRECTIVE	45738	0.0824	1718		80663.42969	2E-24	directives
26	9	Â€	37980	0.0684	0		77982.76563	2E-24	
27	10	COOPERATION	38791	0.0698	1288		69400.61719	2E-24	
28	11	ADOPT	39117	0.0704	2460		63817	3E-24	opting adopts
29	12	RAPPORTEUR	30253	0.0545	29		61679.12109	3E-24	
30	13	PROPOSE	33939	0.0611	1392		59186.46484	4E-24	ses proposing
31	14	UNITE	29628	0.0533	449		56562.39063	4E-24	unites uniting
32	15	AMENDMENTS	31190	0.0562	984		56110.76563	5E-24	
33	16	GENTLEMEN	31081	0.056	1294		54094.40234	5E-24	
34	17	MADAM	28912	0.0521	790		52772.05078	5E-24	
35	18	PRESIDENCY	28331	0.051	1157		49432.74609	7E-24	presidencies
36	19	RESOLUTION	34095	0.0614	3670		49176.78906	7E-24	resolutions
37	20	GENTLEMAN	31271	0.0563	4505		41113.48828	1E-23	gentlemen
38	21	IMPLEMENT	22724	0.0409	1548		36510.83594	2E-23	implementing

Legend from the WordSmith manual:

1. each key word
2. its frequency in the source text(s) which these key words are key in. (Freq. column below)
3. the % that frequency represents
4. its frequency in the reference corpus (RC. Freq. column)
5. the reference corpus frequency as a %
6. keyness (chi-square or log likelihood statistic)
7. p value

8. lemmas (any which have been joined to each other)

(WordSmith Tools Manual, 2015, "KeyWord list display", para. 1)

The file can be accessed in additional electronical resource as part of this paper.

Phase 2: Manual Selection of Words for Glossaries from the Keyword List

Keyword list categorization. The automatic selection of keywords has already been done by the corpus linguistics software as described in phase 1. This would imply that creating glossaries only takes selecting the top 1500 keywords and supplementing them with target language equivalents. However, as we have already mentioned in chapter about glossary templates, the actual process of glossary compilation goes much further. It requires a careful manual selection of words for respective glossaries. Therefore, the most frequent words have to be further categorized in order to be later correctly selected for relevant glossaries. Since this process fully relies on human intuition and background knowledge, the overall objectivity of this process cannot be asserted. Still, we have put some safeguards in place that can minimize the risk and document the whole procedure for a possible replicability in the form of four filters in an Excel sheet.

The first filter distinguishes whether the word is a verb, an internationalism, a proper noun, an abbreviation or as an incorrect word that mistakenly appeared in the results.

The second filter follows Klein's typology of political terms (see chapter about politolinguistics).

The third filter is meant for a preliminary assignment of selected words to respective glossaries (1, 2, 3, 4). Careful consideration of each term should provide a certain degree of reasonable objectivity to the word selection for glossaries 2, 3, and 4., The result of this manual selection can be seen in columns J-M in the keyword list excel file. For this initial corpus screening, EP parallel corpus for English – Slovak was used and analyzed by ParaConc with functions "hotwords" and "collocate frequency data". This proved to be the fastest alternative.

The fourth filter was used for word omissions. After initial screening of the generated keyword list, it became obvious that a considerable number of words has to be left out from the final glossaries. Such words have been considered as having no additional value to the glossary and/or expected to appear in the keyword list. These left-

outs are well documented in the fourth filter in four categories: 1) words that presumably appeared in the generated keyword list by mistake (e.g. Â, the, Â€); 2)) they appeared to be too easy, general or irrelevant (e.g. fishing); 3) they represented proper names (Afghan) 4) politicians (e.g. Putin, Prodi); 5) they were repetitions or flexemes unrecognized by the lemma list (e.g. NGOs). Regarding the last category for unrecognized lemma identification, the final keyword list for the first glossary was sorted alphabetically for another manual check whether all unidentified lemmas were correctly marked as duplicates. More precisely, most identified keywords as verbs in gerund or third person singular, all nouns in plural provided they were also found in singular, and all spelling derivations between US and UK English (e.g. authorize vs. authorize were manually marked as unrecognized lemmas and left out from the final list. On the other hand, present and past participle of verbs that form adjectives were left in the final selection of keywords. This manual check was repeated 3 times to reduce the risk of human error to an acceptable level.

Due to difficulties with grouping lemmas, mistakes in keyword generation and a considerable amount of proper nouns, the total amount of omissions is rather high and accounts for 36,60% of all words in the initial keyword list of the top 3000 words. The statistic can be seen in the table below.

Table 1. Statistical results of the manual categorization.

TOTAL ANALYSED WORDS	**3000**	**%**
Set 1 (initial analysis of 3000 words)		
Abbreviations	133	4.43%
Verbs	554	18.43%
Set 2: Klein's typology (1921 unique keywords)		
institutional vocabulary (1)	187	9.73%
vocabulary of specialist fields (2)	58	3.02%
general interaction vocabulary (3)	1618	84.23%
ideological vocabulary (4)	58	3.02%
Set 3: initial categorization of glossaries (1921 unique keywords)		

Glossary 1 (all words)	1500	78.08%
Glossary 2 (Internationalisms)	470	24.47%
Glossary 3 (detailed)	132	6.87%
Glossary 4 (definitions)	142	7.39%

Set 4 (omissions)		
Incorrectly generated words (1)	220	7.33%
Common words (2)	41	1.37%
Proper names (3)	225	7.50%
Proper names (politicians) (4)	271	9.03%
Lemmas (5)	322	10.73%
Omissions total	1079	35.97%

Percentage of verbs identified in top 3000 keywords. The keyword analysis has shown a slightly higher percentage of verbs as it is usual for GSL conversational English. We have identified 554 verbs in 3000 analyzed words which accounts for 18.43%. Interestingly, corpus evidence used in the 'Longman Grammar of Spoken and Written English' estimates the distribution of verbs in conversations stands at roughly 12,5% (125 000 verbs in every one million words) (Biber, Johansson, Leech, Conrad & Finegan, 1999). Here it is important to bear in mind that these are the results from the top 3000 keywords identified in the corpus analysis and this slight increase only suggests that a) verbs might be slightly more frequent in the top keywords than in the whole keyword list, b) verbs obtained a slightly higher keyness than nouns in the keyword generation process.

Also, it is important to highlight that computing the exact noun to verb ratio, the standard procedure in quantitative corpus analysis, did not lie in focus of the presented study. Therefore future studies in quantitative corpus analysis of the Europarl corpus and its comparison to other corpora are suggested. Such studies could confirm the generally suggested strong tie between genre/text-type and the frequency of nouns and verbs, or in more simple terms, whether texts allowing information orientation promote the use of nouns, while narration is rather characterized by a higher incidence of verbs (McEnery, Xiao & Tono, 2006). Since we have not primarily analyzed the Europarl Corpus by orality markers, here we consider appropriate the mention the results of a comparative annotation analysis of the Europarl corpus focused on part of speech and syntactic function (the concrete statistics from this study can be seen in appendix F):

> The Europarl corpus, in particular, is atypical for speech, and in many regards closer to running text, most likely a consequence of it consisting of formal

> monologue, with an abstract public in mind rather than an individual turn-taker. Thus, the Europarl data boasts the longest words and longest sentences, and scores highest for subordination and infinitive subclauses, as well as the rare past participle subclauses, all of which considerably complicate syntactic trees. ... While the chat corpus consistently scored high on a number of orality markers, both the Enron e-mail data and the Europarl parliamentary transcripts proved to be atypical as representative sources of spoken language data, with - for instance - a low pronoun count in the former and a very high degree of linguistic complexity in the latter (Bick, 2010, p. 727, 729).

Klein's typology in the Europarl corpus. As mentioned in chapter *politolinguistics*, the political vocabulary cannot be fully seen as an LSP, but on the other hand it resembles it by some specialist elements. The initial analysis of the top 3000 identified keywords supports this hypothesis: out of 1921 unique keywords 85% can be regarded as words belonging to the category of general interaction vocabulary. The institutional vocabulary accounts only for 9%, and similarly, the vocabulary of specialist fields for 3%, and the ideological vocabulary for 3%. A possible explanation for rather small amount of words in other categories than general interaction vocabulary could be explained by fact that the keyword generation treated each words individually and did not take collocations into account. According to this, the words that would normally group in a term belonging to anything other than the third category were just identified manually as isolated words ripped out of context and categorized as belonging to the general interaction vocabulary. It is difficult to estimate how the percentage for the respective categories would change but sharp differences are rather unlikely. Generally speaking, we could apply the Pareto principle that 80% of political language consists of general vocabulary and only 20% could be assigned to the other three categories. The full lists of the first, second and fourth category can be found in appendix A. Here we should also point out again that the political language categorization is a product of manual selection dependable on authors' linguistics knowledge and intuition, and thus prone to error.

Comparison of results from the keyword list and the wordlist: proper nouns. In this discussion we would like to point to several findings resulting from the comparison of the generated wordlist and the keyword list regarding to proper nouns. On first impression, proper nouns are strikingly frequent in the top 3000 keywords. The reason

for this can be attributed to the statistical analysis that computed by algorithm very sensitively the unusual high frequency of proper nouns in the EP Corpus compared to the reference BNC corpus. In more simple terms, if some words appeared rarely in the reference BNC corpus and moderately frequently in the Europarl corpus, they acquired higher KW ranking than words with the same number of occurrences in the Europarl corpus as the first group, and the higher frequency in the reference corpus.

There might be two opposite views on this issue. The first view is that this is rather an undesirable outcome because such words cannot be used for learning purposes and they only took the place of other, more frequent and important words that could appear among the top 3000 analyzed words (or 1500 selected unique keywords respectively). This objection cannot be denied – these words truly take place of other more important and more frequently used words. Nevertheless, this highly sensitive keyword analysis brings along some advantages as well as we will see in the next chapter about abbreviations,

To tackle these unexpected results, all proper nouns were identified and left out from the final selection of 1500 keywords. The full list of these words can be seen in the corpus analysis working sheet by selecting the third (proper names as names of countries) and fourth category (names of politicians) in column M in the sheet named Processing filter.

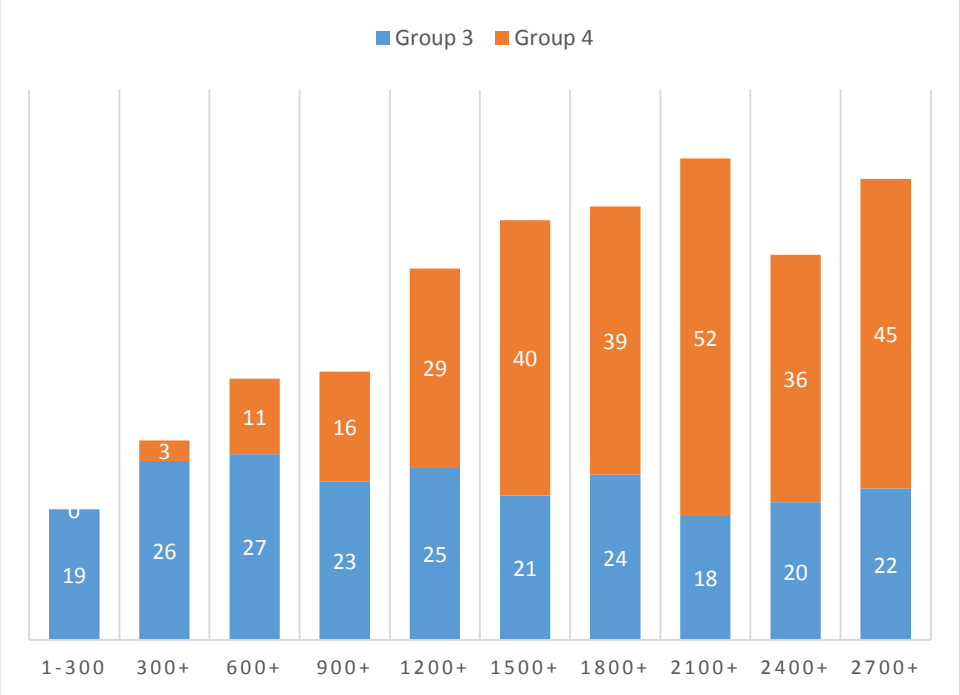
The statistics gathered from proper nouns are also worth mentioning:

As we can see, nine out of ten proper nouns from the random selection would not make it to the HFWL list of the EP corpus. This sends a clear message that keyword list generation may lead to some undesirable results. This problem could be tackled by setting minimal occurrences in the KW to 500+, but such proceeding could lead to leaving out other quite important and less frequent words. If we were to create only a glossary based on the high frequency words compared to a general reference corpus, it would be suitable to apply this setting. In our case, leaving out other less frequent, yet important words was not the intention, therefore we deem the default sensitive setting for the keyword analysis combined with manual selection of undesired words as appropriate.

Table 2. Statistical results of a random sample from the proper nouns identified in the analyzed 3000 words (words above 3000 that would not come to a typical HFWL list are in italics).

Word (roughly every 300. word)	**KW rank**	**HFWL rank**	**Occurrences**	**BNC occurrences in a comparable sample of 60 million words[41]**	**BNC corpus vs. EP corpus: %**
Copenhagen	292	1736	3259	164	5.02%
Hercegovina	590	*24310*	1250	16	1.30%
Albanian	895	*4025*	1321	150	11.36%
Belarusian	1207	*5760*	457	3	0.59%
Azores	1497	*5341*	526	39	7.49%
Basel	1796	*5895*	438	42	9.62%
Beijing	2171	*4078*	851	272	31.98%
Fabra	2430	*9519*	163	1	0.33%
Kosovar	2727	*9915*	150	2	1.44%
Galician	2999	*9324*	170	13	7.62%

Figure 17. The distribution of proper nouns (proper nouns, group 3 + proper nouns group 4 – names of politicians) in 3000 analyzed words. .



[41] BNC corpus contains 100 million words whereas the EP Corpus 54 million. Therefore, we took the number of occurrences from the BNC corpus and calculated 60% to match the size of the EP Corpus.

As we can see in the figure, group 4 (names of politicians) started to appear after first 300 keywords and quickly became dominant after first 1000 keywords. This implies that keywords analysis after 1000 keywords may yield undesirable results if the keyword list settings is not set up to a higher minimal threshold.

Comparison of results from the keyword list and the wordlist: abbreviations and specialized vocabulary used in Glossary 4. In this discussion we would like to point to several findings resulting from the comparison of the generated wordlist and the keyword list regarding abbreviations and specialized vocabulary that served as a basis for creating Glossary 4.

Table 3 shows very similar result to the analysis of proper nouns: eight out of ten words in this category would never appear in the HFWL list of the top 3000 keywords. While in case of proper nouns their appearance on the HFWL list was deemed unwanted, the situation takes a swift turn in case of specialized words and abbreviations used in Glossary 4 that tend to be less frequent, but carry a significant portion of meaning.

Similarly as proper nouns, abbreviations and specialized vocabulary would not normally appear in the 3000 most frequent words identified in the HFWL due to their low frequency. However, thanks to the statistical keyword analysis, they ranked higher in the keyword list because they had even lower frequency in the reference BNC corpus.

Thus we could conclude that identifying specialized terminology using keyword analysis is a double-edged sword. On one hand, the researcher might identify words that would not normally appear in the word frequency list, but he or she must count on manual weeding out undesired words like proper nouns. Should the researcher refrain from a sensible keyword list analysis by setting higher the minimal occurrences limit, he or she might risk not identifying the specialized terminology hidden in the low frequency spectrum.

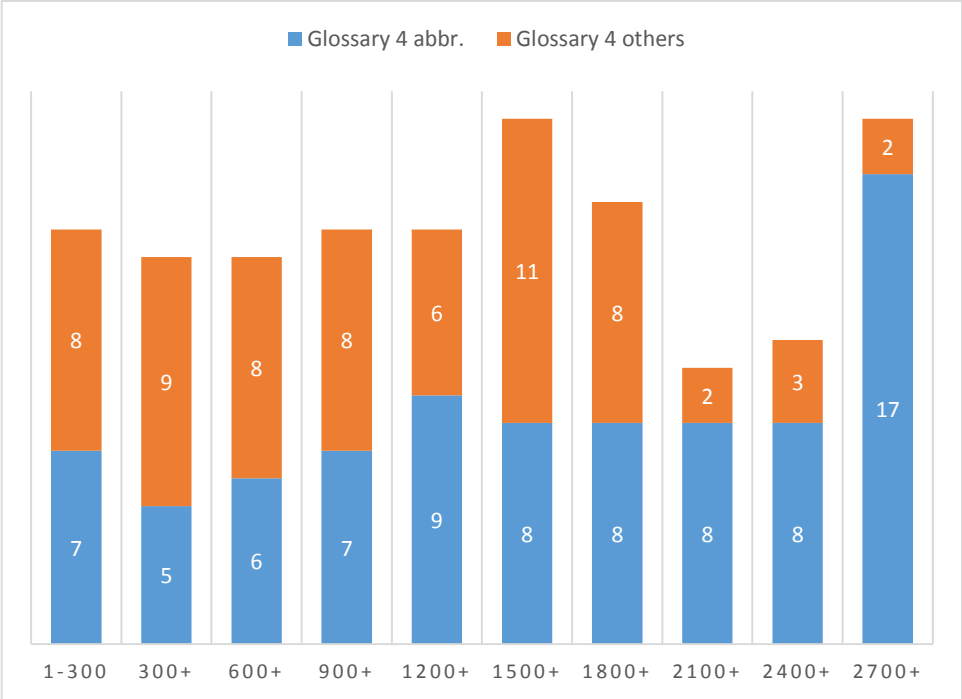
Table 3. Statistical results of a manual selection from the proper nouns identified in the analyzed 3000 words (words above 3000 that would not come to a typical HFWL list are in italics).

Word (roughly every 300. word)	**KW rank**	**HFWL rank**	**Occurrences**	**BNC occurrences in a comparable sample of 60 million words**	**BNC corpus vs. EP corpus: %**
WTO	95	1025	6136	6	0.11%
OSCE	520	*3244*	1263	1	0.09%
GDP	692	2062	2502	488	19.49%
GMO	1135	*5423*	513	1	0.11%
FYROM	1246	*5904*	436	0	0.00%
OECD	1701	*3873*	930	245	26.36%
ACTA	1949	*7358*	276	8	2.74%
Tobin	2183	*7113*	298	22	7.43%
Sharia	2640	*7920*	237	22	9.11%
ecolabel	2787	*10351*	137	2	1.18%

The words for this comparison were manually selected for we did not expect some specialized abbreviations to be found in the BNC corpus. Still, many of these words were very rarely distributed in the BNC corpus which confirms that they belong to the specialized vocabulary of LSP.

Abbreviations and specialized vocabulary seem to be equally distributed through the whole 3000 analyzed keywords. The abbreviations start to dominate in the last 1000 words.

Figure 18. Distribution of specialized vocabulary for Glossary 4 in 3000 analyzed keywords.

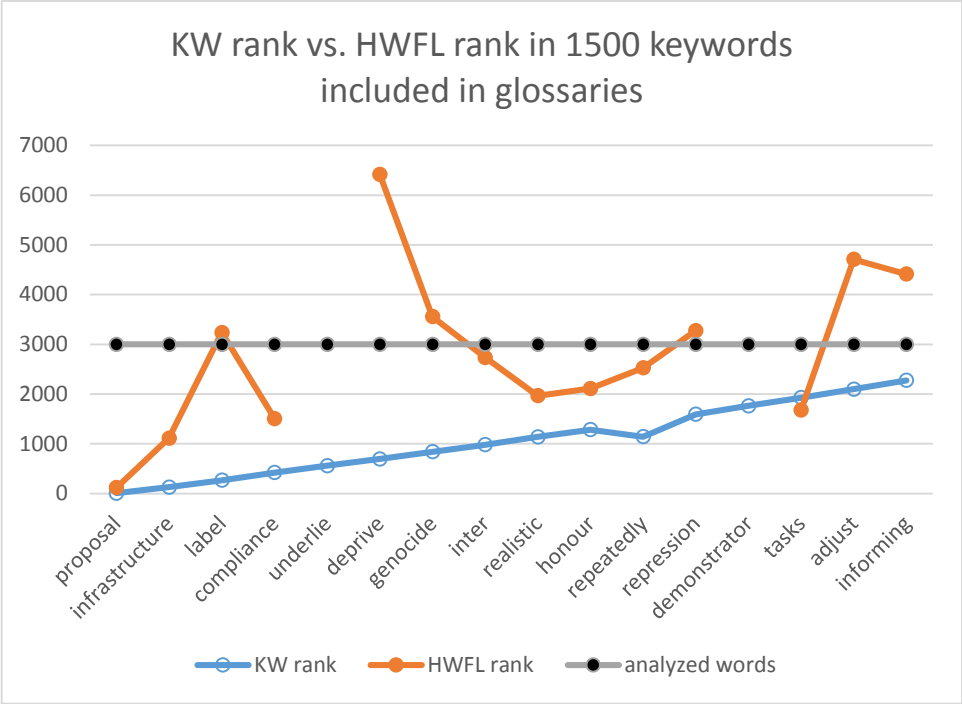


Comparison of results from the keyword list and the wordlist on the selected 1500 keywords. The last comparison that we would like to draw is the comparison of selected keywords to the generated HFWL list and the BNC corpus. In the table below we can see a slight change in trends that could be observed in comparing the HFWL and the keyword list of abbreviations and proper nouns. While in their cases nine out of ten words would never appear in the HFWL of 3000 most frequent words, nearly half of general words included in the top 1500 keywords would also appear in the HFWL. Since the identified keywords are 85% built from general vocabulary (see chapter *vocabulary acquisition*), here categorized according to Klein in the general interaction vocabulary, such an outcome was to be expected. The result supports Zipf's law (see page 56) and categorization of vocabulary by Nation (see page 25-27) which both suggest that the most frequent words creating the basic structures in a language are the more common ones. Please note that the comparison with the BNC corpus in the table below is purely informative as many words may have other semantic meaning thanks to which they gain better ranking (e.g. the word *proposal* in the EP Corpus: mostly in the sense of a suggestion; BNC corpus: other meaning, such as wedding proposal).

Table 4. Statistical results of a manual selection from the top 1500 unique keywords identified in the analyzed 3000 keywords (words above 3000 that would not come to a typical HFWL list are in italics).

Word sample of every 100 word + the 1st keyword	**KW rank**	**HFWL rank**	**Occurrences**	**BNC occurrences in a comparable sample of 60 million words**	**% of EP Corpus in BNC corpus**
proposal	4	116	77463	2226	2.87%
infrastructure	127	1112	10160	528	5.20%
label	267	*3236*	7818	1069	13.68%
compliance	420	1504	3940	691	17.53%
underlie	559	*12678*	1684	102	6.03%
deprive	695	*6413*	2683	177	6.60%
genocide	840	*3557*	1050	66	6.27%
inter	981	2734	1656	306	18.46%
realistic	1138	1965	604	988	163.61%
honour	1283	2111	676	1416	209.53%
repeatedly	1140	2525	1152	668	57.98%
repression	1591	*3277*	744	362	48.70%
demonstrator	1761	*5797*	461	45	9.72%
tasks	1928	1678	231	1939	839.45%
adjust	2098	*4708*	405	576	142.27%
informing	2275	*4411*	615	228	37.14%

Figure 19. KW rank vs. HFWL rank in a sample of 1500 keywords.



Discussion. Through filtering out the 3000 most frequent words of the BNC corpus by using the stop list, the keyword list generator identified all other most frequent words and compared them to the BNC corpus, in which, as you can see, were mostly represented by a substantially lower number. It is hard to say whether the HFWL or they keyword analysis is the better method for selecting subtechnical words (page 25) for general interaction in a specialized field. After careful comparison of the HFWL to the keyword list, the reader can see the actual difference in the general vocabulary between ranking in the HFWL list and the keyword list starts to wane after the first 3000 words (14 out of 16 words in the sample of the 1500 top keywords are ranked in the first 6000 high frequency words). Therefore, to create a large glossary focused on the general interaction vocabulary, it would be wiser to either to set the minimal threshold for a keyword for 300 in a corpus of a similar size, or to solely use the wordlist generator with the stop list. Of course, this is only a speculation based on the results of this corpus analysis. In our view, the decision to use either the wordlist or the keyword list generator for creating a glossary for general interaction vocabulary in a particular LSP context depends on many factors such as the number of words intended for the glossary, the size of the corpus, as well as on the settings used in the wordlist and keyword list generation

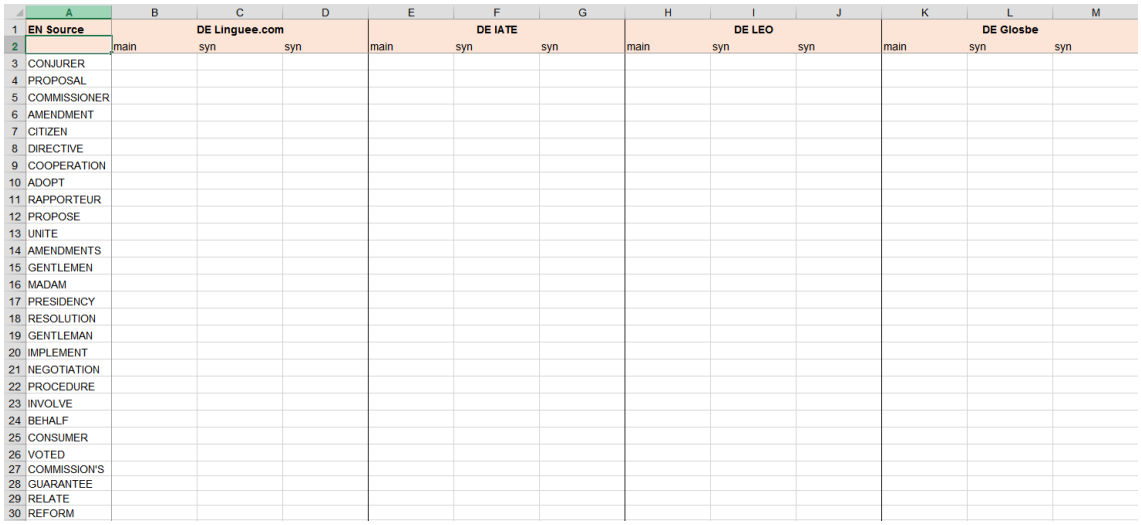
process. Considering that this was a bold experiment, we cannot provide advice or guidelines about the proper approach – we merely wanted to show that it is possible to identify important lexical elements for a glossary compilation from a corpus analysis and to encourage other scholars to use it for similar purposes. Regarding the Europarl corpus analysis in which we have analyzed 3000 words and included 1500 in glossaries, we conclude that identifying the general interaction vocabulary was in the top range of identified words (1500-3000) in sum slightly more effective by using the keyword analysis than by creating a classical HFWL list. Moreover, combining the keyword analysis with the manual selection, we managed to extract important abbreviations and terms that would have otherwise been omitted in a classical HFWL.

Keyword list: target language equivalents. The selection process of words for the glossary set does not only consist of selecting keywords, but also from selecting correct language equivalents that can be used in the glossary. This is an intrinsic task that requires consulting all the language resources introduced in chapter about linguistic sources. We have decided that it is more reasonable to obtain equivalents for a word from the language resources in one table than searching language resources individually during the glossary compilation. It is true that looking up language equivalents in the language resources for all top 1500 keywords, 1500 x 10 language resources x 3 entries = 60 000 entries respectively, is a difficult and time consuming task. Nevertheless, we believe that the additional value created prevails the time and effort invested into this process - in the end, each of the 1500 keywords will require a target language equivalent at least for Glossary 1. Besides, this procedure has two major advantages: Firstly, stocking all target language equivalents from presented language sources together should provide solid reference for decision making in selecting the right ones. Secondly, the created summary of target language equivalents obtained from language resources brings more transparency to the whole decision making process.

For this purposes, a separate Excel sheet was created. There were three columns assigned to each language resource for including the first listed equivalent and two other synonyms if applicable. Some language resources (e.g. IATE) provide translation equivalents for all our target languages, but some do not. This is the case for Linguee.com that provides revised dictionary entries for German, but not for Slovak and Czech (these are only supported by a sentence-aligned parallel corpus), and the dictionary Lingea that

provides translations from English into Slovak and Czech, but not to German. Similarly, some words could not be found in the language resource (this applied mostly to IATE that included specialized terminology rather than general English words). If this occurred, the column is marked as N/A. A layout of the excel sheet can be seen in figure 20 and the file itself can be downloaded from supporting electronic documentation.

Figure 20: Screen shot of the summarizing excel sheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	EN Source	DE Linguee.com			DE IATE			DE LEO			DE Glosbe		
2		main	syn	syn	main	syn	syn	main	syn	syn	main	syn	syn
3	CONJURER												
4	PROPOSAL												
5	COMMISSIONER												
6	AMENDMENT												
7	CITIZEN												
8	DIRECTIVE												
9	COOPERATION												
10	ADOPT												
11	RAPPORTEUR												
12	PROPOSE												
13	UNITE												
14	AMENDMENTS												
15	GENTLEMEN												
16	MADAM												
17	PRESIDENCY												
18	RESOLUTION												
19	GENTLEMAN												
20	IMPLEMENT												
21	NEGOTIATION												
22	PROCEDURE												
23	INVOLVE												
24	BEHALF												
25	CONSUMER												
26	VOTED												
27	COMMISSION'S												
28	GUARANTEE												
29	RELATE												
30	REFORM												

The process of selecting the terms for glossaries from gathered language resources can be simply described as highlighting the most frequent target language equivalents. The selection was later used as a source for glossary compilation of Glossary 1 that later served as the main reference source for the other glossaries.

Phase 3: Completing Glossaries through Additional Relevant Information

This chapter discusses the final phase of glossary compilation that consisted of adding additional information to glossaries 2, 3, and 4. Each of these glossaries follows a certain purpose that was explained in chapter about glossary templates. In this chapter we would like to describe the actual process of glossary compilation. Searching, evaluating and adding relevant information to the glossaries was also an arduous and time consuming task requiring meticulous effort

Additional information provided in all four glossaries can be divided into three categories: statistical information (e.g. number of occurrences, word rank etc.) obtained

from the corpus analysis, additional linguistic information (e.g. synonyms, n-grams, KWIC sentence examples), and reference information about the source of included additional information. Each type of information has its own specific purpose.

Statistical information serves for a reference about importance of a particular word in the source text. We considered it useful to include statistical data in Glossary 3 that provides very detailed information about a narrow selection of the most frequent words. These words might have different ranking in the keyword list, yet very high importance from a lexical point of view as they can be combined into common noun-verb phrases, expressions etc. Furthermore, the decision was made to include basic statistical information in Glossary 4 that is intended to explain various proper nouns and terms identified in the corpus analysis. Such statistical information can give an interesting insight into basic figures obtained from the corpus analysis which would otherwise remain unknown to the target user of the glossaries.

Additional linguistic information represents important linguistic knowledge that may benefit the learner for the lexical reasons mentioned in the previous paragraph. This includes information about admitted, preferred and deterred words in Glossary 2, synonyms, collocations, N-grams and KWIC sentences in Glossary 3 and definitions with additional notes in Glossary 4. This should all serve for highlighting important connection worth remembering.

Reference information is particularly important for the verifiability of compiled information in the glossaries. References to sources are provided either in hyperlinks in glossaries or labeled with numbers linked to their reference.

Glossary 1. The first glossary was compiled from summarizing the Excel sheet described in the previous chapter. While this Excel sheet served as a basis for creating Glossary 1, this glossary was further used as a basis for target language equivalents in glossaries 2, 3, and 4. Glossary 1 was simply built by choosing 3-4 preferred translations from various language sources summarized in the Excel sheet and referring to their source. The words are sorted by their keyword rank, not alphabetically.

Compilation of Glossary 1 required two steps: Firstly identifying the preferred translations and synonyms in the target language, and secondly, creating a format compatible with CAVL programs. As mentioned earlier, our intention was to create three formats for this glossary: an Excel sheet that serves as a documentation of English words,

their target equivalents and used language sources, a two databases for learning purposes: a database for InterpretBank, and a database for Anki. The exact procedure could be described as follows:

1) Preparing the excel sheet for import: Consisted of deleting the information sources, merging the 3-4 columns for each target language into one column by using the function *concacenate* and adding " / " to separate the words merged from individual rows.
2) Importing the adjusted copy of Glossary 1 into Anki: Before the actual import, the table had to be saved in a text file firstly in Excel and subsequently again re-saved in UTF-8 format in Windows Notepad for each language combination (EN-DE, EN-SK, EN-CZ. The combination of more languages, e.g. EN-DE-SK could not be created. When the glossary was successfully imported, it was saved in Anki deck package format (*.apkg). Afterwards, the Anki deck was uploaded to the public database on AnkiWeb[42] under the name Europarl Corpus: The top 1500 most frequent keywords in the European Parliament.
3) Importing the adjusted copy of Glossary 1 into InterpretBank: The process was very similar to importing the glossary into Anki. There was no need to separately create more language combinations as InterpretBank is capable of importing multilingual glossaries. The glossary was imported straight from an Excel sheet and saved in the TermMode's exchange format (*.TMEX).

We are also particularly happy to confirm that Claudio Fantinuoli, the creator of InterpretBank, agreed to share and promote our glossary on the InterpretBank website. We believe that this step will lead to dissemination of the glossary for the target audience of student of conference interpreting who can not only test the functionality of the program, but take the advantage of learning the vocabulary in the MemoryMode jazykom (C. Fantinuoli, personal communication, December 3, 2015).

Glossary 2. The second glossary (Glossary 2), consists of selected internationalisms that may or may not be translated in a similar form across languages. It must be highlighted that the focus by identifying potential problem triggers laid on Slovak and Czech interferences. German had a purely informative function.

[42] https://ankiweb.net/shared/decks/

The whole procedure of compiling the glossary consisted of several steps. Firstly, all keywords that could come into account as internationalisms were manually selected by assigning them to glossary category 2 during the initial corpus analysis. The manual evaluation had rather low sensibility to minimize the risk of omitting important words – it was found more feasible to select improper keywords with lower chance to become internationalisms than risking leaving out the more important ones in the initial selection. The identified 384 words from the unique 1500 keywords in this initial selection were further divided into following categories:

1. International word codified in the target language (Slovak/Czech) for which there is no other equivalent, thus no peril of incorrect usage (e.g. *democracy*);
2. International word codified in the target language, but with a potential for semantic shift (e.g. the word *assistance* can be suitable for some context in the target languages, but for some not). The Speaker must be aware of semantic differences while using these words to avoid misunderstandings;
3. High risk of interference or false friends. This category includes words that should not be used in an international form, but should rather be replaced by more suitable target language equivalent;
4. Low risk of interferences. These words do not pose high risk as the third category and their usage would sound unnatural except for usage in specialized fields;
5. Minimal risk of interferences. Words that sound very unnatural when rendered as Slovak internationalisms. Still, there might be a chance that the speaker uses them due to high cognitive overload.

The categories were revised 2 times by the author and the selection of words from category 2 and 3 (164 words) were further consulted with PhDr. Jana Rejšková from the Centre of Translation studies at Faculty of Philosophy and Arts of the Charles University in Prague. The reason of the consultation was to identify which of the selected words are prone to be used as internationalisms and could not be used in formal Czech language, and/or which of them pose a risk of incorrect usage for interpreting students. Ms. Rejšková further divided all words from category 2 and 3 into:

1) Words that could be used interchangeably in Czech language;
2) Words that could be used as internationalism only in certain context while in other context should be avoided;

3) Words that should under no condition be used as derived internationalisms in the Czech language.

We must express our sincere thanks to both Ms. Rejšková for her kind support of our endeavor to create an effective tool to promote correct language usage.

Since Slovak and Czech are very similar languages, the result of this consultation was incorporated in Glossary 2 for both languages and served as a basis for further investigation. The investigation consisted of searching linguistic resources issued by the *Ľ. Štúr Institute of Linguistics* for Slovak, and the *Institute of the Czech Language* for Czech[43].

Compiling the glossary required considerable attention in the scrupulous investigation of codified, admitted and deprecated terms. The final version of Glossary 2 consists of 51 words that a) might have a different meaning in the target language, b) should not be used as internationalisms. Appendix B contains a sample of the glossary and the full list of words. The full version of the glossary can be downloaded from the electronic link included in appendix A.

The gradual weeding out procedures applied in Glossary 2 was not the case for Glossary 3, for which we intended to include only the top 100 keywords, and Glossary 4 in which we included abbreviations from the top 1500 unique keywords.

Glossary 3. Glossary 3 is by far the most comprehensive glossary from the whole glossary collection. The ambition was to provide an in-depth insight into particularities of narrow selection of words that were intuitively identified as words rich for collocations. We believe that this glossary can provide specific information about words usage and consequently create higher confidence in speakers who study the glossary and acquire them in their active vocabulary. Although mainly aimed at English language (definitions, number of synonyms, N-grams, collocations and context usage), Glossary 3 provides

[43] Mainly consulted sources:
SK: http://slovnik.juls.savba.sk as the main codified dictionary for Slovak, periodics; journals: Linguistic Journal, The Culture of the Word, Slovak Speech
CZ: http://ssjc.ujc.cas.cz/, http://www.neologismy.cz/, http://bara.ujc.cas.cz/ssjc/, http://bara.ujc.cas.cz/psjc/

useful information in three other languages as well. This was achieved by consulting several linguistic resources.

The selection of words for Glossary 3 went through three stages. Firstly, the words were categorized by the initial analysis of the top 3000 keywords. Secondly, the selected words were sorted alphabetically to compare words next to each other in various inflections (e.g. endorsement, endorse, endorsed) and consequently to decide which part of speech the word would be most suitable for representing its semantic group in the glossary. Thirdly, the selection was examined once again to narrow down the number of words to the final 100. For this purpose, the filter in column R was set up in which we firstly marked all the top 1500 keywords in 5 and then in 9. These attributes allowed us to swiftly change the selection during the glossary compilation process if some words were later regarded as less important. Now the column R contains only mark 9 that was used for the final selection of the most interesting 100 keywords from the top 1500 unique keywords.

The created template described in chapter *glossary templates* was used for each word in the glossary. The template was slightly changed by leaving out some categories for which there was no reliable information to be found. For each word in the glossary, we firstly added language equivalents and synonyms for a word in target languages from Glossary 1. Then we looked for English synonyms in the online Collins dictionary, definitions (for the political discourse non-relevant ones were left out) and English collocations in the online Oxford dictionary. Then we combined the dictionary search through consulting Linguee.de parallel corpora resources and the Leo dictionary. The results provided us with suggestion for the most common collocations in English and German. Afterwards we tried to find their language equivalents in Slovak and Czech which in fact proved to be the biggest challenge in compiling Glossary 3. Finding German translations was not difficult due to a large database of Linguee.de for this language combination, but for Slovak and Czech this process showed to be rather difficult. Therefore, we applied cross-search (English>>German>>Slovak/Czech) and occasionally had to confirm the results with Google if a common phrase for Slovak and Czech could not be found in Linguee. From our experience we daresay that working with German-English parallel corpora on Linguee.de is thanks to a large database much more

convenient and effective than using it for Slovak and Czech. This collocation search was a time consuming process, but we believe that the result is worth it.

For identifying N-grams, the demo version of ParaConc was used instead of AntConc and WordSmith due to its quick response time. Clearly, using WordSmith and AntConc with the full corpus of 54 million words would be more suitable than searching the parallel Slovak-English corpus of 15 million words, but initial comparison of all three programs showed an insignificant difference between using the whole EP Corpus and a 3,5 times smaller parallel EN-SK corpus by such vast amount of words.

The process of searching words in context also consisted of searching the Europarl corpus through ParaConc in which the interesting results in KWIC were further investigated in context and later copied to the glossary. We added 3-4 sentences for each word to provide better understanding of its usage in context. The glossary also includes references to each language resource and basic statistical information from the quantitative corpus analysis (word occurrences, wordlist rank, and keyword rank).

Glossary 3 does not contain basic grammatical information such as part of speech category, plural forms in nouns and present/past participle in verbs. We concluded that this basic information is known to the advanced English learner and thus redundant.

Glossary 4. Glossary 4 can be seen as a typical LSP glossary displaying the term, translation equivalents in target languages, English definition and some additional information if necessary. The glossary contains 97 out of 143 keywords that were initially identified as appropriate in the first analysis of 3000 keywords The rest of the keywords was either later assessed as inappropriate or were outside the spectrum of 1500 unique keywords. The original aim was to keep the glossary compact, but then we have come to the conclusion that displaying the key information on the left and additional information to the term on the right would be the best procedure for striking the balance between information coverage and compactness.

Glossary 4 contains reference to ensure better verifiability of provided information. The glossary is also intentionally not sorted alphabetically, but by the keyword list rank of respective words so the reader can start studying the most important words first. It also occasionally offers correct language usage of the selected term. The major information source for language equivalents was the IATE database that includes admitted, preferred and deterred words to some terms. IATE proved to be the best source

for language equivalents for various proper nouns used in the EU context. Definitions and additional information for glossary words were gathered from the IATE, official websites of institutions or initiatives represented by the glossary word, and in other relevant and reliable information sources found on the internet. Glossary 4 mainly consists of abbreviations and the vast majority of the words in this glossary would normally not appear in the HFWL. Also, it contains reference to some obsolete terms identified in the corpus analysis and provides information about the preferred usage and the current state of affairs.

Based on these results, we conclude that Glossary 4 mostly benefited from the glossary compilation procedure based on the keyword list generation. However, a note of caution is due here since many terms consisting of two or more words could not be identified through the keyword analysis and consequently did not appear in the results (e.g. the European Union, the European Commission etc.). Even though such terms are mostly part of the fundamental vocabulary in political context, the limitation of corpus-based lexical keyword analysis cannot exclude the possibility that some less frequent, yet important compound words or terms were left out. It would be possible to tackle this issue through word cluster examination, but this procedure did not lie in the scope of this paper.

Discussion

The glossaries compilation required in effect considerable more time and effort than was initially anticipated. Copying about 60,000 dictionary entries in an Excel sheet was a monotonous and heavy, yet inevitable task. Similarly, creating the glossaries required meticulous effort in searching and comparing relevant information. The biggest challenges encountered by compiling Glossary 1 were selecting the correct synonyms, by compiling Glossary 2 consulting language resources for preferred, admitted and deterred terms (the language resources available online showed to be ineffective in many cases), in Glossary 3 it detailed search for identical collocations in all four languages together with entering a substantial amount of information, and finally, in Glossary 4 the biggest challenge was to strike the balance between the amount of information given for creating a better overview about the respective term and the compactness of the glossary. In the end, the final results of our work is following: Glossary 1 contains the top 1500 English keywords, Glossary 2 51 tricky internationalism, Glossary 3 100 manually identified

keywords of high importance, and Glossary 4 contains 100 abbreviations or terms related to the EU political discourse.

We stay positive that creating LSP glossaries from a keyword list based on a quantitative corpus-based lexical analysis is a viable, if not advisable option, but some drawbacks suggest that there is ample room for improvement and need for caution by using this method. Here we would particularly like to address the following issues: the up-to-dateness of this corpus analysis from the perspective of a discourse & lexical analysis, lemmatization, the interpretation of statistical occurrences of identified keywords, and finally, the pros and cons of the keyword analysis.

Speaking about the up-to-dateness, the Europarl corpus consists of texts over a long time spanning from April 1996 until November 2011. The corpus could hence serve as a very useful resource for discourse analysis of the European matters across various time periods. It is reasonable to expect that main issues discussed in 1996 could be different from the year 2000 which could be statistically proven by applying the keyword list analysis. Our keyword analysis could serve well as a springboard for identifying topic words in the Europarl corpus. This could be achieved by implementing the results of our quantitative corpus analysis, the manual selection of the general interaction vocabulary, and combining the selection with the stop list of the 3000 most frequent words in the BNC corpus. A further study with more focus on the political discourse analysis could reveal interesting findings and is therefore suggested. However, one important thing to keep in mind here is that the keyword analysis ranks words on their statistical keyness and not on the number of occurrences. Therefore, discourse analysis researchers might rather want to compare wordlists than keyword lists. Discourse analysists must also take in consideration two shortcomings of the Europarl corpus: the lack of up-to-date data (December 2011 – present) and possibly a big size of the corpus whose analysis may lead nevertheless to overly general results. Speaking about the currency of the corpus, unfortunately, there are no more releases planed since the proceedings are no longer being translated into all official languages (Koehn, personal communication, April 10, 2015).

Regarding the up-to-dateness from the lexical perspective, we conclude that the lack of recent data (December 2011 – present) should not pose a problem. The only concern is the obsoleteness of some terms included in Glossary 4 and at the same time the absence of new terms which have appeared recently or have started to be used more

frequently in political debates (ISIL, TTIP etc.). Still, we can proceed on the assumption that the learner studies on his or her own initiative over the current course of events. After all, from the learners' perspective it is important to firstly grasp the general interaction vocabulary for active language use and then to learn specific LSP vocabulary important for the respective topic.

In the case of lemmatization, we must admit that the lemmatization process did not result in leaving the words from the generated word list and keyword list as originally intended. The outcome can be regarded from two perspectives as either positive or negative. The positive is that since the wordlist and the keyword list did not treat the words in one group and did not list only one representative of the whole morphological family (e.g. propose, proposal, proposed), all such words with high keyness could appear in Glossary 1. As mentioned previously, Glossary 1 consists only of the word in the source language and few equivalents in each target language. Should the morphologically similar words be represented only with one word from the whole group, the learner would be required to think of all words from the same morphological group in the target language which would go against the original intention to help automatize the political vocabulary. Therefore, on one hand, there are more similar words in Glossary 1, on the other hand they can be easier learned with their exact target language equivalents. There could be a clear objection raised that the advanced learner should already be able to form morphologically similar words after learning only one of them, but as we mentioned in chapter about vocabulary acquisition, what is not automatized, cannot be fully applied. For this reason, learning these words by heart is advisable for advanced learners without previous experience in political language who want to dive into and acquire the general interaction vocabulary. The not-lemmatization of the keyword list also created room for deliberate decision which words of the same morphological group will stay and which will not.

Speaking about the disadvantage of the unsuccessful lemmatization process, it is crystal clear that these words created additional burden (initial categorization of the keyword list, careful examination of alphabetically sorted keyword lists, unnecessary translation of these words etc.). It is difficult to decide whether the benefits outdo the additional effort. In any case, we are not entitled to provide a statement on this because

the glossaries and their effectiveness is based purely on theoretical assumptions and there have been never experimentally tested in practice.

As regards statistical significance of the occurrence of words included in the top 1500 keywords, the reader might be under the impression that firstly, the words included in glossaries occur so rarely that it makes no sense to include them in glossaries, and secondly, the selection of the top keywords should have been limited to a fraction of this amount. The statistical data obviously support this view. If we consider the first keyword "propose" with 77,463 occurrences in 54 million words of the Europarl corpus, it shows us that the top keyword from our list appears on average once in every 696 words. The word "stringent", the 750th word on the keyword list with the KW rank 1058, appears in the whole corpus only 1392 times, which shrinks its presence on average once in vast 51,000 words. If we consider the average speed of spoken language at 120 words per minute, the word would come up every 7 hours. The last word "recourse" on Glossary 1 with the KW rank 2266 and surprisingly higher frequency of 2342 occurrences appears on average once every 23,000 words. These statistical results might seem tremendously discouraging for learners willing to automatize words in glossaries on the first look. However, one must consider the importance of words in the overall context, their more often repetition in some contexts. Perhaps more importantly, these results should be compared to the relatively low frequency of functional and fundamental words. For instance, the most basic word "the", an almost indispensable article to any sentence yet carrying no meaning, appears in the BNC corpus once every 16 words. The 10th most frequent word in the BNC corpus "it" appears once in every 108 words, and the 100th most frequent word once in every 1043 words. Considering that these and remaining 2997 were used to create the stop list of 3000 other most frequent words that were filtered out from the keyword analysis as too general or simple, it is no wonder that the words included in top 1500 keyword list show rather astonishing low frequency. Therefore, it is important to look at statistics wisely and with certain caution and not to jump to conclusions that these words are too rare to create a useful language resource for learners. The words included in the top 1500 keywords might show rather low frequency in absolute terms, but are important carriers of meaning and as the keyword analysis shows, play an important role in the political discourse of the European Parliament that can be partially considered as a language for special purposes.

The last thing to which we would like to give particular attention to in this chapter is the effectiveness of the keyword analysis and its application in glossary compilation. Here, caution is also advised. As mentioned in the theoretical part, the keyword list is a method that yields statistical results according to set parameters. The decision to use the keyword list analysis depends on many factors that greatly influence results. Yet, even if we compare the best case scenario of an optimal investigation and reference corpus with the most optimal settings, there are still some perils to be aware of. In the theoretical part, we listed on page 112 several problems arising from the statistical nature of computing keywords in quantitative corpus analysis. Here at the end of the practical part, we should like to provide a warning that would otherwise stay unnoticed if left in the first theoretical part. It is a caution of statistical corpus analysis of keyword list provided by the author of the statistical keyword list analysis R. Scott:

> Suppose you process a text about a farmer growing 3 crops (wheat, oats and chick-peas) and suffering from 3 problems (rain, wind, drought). If each of these crops is equally important in the text, and each of the 3 problems takes one paragraph each to explain, the human reader may decide that all three crops are equally key and all three problems equally key. But in English these three crop-terms and weather-terms vary enormously in frequency (chick-peas and drought least frequent). WordSmith's KW analysis will necessarily give a higher keyness value to the rarer words. So it is generally unsafe to rely on the order of KWs in a KW list (Scott, 2015, para. 5).

This elaboration clearly shows that the keyword analysis might be unreliable for in-depth discourse and lexical analysis due to *false positives*. However, from the statistical point of view, there are no false positives or neglected keywords, only words with lower and higher keyness based on the investigation and reference corpus that are sorted in the keyword list accordingly. The keyness by itself is not a problem, it is the human interpretation that might lead to erroneous conclusions. Nevertheless, it must be underlined that such risk was not primarily a concern of our corpus analysis - the bigger the corpus and the number of analyzed words is, the lesser the risk of false interpretation should be. Still, we consider it wise to conclude our discussion with highlighting the risks of the method for glossary compilation that we tried to defend throughout the whole thesis.

Conclusion

This master thesis has argued that the utilization of corpus linguistics in general and the keyword analysis in particular can be a useful way to create LSP glossaries. The decision to use corpus linguistics for glossary compilation resulted from the original endeavor to create a useful tool that could help students with the acquisition of EU related vocabulary. We have hypothesized that a multilingual glossary of the most frequent words from the political discourse of the European Parliament could help master the most important vocabulary of the EU parliamentary context. Throughout the thesis the inductive approach was applied - the search after an optimal vocabulary learning tool represented the central research question that we tried to address by consulting relevant scientific disciplines and applying their knowledge.

Based on the review of the theoretical knowledge from English teaching, we confirmed that learning high frequency vocabulary lists is a common, useful and preferred practice and that foreign language speakers need to expand their active vocabulary in order to produce spoken language output more effectively. Further investigation into speech science to determine whether vocabulary automatization may lead to improved spoken language output supported our assumption. After examination of different speech production models and linking them to relevant cognitive models, we postulated a synthetized model that incorporates both the speech production process and the concept of shared attention. This model shows that increasing automatization on one subtask of the speech production leads (in our case on the lexical level) to redistribution of saved cognitive capacities to other subtasks, and may consequently result in improved speech production quality.

The abovementioned created a solid theoretical background in favor for creating a high frequency word list from the EU political context. Further investigation seeking to answer the inherent question of how to create this tool brought us to the Europarl corpus that contains verbatim reports of the proceedings from the European parliament, and to corpus linguistics as a method for utilization corpus resources. We have learned that generating a keyword list would be in many respects more suitable for LSP purposes than classical high frequency word lists as it can fully harness both the potential of statistical

analysis and the sheer size of the Europarl corpus. Yet, it also carries certain perils that a corpus researcher must be aware of.

After conducting the quantitative corpus analysis of the Europarl corpus, it was important to correctly interpret and utilize the corpus analysis results. For this reason, scientific disciplines as lexicology, terminology and language for special purposes with a special emphasis on politolinguistics were consulted. The gathered theoretical knowledge helped us make sound decisions which included creating more glossaries to address some specific issue such as the selection of keywords for glossaries, and the template & distribution of glossaries.

We understand that we have only scratched the surface of corpus linguistics and that it is much more complex than the elaboration introduced in this master thesis. We also cannot state that we have successfully overcome all hidden risks and issues that we confronted in the process of glossary compilation because a) the expertise in consulted scientific disciplines in our inductive problem-solving approach to the research question is missing, b) the risk of human error and false interpretation cannot be fully eliminated. Still, we do hope that our inductive approach directed us the right way in our objective to provide all necessary safeguards for the correctness of the glossary.

In regard to evaluating the corpus linguistics as a method for LSP glossary compilation, we stay positive that it has a wide application not only in terminology and lexicology, but also in language teaching as well as in translation and interpreting studies. In fact, we highly recommend incorporating corpus linguistics in classes for students of translation and interpreting in a form of a practical training that seeks to solve some questions closely related to translation and interpreting studies. From the student perspective it can be said that all things that are covered only on a theoretical basis in form of lectures without practical trial exercise during the studies will soon be forgotten. Corpus linguistics is in our mind a method in which students should be actively engaged and trained.

It must be highlighted that the Europarl corpus is an important resource that leaves ample room for further studies including in-depth lexical, semantic and discourse analysis. Our aim was only to extract the most frequent words and use them as a basis for glossary compilation, but the corpus offers a fruitful area for further work. Unfortunately, apart from machine translation and natural language processing for which the corpus was

originally compiled, only little attention was devoted to the Europarl corpus from the linguistic community. Therefore, further studies on this matter are suggested.

The result of our master thesis is a set of four glossaries that should cover the most frequent vocabulary from the European parliament. The first glossary consisted of 1500 unique keywords with several target equivalents in each language (EN/DE/CZ/SK). The second glossary deals with internationalisms that may pose a risk in correct language use. It describes the correct usage of 51 words prone to false usage as mistaken internationalisms. The third glossary can be regarded as the most comprehensive glossary as it provides in-depth insight into the correct use of the keyword in question. It includes word definition, synonyms, and collocations in each language, the word use in context as well as some statistical data. The final fourth glossary was created as an answer to keyword analysis findings that there is a considerable number of words that would normally not appear in a high frequency wordlist but are important carriers of meaning and clearly require further elaboration. Therefore, this glossary provides explanation of such keywords in the form of definitions and additional information supported by statistical information and target language equivalents. Glossary 1 is the main glossary that includes all 1500 top keywords while glossaries 2, 3, and 4 consist of a narrow selection and fulfill a specific purpose. In regard to future lexical studies of the Europarl corpus aimed at similar practical learning application, further investigation into the use of internationalism and their correct usage in respective target languages would be worthwhile.

The results of the problem-oriented theoretical review of relevant scientific disciplines confirmed our hypotheses, namely that a) a glossary of the most frequent words can be a useful tool for language learning, b) learning vocabulary and its subsequent automatization may lead to better speech production; and partially confirmed the third hypothesis that c) political language of the European Parliament is distinct through its special vocabulary and it can be classified as a language for special purposes.

We believe that through validating the research hypotheses and creating the glossary set we have fulfilled our research objective and that our work will be an asset not only for students interested in the most frequently used words in the European Parliament, but also for the academic community for the problem-oriented approach adopted in this paper.

List of References

Aboelela, S. W., Larson, E., Bakken, S., Carrasquillo, O., Formicola, A., Glied, S. A., . . . Gebbie, K. M. (2007). Defining Interdisciplinary Research: Conclusions from a Critical Review of the Literature. *Health Services Research* (42), 329–346. doi:10.1111/j.1475-6773.2006.00621.x

AFP. (2014, March 8). Pentagon studying Putin's body language. *The Telegraph*. Retrieved 12 21, 2015, from http://www.telegraph.co.uk/news/worldnews/europe/russia/10684769/Pentagon-studying-Putins-body-language.html

Alsaif, A. (2012). Human and Automatic Annotation of Discourse Relations for Arabic (Doctoral dissertation or master's thesis). Retrieved 12 21, 2015, from http://etheses.whiterose.ac.uk/3129/1/Amal_PhD_thesis_November.pdf

Anagnostou, N. K., & Weir, G. R. (2007). Average collocation frequency as an indicator of semantic complexity. *ICTATLL Workshop 2007 Preprints*, (pp. 43-48). Hiroshima.

Anderwald, L., & Szmrecsanyi, B. (2008). Corpus linguistics and dialectology. In A. Lüdeling, & M. Kytö, *Corpus linguistics: an international handbook* (II ed., pp. 1126-1140). Berlin/New York: de Gruyter.

Anthony, L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. *Proceedings of the International Professional Communication Conference*, (pp. 729-737).

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research, 30* (2), 141-161.

Archer, D. (2009). *What's in a word-list?: Investigating word frequency and keyword extraction.* Farnham: Ashgate Pub.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In honour of John Sinclair* (pp. 233-250). Amsterdam/Philadelphia: John Benjamins. doi:10.1075/z.64.15bak

Baker, M. (1994). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target, 7* (2), 223-243. doi:10.1075/target.7.2.03bak

Baker, M. (2001). *Routledge Encyclopedia of Translation Studies.* London/New York: Routledge.

Baker, P. (2004). Querying keywords : questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics, 32* (4), 346-359.

Baker, P. (2006). *Using Corpora in Discourse Analysis.* London/New York: Continuum.

Baker, P. (2012). Acceptable bias?: Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies, 3* (9), 247-256.

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Barlow, M. (n.d.). *Online Searches*. Retrieved 12 21, 2015, from Web concordances: http://www.athel.com/web_concordance.html

Barlow, M. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics, 4* (1), 173–184.

Barlow, M. (2002). ParaConc: Concordance software for multilingual parallel corpora. *Language Resources for Translation work and Research, Proceedings of the Third International Conference on Language Resources and Evaulation.* Las Palmas. Retrieved 12 21, 2015, from http:// www.mt-archive.info/LREC-2002-Barlow.pdf

Bartoňková, H. (2002). *Andragogika.* Olomouc: Vydavatelství Univerzity Palackého v Olomouci.

Basil, M. D. (2012). Multiple Resource Theory. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 2384-2385). New York: Springer US. doi:10.1007/978-1-4419-1428-6_25

Baumann, K.-D. (2013). Fachkommunikative Dolmetschkompetenz aus interdisziplinärer Perspektive. In K.-D. Baumann, & H. Kalverkämper, *Theorie und Praxis des Dolmetschens und Übersetzens in fachlichen Kontexten* (TRANSÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens ed., Vol. 63, pp. 31-49). Berlin: Frank & Timme.

Bendazzoli, C. (2010). The European Parliament as a Source of Material for Research into Simultaneous Interpreting: Advantages and Limitations. In L. N. Zybatow (Ed.), *Translationswissenschaft — Stand und Perspektiven.* Frankfurt am Mein: Peter Lang.

Bendazzoli, C., & Sandrelli, A. (2009). Corpus-based Interpreting Studies: Eary Work and Future Prospects. *Revista Tradumàtica 07 L'aplicació dels corpus lingüístics a la traducció*. Retrieved from http://www.fti.uab.es/tradumatica/revista/num7/articles/08/08art.htm

Berber-Sardinha, T. (1999). Using KeyWords in textanalysis: Practical aspects. *DIRECT Papers* (42), 1-9.

Berber-Sardinha, T. (2000). Comparing Corpora with WordSmith Tools: How Large Must the Reference Corpus Be? *In Proceedings of the Workshop on Comparing Corpora, Volume 9*, (s. 7-13). Hong Kong.

Berns, M. (2010). *Concise Encyclopedia of Applied Linguistics.* Oxford: Elsevier.

Bhela, B. (1999). Native language interference in learning a second language: Exploratory case studies of native language interference with target language usage. *International Education Journal, 1* (1). Retrieved from http://ehlt.flinders.edu.au/education/iej/articles/v1n1/bhela/bhela.pdf

Bianchi, F. (2012). *Culture, corpora and semantics. Methodological issues in using elicited and corpus data for cultural comparison.* Lecce: Università del Salento - Coordinamento SIBA.

Biber, D., Connor, U., Upton, A., Anthony, M., & Gladkov, K. (2007). Discourse on the Move: Using corpus analysis to describe discourse structure. In *iscourse* (s. 121-151). Amsterdam: John Benjamin.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Pearson Education.

Bick, E. (2010). Degrees of orality in speech-like corpora: Comparative annotation of chat and e-mail corpora. *In Proc. of the 24th Pacific Asia Conference on Language, Information and* (pp. 721—729). Sendai: Waseda University.

Blecha, J. (2012). *Building Specialized Corpora (Master's diploma thesis).* Brno: Masaryk University.

Boulton, A. (2010). Data-Driven Learning: On Paper, in Practice. In T. Harris, & M. M. Jaén (Eds.), *Corpus Linguistics in Language Teaching* (pp. 17-52). Bern: Peter Lang.

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora.* London/New York: Routledge.

Budin, G. (2010). Socio-terminology and computational terminology – toward an integrated, corpus-based research approach. In R. e. De Cilia (Ed.), *Discourse, Politics, Identity* (s. 21-31). Tübingen: Stauffenburg Verlag.

Burkhardt, A. (1996). Politolinguistik. Versuch einer Ortsbestimmung. In 1996, J. Klein, & H. Diekmannshenke (Eds.), *Sprachstrategien und Dialogblockaden. Linguistische und politikwissenschaftliche Studien zur politischen Kommunikation* (pp. 75-100). Berlin: De Gruyter.

Cartoni, B., Zufferey, S., & Meyer, T. (2013). Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics, 27* (1), 23-42. doi:10.1075/bjl.27.02car

Cedroni, L. (2010, 11 10). Politolinguistics: towards a new analysis of political discourse. Retrieved 12 21, 2015, from http://sspnet.eu/2010/11/politolinguistics-towards-a-new-analysis-of-political-discourse-2/

Cervatiuc, A. (2008a). ESL Vocabulary Acquisition: Target and Approach. *The Internet TESL Journal, 16*(1). Retrieved from http://iteslj.org/Articles/Cervatiuc-VocabularyAcquisition.html

Cervatiuc, A. (2008b). *Highly Proficient Adult Non-Native English Speakers' Perceptions of their Second Language Vocabulary (PhD thesis).* Ottawa: Library and Archives Canada = Bibliothèque et Archives Canada.

Clark, A., Fox, C., & Lappin, S. (Eds.). (2010). *The Handbook of Computational Linguistics and Natural Language Processing.* Wiley-Blackwell.

Cobb, T. (n.d.). *Why & how to use frequency lists to learn words*. Retrieved 12 21, 2015, from Vocab Research Resources: http://www.lextutor.ca/research/

Cole, J., & Hasegawa-Johnson, M. (2012). Corpus phonology with speech resources. In C. Fougeron, M. Huffman, & C. Abigail (Ed.), *Handbook of Laboratory Phonology. Comp* (pp. 431-440). Oxford: Oxford University Press.

Conference, P. (n.d.). *Political Linguistics Conferences*. Retrieved from Political Linguistics: http://pl.ils.uw.edu.pl/

Conrad, S. (2000). Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century? *TESOL Quarterly, 34* (3), pp. 548-560.

Corpus. (n.d.). In Oxford British & World English Dictionary. Retrieved from http://www.oxforddictionaries.com/definition/english/corpus?searchDictCode=all

Coxhead, A. (1998). An Academic Word List. ELI Occasional Publications #18, School of Linguistics and Applied Language Studies. *ELI Occasional Publications, 18.*

Crookes, G. (1991). Second language speech production research: A methodologically oriented review. *Studies in Second Language Acquisition, 13* (02), 113—32. doi:10.1017/S0272263100009918

Čermák, F. (2003). Today's Corpus Linguistics: Some Open Questions. *International Journal of Corpus Linguistics, 7* (2), 265-282.

Dash, N. S. (2010). Corpus linguistics: a general introduction. *Presented in the Workshop on Corpus Normalization, Linguistic Data Consortium for the Indian Languages (LDCIL), Central Institute of Indian Languages*, (pp. 01-25). Mysore. Retrieved from www.ciil.org/ldc-il/

Davletbaeva, D. N. (2010). *Lectures on English Lexicology.* Kazan: The Academy of Sciences.

Day, R. R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language, 14* (2). Retrieved 12 21, 2015, from http://nflrc.hawaii.edu/rfl/October2002/

de Kok, D., & Brouwer, H. (2011). *Natural Language Processing for the Working Programmer.* nlpwp.org.

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition, 19,* 195–222.

Delcloque, P. (2000). *An Illustrated History Of Computer-Assisted Language Learning: The History of CALL Web Exhibition*. Retrieved from History of CALL: http://www.ict4lt.org/en/History_of_CALL.pdf

Dictionary, O. E. (n.d.). *Where Does the Word Robot Come From?* Retrieved 12 21, 2015, from http://www.englishlanguagefaqs.com/2012/09/where-does-word-robot-come-from.html

Dieckmann, W. (n.d.). Deutsch politisch — politische Sprache im Gefüge des Deutschen. In J. Kilian (Ed.), *Sprache und Politik: Deutsch im Demokratischen Staat* (pp. 11-30). Mannheim: Dudenredaktion.

Diemer, S. (2011). Corpus linguistics with Google? *Proceedings of ISLE 2.* Boston.

Dillon, G. (2015, May). *Corpus Resources*. Retrieved from Resources for English language study: http://courses.washington.edu/englhtml/engl560/corplingresources.htm

Ellis, N. E. (1995). The Psychology of Foreign Language Vocabulary Acquisition: Implications for CALL. (D. Green, & P. Meara, Eds.) *International Journal of Computer Assisted Language Learning (CALL), 8*, 103-128.

Ellis, R. (1997). *Second Language Acquisition.* Oxford: Oxford University Press.

Elmes, D. (n.d.). Anki 2.0 User Manual. Retrieved 12 21, 2015, from http://ankisrs.net/docs/manual.html

Europa.eu. (n.d.). *About EuroVoc*. Retrieved 12 21, 2015, from EuroVoc: http://eurovoc.europa.eu/drupal/?q=abouteurovoc

European Commission. (2014). *The European Union Explained: How the EU works.* Luxembourg: Luxembourg: Publications Office of the European Union. doi:10.2775/11255

Fagan, D. S. (2005). Using Corpus Linguistics to Teach ESL. *Proceedings of the CATESOL State Conference.* San Diego: San Diego State University. Retrieved from http://64.8.104.26/Fagan.pdf

Fairclough, N. (2000). *New Labour, New Language?* London: Routledge.

Fantinuoli, C. (n.d.). *InterpretBank: Software for Interpreters*. Retrieved 12 21, 2015, from http://www.interpretbank.de/

Farrell, L., Osenga, T., & Hunter, M. (2013). *Comparing the Dolch and Fry High Frequency Word Lists.* Retrieved 12 21, 2015, from Readsters: http://www.readsters.com/wp-content/uploads/ComparingDolchAndFryLists.pdf

Felber, H., & Budin, G. (1989). *Terminologie in Theorie und Praxis.* Tübingen: Narr.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In T. P. Society, *Studies in Linguistic Analysis* (pp. 1-32). Oxford: Blackwell UK.

Flood, A. (2014, February, 20). Wikipedia 1,000-volume print edition planned. *The Guardian*. Retrieved 12 21, 2015, from http://www.theguardian.com/books/2014/feb/20/wikipedia-1000-volume-print-edition-crowdfunding

Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage.* Boston: Houghton Mifflin.

Frequency Lists. (2009). Retrieved from British Natinal Corpus: http://www.natcorp.ox.ac.uk/using/index.xml?ID=freq

Fry, E. (2004). *The vocabulary teacher's book of lists.* San Francisco: Jossey-Bass.

Fry, E., & Kress, J. E. (2006). *The Reading Teacher's Book Of Lists.* Jossey-Bass.

Gabrielatos, C. (2005, March). Corpora and Language Teaching: Just a fling or wedding bells? *The Electronic Journal for English as a Second Language, 8* (4). Retrieved from http://www.tesl-ej.org/ej32/a1.html

Gabrielatos, C. (2013). Corpus Linguistics 2 Keyword Analysis. *In Dubrovnik Fall School in Linguistic Methods.* Retrieved 12 21, 2015, from http://webcache.googleusercontent.com/search?q=cache:LgFyvCTataEJ:repository.edgehill.ac.uk/5932/1/NTNU.Dubrovnik.Keyness.pdf+&cd=1&hl=en&ct=clnk&gl=sk

Gacek, M. (2015). *Softwarelösungen für DolmetscherInnen (Master's diploma thesis).* Vienna: University of Vienna. Retrieved from http://othes.univie.ac.at/35667/

Gardner, D., & Davies, M. (2013). A New Academic Vocabulary List. *Applied Linguistics, 35* (3), 305-327. doi:10.1093/applin/amt015

Garside, R., Leech, G., & McEnery, T. (Eds.). (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longmann.

Gatto, M. (2014). *Web As Corpus: Theory and Practice.* London/New York: Bloomsbury.

Gile, D. (1999). Testing the Effort Models' tightrope hypothesis in simultaneous interpreting - A contribution. *Journal of Linguistics, 23* (23), 153-172.

Glosbe. (n.d.). *About Glosbe*. Cit. 21. 12 2015. Retrieved from http://blog.glosbe.com/post/85119355711/about-glosbe

Goulden, R., Nation, P., & Read, J. (1990). How Large Can a Receptive Vocabulary Be? *Applied Linguistics, 11* (4), 341-363.

Graham, A. (2008). The Effects of Homography on Computer-generated High Frequency Word Lists. *All Theses and Dissertations, Paper 1617.*

Granger, S., & Paquot, M. (2012). *Electronic Lexicography.* Oxford: OUP .

Gries, S. T. (2011). Methodological and interdisciplinary stance in Corpus Linguistics. In V. Viana, S. Zyngier, & G. Barnbrook, *Perspectives on Corpus Linguistics* (s. 81-97).

Gries, T. S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437. doi:10.1075/ijcl.13.4.02gri

Gries, T. S., & Berez, A. L. (to appear). Linguistic annotation in/for corpus linguistics. In N. Ide, & J. Pustejovsky (Ed.), *Handbook of Linguistic Annotation.* Berlin/New York: Springer.

Harvard Transdisciplinary Research in Energetics and Cancer Center. (2015, 12 21). Retrieved from Harvard T.H. Chan, School of Public Health: http://www.hsph.harvard.edu/trec/about-us/definitions/

He, Y. (2010). *A Study of L2 Vocabulary Learning Strategies (thesis).* Kristianstad: The School of Teaching Education, Kristianstad University.

Hirschman, L., & Sager, N. (1982). Automatic information formatting of a medical sublanguage. In R. Kittredge, & L. Lehrberger (Eds.), *Sublanguage; studies of language in restricted semantic domains* (pp. 27-80). Berlin/New York: Walter de Gruyter.

Hjřrland , B., & Nicolaisen , J. (Eds.). (2005). *The Epistemological Lifeboat: Epistemology and Philosophy of Science for Information Scientists.* Retrieved 12 21, 2015, from http://www.iva.dk/jni/lifeboat/

Höller, B. (2008). Die FremdgängerInnen. In M. Kaiser-Cooke (Ed.), *Das Entenprinzip. Translation aus neuen Perspektiven* (pp. 81-141). Frankfurt am Main: Peter Lang.

Holmes, J. (2013). *An Introduction to Sociolinguistics.* New York: Routledge.

Hubbard, P. (Ed.). (2009). *Hubbard, P. (Ed.) (2009). Computer Assisted Language Learning: Vol 1 (Critical Concepts in Linguistics). London: Routledge.* London: Routledge.

Hugh Bernard Fox, I. (1993). *A study of ESL teachers and the relationship between their attitudes about computer-assisted language learning and their expectations of minority students (PhD thesis).* Kingsville: Texas A&M University.

Hutschins, J. (2010). Machine translation: a concise history. (C. S. Wai, Ed.) *Journal of Translation Studies, 13* (1-2), 29-70.

Hyland, K., Meng Huat, C., & Handford, M. (2012). *Corpus Applications in Applied Linguistics.* London: Continuum.

Chamizo-Domínguez, P. (2006). False Friends. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (pp. 425-429). New York: Elsevier.

Chomsky, N. (1962). A transformational approach to syntax. In A. A. Hill (Ed.), *In Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English, May 9–12, 1958* (pp. 124–148). Austin: Univ. of Texas Press.

Chomsky, N. (2004). (Interviewed by Andor, Jozsef). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics, 1* (1), 93-111.

Chong, I., & Burger, A. (2011). Receptive Vocabulary. In S. Goldstein, & J. A. Naglieri (Eds.), *Encyclopedia of Child Behavior and Development* (pp. 1231-1231). New York: Springer US. doi:10.1007/978-0-387-79061-9_2359

Choueka, Y. (1988). Looking for needles in a haystack. *RIAO 88 Conference Proceedings,* (pp. 609–623). Cambridge.

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning Vocabulary Through Reading. *21*, pp. 767-787. Retrieved from http://www.jstor.org/stable/1163000

Johansson, S. (1995). Mens sana in corpore sano: On the Role of Corpora in Linguistic Research. *he European English Messenger, 4* (2), 19-25.

Kahneman, D. (1973). *Attention and Effort.* New Jersey: Prentice Hall.

Kaiser-Cooke, M. (2008). Warum ich sehe, was du siehst. In M. Kaiser-Cooke (Ed.), *Das Entenprinzip. Translation aus neuen Perspektiven* (pp. 13-17). Frankfurt am Main: Peter Lang.

Kezhen, L. (2015). The Use of Concordance Programs in English Lexical Teaching in High School. *Higher Education of Social Science, 8* (1), 60-65. doi:10.3968/6267

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics, 6* (1), 1-37.

Kilgarriff, A. (2006). *BNC database and word frequency lists*. Retrieved from http://www.kilgarriff.co.uk/bnc-readme.html

Kilgarriff, A. (2009). *Simple Maths for Keywords.* Liverpool: Lexical Computing Ltd.

Klein, J. (Ed.). (1989). *Politische Semantik: bedeutungsanalytische und sprachkritische Beiträge zur politischen Sprachverwendung.* Opladen: Westdeutscher Verlag.

Klein, J. (2014). *Grundlagen der Politolinguistik: Ausgewählte Aufsätze.* Berlin: Frank & Timme.

Klosa, A. (2007). Korpusgestützte Lexikographie: besser, schneller, umfangreicher. In W. Kallmeyer, & G. Zifonun (Eds.), *Sprachkorpora. Datenmengen und Erkenntnisfortschritt* (pp. 105-122). Berlin/New York: de Gruyter.

Knowles, G., & Zuraidah Mohd Don. (2004). The notion of a "lemma" : headwords roots and lexical sets. *International journal of corpus linguistics, 9*(1).

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit X.* Phuket.

Kolb, D. A., & Boyatzis, R. E. (2001). Experiential Learning Theory: Previous Research and New Directions. In R. J. Sternberg, & L.-f. Zhang (Eds.), *Perspectives on cognitive, learning, and Cognitive Styles.* L. Erlbaum.

Kormos, J. (2014). *Speech Production and Second Language Acquisition.* London/New York: Routledge.

Krieger, D. (2003). Corpus Linguistics: What It Is and How It Can Be to Teaching. *The Internet TESL Journal, 9* (3). Retrieved from http://iteslj.org/Articles/Krieger-Corpus.html

Krüger, E. (2013). Lob des bilateralen Dolmetschens. Eine didaktische Betrachtung. In K.-D. Baumann, & H. Kalverkämper, *Theorie und Praxis des Dolmetschens und Übersetzens in fachlichen Kontexten* (TRANSÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens ed., Vol. 63, pp. 285-295). Berlink: Frank & Timme.

Kubišová, J. (2015, June 17). Je vám bližšie selfie alebo „svojka“? Ako hovoríte, tak bude. *Aktuality.sk*. Retrieved 12 21, 2015, from http://www.aktuality.sk/clanok/277470/pouzivate-selfie-alebo-svojku-ako-hovorite-tak-bude/

Lamy , M. N., & Klarskov Mortensen, H. J. (2012). Using concordance programs in the Modern Foreign Languages classroom. (G. Davies, Ed.) *Module 2.4 in Davies G. (ed.) Information and Communications Technology for Language Teachers (ICT4LT)*. Retrieved 12 21, 2015, from http://www.ict4lt.org/en/en_mod2-4.htm

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik, *Directions in Corpus Linguistics* (s. 105-22). Berlin: de Gruyter.

Leech, G. (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing, 8* (4), 275-281.

Leech, G. (2005). Adding Linguistic Annotation. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow Books. Retrieved 12 21, 2015, from http://ahds.ac.uk/linguistic-corpora

Leech, G. (2014). The state of art in corpus linguistics. In K. Aijmer, & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 8-29). London: Longman.

Leech, G., Rayson, P., & Wilson, A. (2001). *Companion Website for: Word Frequencies in Written and Spoken English based on the British National Corpus.* Retrieved from http://ucrel.lancs.ac.uk/bncfreq/lists/2_2_spokenvwritten.txt

Leibert, R. E. (1991). The Dolch List Revisited - An Analysis of Pupil Responses Then and Now. *Reading Horizons, 31*(3).

Léon, J. (2005). Claimed and unclaimed sources of corpus linguistics. *Henry Sweet Society Bulletin* (44), 36–50.

Levelt, W. (1999). The neurocognition of language. In C. Brown, & P. Hagoort (Eds.). Oxford: Oxford Press.

Lexicology. (n.d.). In Oxford British & World Dictionary. Retrieved from http://www.oxforddictionaries.com/definition/english/lexicology

Liberman, M. (2008). *Comparing the Vocabularies of different languages*. Retrieved from Language Log: http://itre.cis.upenn.edu/~myl/languagelog/archives/005514.html

Lindermann, D. (2013). Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair. *5th International Conference on Corpus Linguistics* (pp. 105-122). Alicante: Elsevier.

Lingea. (n.d.). *Lexicon 5 Anglický slovník Platinum*. Cit. 21. 12 2015. Retrieved from http://www.lingea.sk/lexicon5-anglicky-platinum.html

Lingea. (n.d.). *O firme*. Cit. 21. 12 2015. Retrieved from http://www.lingea.sk/o-firme

Linguee. (n.d.). *About Lnguee*. Cit. 21. 12 2015. Retrieved from http://www.linguee.com/english-french/page/about.php

Lipka, L. (1992). *An outline of English lexicology.* Tübingen: Max Niemeyer.

Longman Communication 3000. (18. March 2003). *Longman English Dictionary - LDOCE*. Pearson ESL.

Makarová, V. (2004). *Tlmočenie : Hraničná oblasť medzi vedou, skúsenosťou a umením možného.* Bratislava: Stimul.

McEnery, T., & Gabrielatos, C. (2006). English corpus linguistics. In B. Aarts, & A. McMahon (Eds.), *The Handbook of English Linguistics* (pp. 33-71). Oxford: Blackwell.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Yukio, T. (2006). *Corpus-based Language Studies: An Advanced Resource Book.* London/New York: Routledge.

Meara, P., & Jones., G. (1990). *The Eurocentres Vocabulary Size Tests.* Zurich: Eurocentres.

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

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL - International Journal of Applied Linguistics*(107/108), 17-34.

Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics.* Oxford: OUP.

Morley, G. D. (2000). *Syntax in Functional Grammar*. London/New York: Continuum.

Mukherjee, J. (2005). *English ditransitive verbs: Aspects of theory, description and a usage-based model.* Amsterdam/New York: Rodopi.

Müller, F., & Waibel, B. (n.d.). *Corpus linguistics - an introduction*. Retrieved from uni.freiburg.de: http://www.anglistik.uni-freiburg.de/seminar/abteilungen/sprachwissenschaft/ ls_mair/corpus-linguistics.

Myers, R. (2015, April 24). Political jargon: a plain language guide. *GQ Magazine*. Retrieved 12 21, 2015, from http://www.gq-magazine.co.uk/comment/articles/2015-04/24/political-jargon-explained-guide

Nagy, W. E., & Stahl, S. A. (2007). *Teaching Word Meanings.* London/New York: Routledge.

Nation, P. (2001). How many high frequency words are there in English? In M. Gill, A. W. Johnson, L. M. Koski, R. D. Sell, & B. Wårvik (Eds.), *Language, Learning and Literature: Studies Presented to Håkan Ringbom English Department Publications 4* (pp. 167-181). Turku: Åbo Akademi University.

Nation, P. (2001b). *Learning Vocabulary in Another Language.* Cambridge University Press.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In M. McCarthy, & N. Schmitt (Eds.), *Vocabulary: Description, acquisition, pedagogy* (pp. 6-19). New York: Cambridge University Press.

Nesselhauf, N. (October 2005). *Corpus Linguistics: A Practical Introduction.* Retrieved from as.uni-heidelberg.de: http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf

Nicolescu, B. (1999). The Transdisciplinary Evolution of Learning. *Talk at the American Educational Research Association (AERA), Annual Meeting, Round-Table "" Overcoming the Underdevelopment of Learning : a Trandsdisciplinary View"", with the participation of Leon Lederman (Nobel Prize of Physics), Jan Visser, Ron Burne.* Montréal.

Tognini-Bonelli, T. (2001.). *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Paquot, M., & Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In M. Hundt, D. Schreier, & A. H. Jucker (Eds.), *Corpora: Pragmatics and Discourse* (pp. 243-265). Amsterdam: Rodopi.

Paroubek, P. (2008). Evaulating Part-of-Speech Tagging and Parsing. In L. Dybkjær, H. Hemsen, & W. Minker (Eds.), *Evaluation of Text and Speech Systems* (pp. 99-124). Dordrecht: Springer.

Pastor, G. C., & Seghiri, M. (2007). Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness. *Translation Journal, 11* (3).

Pearson, J. (1998). *Terms in context.* Amsterdam/New York: John Benjamins.

Perez-Sabater, C. (2011). Active Learning to improve long-term knowledge retention. *Proceedings of the XII Simposio Internacional de Comunicación Social*, (pp. 75-79). Santiago de Cuba.

Picht, H., & Draskau, J. (1985). *Terminology: An introduction.* Guildford: University of Surrey.

Pöchhacker, F. (1994). *Simultandolmetschen als komplexes Handeln.* Tübingen: Gunter Narr.

Pöchhacker, F. (1995). Slips and Shifts in Simultaneous Interpreting. In Y. Gambier, & J. Tommola, *Translation and Knowledge* (pp. 57-100). Turku: University of Turku.

Preller, A. G. (1967). Some Problems Involved in Compiling Word Frequency Lists. *The Modern Language Journal* (51), 399–402. doi:10.1111/j.1540-4781.1967.tb06724.x

Quinlan, P. T., & Dyson, B. J. (2008). *Cognitive Psychology.* Essex: Pearson Education.

R.L.G. (2014, January 30). The president's words of choice. *The Economist*. Retrieved 12 21, 2015, from http://www.economist.com/blogs/democracyinamerica/2014/01/politics-and-linguistics

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison (PhD thesis).* Lancaster: Computer Science Computer Department, Lancaster University.

Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: someanalyses ofthe conversational component of the Corpus. *International Journal of Corpus Linguistics, 2* (1), 132-152.

Robot. (n.d.). *In Online Etymology Dictionary*. (D. Harper, Ed.) Retrieved 12 21, 2015, from http://www.etymonline.com/index.php?term=robot

Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Exploration of MICUSP. *Journal of Writing Research*(2), 99-127.

Russo , M., Bendazzoli, C., Monti, C., Sandrelli, A., Baroni, M., Bernardini, S., . . . Mead, P. (2011). European Parliament Interpretation Corpus (EPIC).

Saldanha, G. (2009). Principles of Corpus Linguistics and Their Application to Translation Studies. *Tradumática, 7*.

Scott, M. (1997). PC analysis of key words - and key key words. *System, 25*(2), 233-245. doi:http://dx.doi.org/10.1016/S0346-251X(97)00011-0

Scott, M. (2016). *WordSmith Tools Manual.* Stroud: Lexical Analysis Software. Retrieved from http://lexically.net/downloads/version6/HTML/index.html?getting_started.htm

Serrander, U. (2011). *Bilingual Lexical Processing in Single Word Production: Swedish learners of Spanish and the effects of L2 immersion (PhD thesis).* Uppsala: Upsala Universitet.

Shlesinger, M. (2009). Towards a definition of Interpretese: An inermodal, corpus-based study. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogast (Eds.), *Efforts and Models in Interpreting and Translation Research: A tribute to Daniel Gile* (pp. 237-255). Amsterdam/Philadelphia: John Benjamins.

Schiemd, J. (1993). Qualitative and quantitative research approaches to English relative constructions. In C. Souter, & E. Atwell, *Corpus Based Computational Linguistics.* Amsterdam: Rodopi.

Sinclair, J. (Ed.). (1987). *Collins Cobuild English Language Dictionary.* London: Collins.

Sinclair, J. (1991). *Corpus concordance collocation.* Oxford: Oxford University Press.

Sinclair, J. (1995). Corpus typology – a framework for classification. In G. Melchers, & B. Warren (Eds.), *Studies in Anglistics* (pp. 17-33). Stockholm: Almqvist & Wiksell.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse.* London: Routledge.

Sinclair, J. (2005). Corpus and Text-Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxford: Oxford Books. Retrieved 12 21, 2015, from http://icar.univ-lyon2.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf

Singh, R. A. (1982). *An Introduction to Lexicography.* Central Institute of Indian Languages.

Skiba, R. (1997). Code Switching as a Countenance of Language Interference. *The Internet TESL Journal, 3* (10). Retrieved from http://iteslj.org/Articles/Skiba-CodeSwitching.html

Smith, M. (2015, February 26). Unparliamentary language: The rude words banned from the House of Commons. *Mirror*. Retrieved 12 21, 2015, from http://www.mirror.co.uk/news/uk-news/unparliamentary-language-rude-words-banned-5234983

Smith, S., Kilgarriff, A., & Sommers, S. (2008). Making better wordlists for ELT: Harvesting vocabulary lists from the web using WebBootCat. *Conference and Workshop on TEFL and Applied Linguistics.* Taoyuan.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*, 72-110.

Stavridou, I., & Ferreira, A. (2010). *Multi- Inter- and Trans-disciplinary research promoted by the European Cooperation in Science and Technology (COST): Lessons and experiments. [Research Report]*. Retrieved 12 21, 2015, from https://hal.inria.fr/inria-00512712v1/document

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., & Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation Journal, 48*(4), 679-707. doi:10.1007/s10579-014-9277-0

Stoykova, V., Simkova, M., Majchrakova, D., & Gajdosova, K. (2015). Detecting Time Expressions for Bulgarian and Slovak Language from Electronic Text Corpora. *Procedia - Social and Behavioral Sciences, The Proceedings of 5th World Conference on Learning, Teaching and Educational Leadership* (s. 257–260). Elsevier. doi:10.1016/j.sbspro.2015.04.178

Strauß, M. (2011). *Politolinguistik: Lexikalische und pragmatische Analyse politischer Sprache: Am Beispiel der parlamentarischen Rede Merkels zum Haushaltsgesetz 2010.* München: GRIN Verlag.

Stubbs, M. (1996). *Text and corpus analysis: computer-assisted studies of language.* Oxford: Blackwell.

Styles, E. A. (2006). *The Psychology of Attention.* Sussex/New York: Psychology Press.

Sudor, K. (2015, July 1). Sibyla Mislovičová: Slová neni a sranda spisovnými nebudú, selfie asi áno. *Denník N*. Retrieved 12 21, 2015, from https://dennikn.sk/174423/sibyla-mislovicova-slova-neni-a-sranda-spisovnymi-nebudu-selfie-asi-ano/

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME Journal*(32), 179-200.

Terminology. (n.d.). In Oxford British & World Dictionary. Retrieved from http://www.oxforddictionaries.com/definition/english/terminology?searchDictCode=all

Teubert, W. (2004). *Applied Corpus Linguistics: A Multidimensional Perspective.* (T. A. Upton, & ,. U. Connor, Eds.) Amsterdam/New York: Rodopi.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10* (1), 1-13.

Thomas, M. (2009). *Handbook of Research on Web 2.0 and Second Language Learning.* IGI Globa.

Thompson, E. (2011, December 4). The 106 things you can't say in Parliament. *iPolitics*. Retrieved 12 21, 2015, from http://ipolitics.ca/2011/12/14/the-106-things-you-cant-say-in-parliament/

Townson, M. (1992). *Mother-tongue and Fatherland: Language and Politics in German.* Manchester: Manchester University Press.

Transdisciplinary Inquiry: incorporating holistic principles. (2015, 12 21). Retrieved from Holistc Education Network: http://www.hent.org/transdisciplinary.htm

Tribble, C. (2000). Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals. In T. McEnery, & L. Burnard (Ed.), *Rethinking Language Peda-gogy from a Corpus Perspective* (s. 75-90). Frankfurt: Peter Lang.

Van de Walle, J., & Willems, K. (2006). Zipf, George Kingsley (1902–1950). In K. Brown (Ed.), *Encyclopedia of language & linguistics* (pp. 756-757). New York: Elsevier.

van Dijk, T. A. (2004). Text and Context of Parliamentary Debates. In P. Bayley (Ed.), *Cross-Cultural Perspectives on Parliamentart Discourse* (pp. 339-372). Amsterdam: Benjamins.

Vanopstal, K., Vander, R. V., Laureys, G., & Buysschaert, J. (2009). Vocabularies and Retrieval Tools in Biomedicine: Disentangling the Terminological Knot. *Journal of Medical Systems, 33*(4), 527-543. doi:10.1007/s10916-009-9389-z

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: phrase and topic discovery, with an application to information retrieval. *Proc. The 2007 edition of the IEEE International Conference on Data Mining series (ICDM'07)*, (pp. 697-702).

Waring, R., & Nation, P. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19). New York: Cambridge University Press.

Weber, K. (2011). *A genre analysis of the American presidential (Mag.phil thesis'*. Retrieved from http://othes.univie.ac.at/14956/1/2011-05-30_0307593.pdf

Weninger, C. (2010, October). The lexico-grammar of partnerships: Corpus patterns of 'facilitated agency'. *Text & Talk - An Interdisciplinary Journal of Language, Discourse & Communication Studies, 30*(5), 591-613. Retrieved from http://www.reference-global.com/toc/text/2010/30/5?ai=sb&ui=w6&af=H

Widdowson, H. G. (2002). On the limitations of linguistics applied. *Applied Linguistics, 21*(1), 3-5.

Wilkinson, M. (2011). WordSmith Tools: The best corpus analysis program for translators? *Translation Journal, 15*(3). Retrieved 12 21, 2015, from http://translationjournal.net/journal/57corpus.htm

Williams, R. (1976). *Keywords. A Vocabulary of Culture and Society.* London: Fontana.

Wright, S. E., & Budin, G. (Eds.). (1997). *Handbook of terminology management.* Amsterdam/Philadelphia: John Benjamins.

Xiao, Z. (2010). Corpus creation. In N. Indurkhya, & F. Damerau (Eds.), *The Handbook of Natural Language Processing* (2nd ed., pp. 147-165). London: CRC Press.

Young, M. S., & Stanton, N. A. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science, 3*(2), 178-194. doi:10.1080/14639220210123789

Zahurul, I. & Mehler, A. (2012). Customization of the Europarl Corpus for Translation Studies. (2012). *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (s. 2505-2510). Istanbul: European Language Resources Association.

Zufferey, S., & Cartoni, B. (2012). English and French causal connectives in contrast. *Languages in Contrast, 12* (2), 232–250.

Zufferey, S., & Cartoni, B. (2014). A multifactorial analysis of explicitation in translation. *Target, 26* (3), 361-384.

Žigo, P. (2001). Lexikálne prevzatia na slovensko-rakúskom pomedzí. *Slavica Slovaca, 36* (1), 3-12.

Appendix A

The hyperlink to corpus analysis files & glossary:

http://bit.ly/EP_corpusanalysis_glossaries_files

The *.zip archive contains:

Files

- BNC World wordlist used as a reference corpus
- Stop list created from BNC spoken subcorpora (3000 words)
- Generated wordlist from the Europarl corpus (vyrious formats and versions)
- Generated keyword list from the Europarl corpus (vyrious formats and versions)
- Corpus analysis working sheet (contains filter used in keywords selection & categorization

Glossaries

- Glossary 1 (*.xls, *.apkg, *.TMEX)
- Glossary 2 (*.docx)
- Glossary 3 (*.docx)
- Glossary 4 (*.docx)

The indication of version updates can be found in the text file *Readme.txt.* The files can be updated upon request.

Appendix B

Custom Word Lists from the EP Corpus Analysis
Content

1 Filter 1: abbreviations and verbs

1.1 Abbreviations

PPE, ACP, FR, PL, NOS, NGOS, ECU, PT, MRSÂ, NL, BSE, ECB, PSE, IGC, GMOS, ECOFIN, OSCE, HU, COD, NATO, EUROS, NGL, CNS, ES, EIB, SK, EMU, USD, GDP, VISAS, PHARE, CFP, PNR, ELDR, HÃ, PÃ, LÃ, EL, SIS, EDF, GSP, SME, NEO, FI, ILO, CÃ, GMO, EGF, FYROM, DG, LEZ, EPP, INTERREG, ECR, BG, ECO, FEIRA, INI, GRÃ, ICT, EEAS, EEC, ESDP, INTRA, IND, ALE, GBP, ITER, VAT, COREPER, ERDF, NGO, NSCH, UKIP, NAFO, EFSA, ECSC, EAGGF, EPLP, ALTENER, SSEL, IMO, EURES, EIT, EMAS, GM, TACS, DRC, EP, KLA, MFF, DAPHNE, ICCAT, CMO, UCITS, ICAO, FTA, GATS, MDGS, ETS, GNSS, KIVU, EASA, ESMA, MAGP, AARHUS, IUU, SYN, SADC, FIFG, UNMIK, ERTMS, EGNOS, UPE, ICTY, UNHCR, CSDP, CCCTB, TSE, EMCDDA, ECJ, SAPARD, ECHR, NEPAD, JPA, ECALL, KFOR, HIPC

1.2 Verbs

RATIFY, TACKLE, EMPHASISE, REGULATE, COMPLY, EXPORT, ALLOCATE, DEFEND, FACILITATE, SIMPLIFY, REGRET, FULFIL, GOVERN, CONFIRM, PARTICIPATE, ENTITLE, CLARIFY, UNDERLINE, PURSUE, REITERATE, ASSURE, SOLVE, RESTRUCTURE, ABSTAIN, OBLIGE, DECLARE, DESERVE, THREATEN, ACKNOWLEDGE, JUSTIFY, REVISE, ADAPT, INCORPORATE, DISCHARGE, RESTRICT, DELAY, REPEAT, UNDERMINE, COOPERATE, REINFORCE, COMPLICATE, CONSTITUTE, ADVOCATE, PROPOSES, DISAPPOINT, INVEST, SUCCEED, DISABLE, SUSPEND, CONSOLIDATE, ANNOUNCE, BURDEN, PUBLISH, PATENT, ATTACH, INSIST, VIOLATE, PERMIT, ASSOCIATE, ENLARGE, ARISE, ENVISAGE, EXCLUDE, MODIFY, ELIMINATE, ABOLISH, ENSHRINE, INVITE, STIPULATE, POSTPONE, UPHOLD, RECOMMEND, ADVERTISE, PROHIBIT, INITIATE, EMERGE, DEVOTE, OBTAIN, QUALIFY, REPRESENT, CONVINCED, COUNTERFEIT, APPAL, ACCOMPANY, OBSERVE, WEAKEN, WARN, AWAIT, REFUSE, PREVAIL, UNDERLIE, WITHDRAW, SATISFY, COMBINE, EMPLOY, PROCEED, CONSULT, POSE, DEPEND, ANALYSE, REAFFIRM, ENTAIL, OPT, TREAT, STRENGTHENED, INTERVENE, INTENDS, DEVASTATE, RESTORE, REMOVE, DUMP, RECYCLE, COMPARE, DISTRIBUTE, PRESERVE, STRIVE, CORRESPOND, FORMULATE, RENEW, DESTROY, COMMEND, EVALUATE, UNIFY, RESUME, ACCUSE, DEPRIVE, IGNORE, SURROUND, EXPLOIT, HARMONISE, CONFINE, DEEPEN, DEDICATE, DIVIDE, ENFORCE, BELONG, STATED, TORTURE, CONNECT, ACCEDE, ADHERE, INTENSIFY, EQUIP, APPRECIATE, DISTORT, DEEM, DEPLORE, CONSIDER, OVERCOME, DENOUNCE, REPLACE, APPLAUD, EXERT, RESOLVED, RESERVE, APPOINT, EMPHASIZE, EXACERBATE, CLOTHE, OVERWHELM, PAVE, ABANDON, ENDANGER, PUNISH, DESIGNATE, CLONE, CONCENTRATE, EXPOSE, ALLOCATED, ISOLATE, REACT, DISREGARD, CONFUSE, OCCUPY, RECONCILE, DENY, TRANSLATE, GENERATE, INSPIRE, UNDERPIN, CONSTITUTES, CONFRONT, DEPLOY, DISPLACE, ALLEGE, DISAGREE,

FINANCE, EXCEED, ERADICATE, WORSEN, PERPETRATE, OUTLINE, IMPRISON, ENTRUST, PERSECUTE, HINDER, MOBILISE, ASSIST, ACCELERATE, STIMULATE, ANTICIPATE, DISPUTE, MOTIVATE, SMUGGLE, CLAUSE, INFRINGE, CENSURE, CONTAMINATE, LEGISLATE, CONSIST, TIGHTEN, REASSURE, SPECIFY, ASTONISH, ORIGINATE, SEEKS, STIPULATES, THREATS, POPULATE, HARMONIZE, EMBARK, FORESEE, RETAIN, ROOT, SCOURGE, DELETE, VERIFY, FEED, SUSTAIN, HONOUR, UNDERESTIMATE, TRANSFER, ADJOURN, DEPLETE, CROP, AGREES, DETERIORATE, CONTAINS, INTERPRET, TOLERATE, MODERNISE, WITNESS, REMIT, INJURE, MANUFACTURE, HAMPER, PROMOTE, ABIDE, DETAIN, JEOPARDISE, SOVEREIGN, OWE, ARRIVE, ASSUME, DEVISE, AFFIRM, EXPAND, CONSUME, SLAUGHTER, CONTRADICT, BOOST, RESPECTS, REPORT, COMPEL, DIFFER, RECTIFY, BENCHMARK, ALIGN, LAG, ORGANIZE, PERTAIN, EDUCATE, CRITICISE, INCUR, RECAST, UNDERGO, REOPEN, DETAINEE, EXPIRE, RAIL, PRIORITISE, CONTRIBUTES, ENDORSES, BROADEN, CONVENE, CREATE, UNDERLINES, LIBERALISE, AMAZE, SEIZE, DISSEMINATE, CONTRAVENE, LAY, ACCOMPLISH, PLEDGE, DIGNIFY, DISAPPEAR, ENRICH, PRECEDE, POLISH, ADAPTING, INFECT, REFRAIN, UNDERMINES, PRESCRIBE, REFUND, GRATIFY, MONITORED, DERIVE, PROCLAIM, CODIFY, MODERNISING, DIVERSIFY, REBUILD, ATTAIN, MITIGATE, OPPRESS, DICTATE, COMPENSATE, DEFINING, ENACT, PROMPT, CONCLUDED, PROVEN, WIDEN, RATIFYING, RECONSIDER, COLLABORATION, STAKE, WITNESSING, FRUSTRATE, CEASE, EMPOWER, ALLEVIATE, RELAUNCH, PREMISE, OBSTRUCT, RESIDE, RELOCATE, INVOKE, PROSECUTE, SACRIFICE, CONFER, MISUSE, ORGANISE, LEND, REFER, COMPETE, CONCUR, EXPEL, CLASSIFY, FREEZE, SITUATE, STRIKE, ENCOMPASS, ENTAILS, VETO, STABILISE, STEM, DISCARDS, NOTIFY, PENALISE, FLOUT, PARALYSE, EXEMPT, VIOLATES, TRANSMIT, DISCARD, ADJUST, PRESIDE, IMPEDE, FALSIFY, ASPIRE, FLEE, DISTURB, ALLOT, LICENSE, AFFLICT, ENVISAGED, SWIFT, HORRIFY, COMPILE, EXECUTE, EVOLVE, ACQUIRE, CONFIRMS, FOCUSE, ERODE, RESTATE, REPEAL, IMPLY, PRESUPPOSE, TOY, SADDEN, CONVEY, ESCALATE, COMPRISE, RECOGNIZE, FOSTER, ECHO, EXCITE, REVISIT, IMPOVERISH, ALLUDE, DIMINISH, DECREASE, CONCEAL, WAIVE, CLEANSE, CONFISCATE, RESOUND, INSTIGATE, OVERSHADOW, SUM, TRANSFORM, TRAMPLE, DEPORT, IMPACTS, ABDUCT, LOCATE, ADOPTS, SHADOW, CIRCUMVENT, CIRCULATE, LAPSE, REINSTATE, SUPERVISE, ADVOCATES, REIMBURSE, FORGE, LULL, RAISES, CURTAIL, SQUANDER, HARM, ACCUSTOM, STEER, CLARIFIES, COLLABORATE, ASSERT, ADMINISTER, REAFFIRMS, DISBURSE, INCITE, RAVAGE, DIFFERENTIATE, NEGLECT, CULMINATE, HECKLE, RESOLUTE, HUMILIATE, DISCOURAGE, COMMUNICATE, ELAPSE, REUNITE, AUTHORIZE, RENDER, FIRE, ENTRENCH, EXAGGERATE, APOLOGISE, ILLUSTRATE, AUTHORISE, MISUNDERSTAND, POSES, DOWNGRADE, CROW, CERTIFY, FORBID, BROADCAST, CALCULATE, TEMPT, EXTRADITE, BRING, ENVISAGES, DECENTRALIZE, REDISTRIBUTE, COIN, TIRE, CONVICT,

ENSURE, REITERATES, APPROVES, INCLINE, GADDAFI, OMIT, OUGHT, ASTOUND, PREJUDGE, DOMINATE, RENOUNCE, PLUNDER, UNLEASH, URGES, LEVY, DELUDE, WITHHOLD

2 Filter 2: Klein's typology: list of categories except general interaction vocabulary

2.1 Group 1

COMMISSIONER, AMENDMENT, DIRECTIVE, RAPPORTEUR, PRESIDENCY, IMPLEMENT, COMMISSION, ENLARGEMENT, COLLEAGUE, COHESION, BUDGETARY, DRAFT, TABLED, UNION, SUMMIT, LEGISLATIVE, GUIDELINE, REPRESENTATIVE, CONVENTION, AMEND, WTO, MEP, HONOURABLE, SMES, EUROPEANS, INTERGOVERNMENTAL, SUBSIDIARITY, VISA, PLENARY, DELEGATION, PACT, PPE, ACP, RATIFY, TREATY, DECLARATION, MANDATE, PARAGRAPH, NOS, INTERINSTITUTIONAL, ABSTAIN, PROTOCOL, PACKAGE, CHARTER, EUROPOL, KYOTO, OMBUDSMAN, BILATERAL, CONSTITUTE, REFERENDUM, PROCEEDING, ACQUIS, ECB, RATIFICATION, PSE, CONSTITUTIONAL, IGC, MULTILATERAL, PRESIDENTS, DOSSIER, ALDE, OLAF, ECOFIN, OSCE, NATO, GUE, COUNCIL, COMITOLOGY, CFSP, INCENTIVE, FRONTEX, EIB, EMU, VERTS, EUROSTAT, EUROZONE, AMENDED, MEDA, VIS, PHARE, CONFEDERAL, UEN, EURATOM, CFP, ENFORCEMENT, PNR, TACIS, DIRECTORATE, EUROJUST, GALILEO, SIS, NATURA, GSP, SME, ILO, TROIKA, EGF, ERASMUS, EURODAC, FYROM, DG, NABUCCO, EPP, INTERREG, ECR, UCLAF, ASEM, ICT, EEAS, EEC, ESDP, SOCRATES, RECTIFY, DOSSIERS, SOLVIT, PLURALISM, EPAS, IND, ASEAN, QUAESTORS, MULTILATERALISM, INTERPARLIAMENTARY, BUREAU, OECD, ESF, EUROBAROMETER, TRANSNISTRIA, EUROGROUP, ITER, UNILATERALLY, COREPER, ERDF, EUROMED, GNI, EFD, UKIP, MUNDUS, ACTA, ENP, NAFO, EFSA, ECSC, EUROVIGNETTE, MONTERREY, EAGGF, TOBIN, EURES, EIT, EMAS, TACS, DRC, EP, KLA, MFF, DAPHNE, ICCAT, UCITS, ICAO, FTA, INTERGROUP, GATS, MDGS, HEMICYCLE, ETS, GNSS, EASA, ESMA, AARHUS, SADC, FIFG, UNMIK, ERTMS, ECOLABEL, EGNOS, UPE, ICTY, UNHCR, CSDP, CCCTB, TSE, EMCDDA, ECJ, SAPARD, ECHR, NEPAD, ECALL, HIPC

2.2 Group 2

FISHERY, EMISSION, AGRICULTURAL, FARMER, IMPOSE, MONETARY, SUBSTANCE, TOURISM, ECU, BIODIVERSITY, INFRINGEMENT, TOBACCO, AVIATION, REPERCUSSION, BIOFUELS, FISCAL, COUNTERFEITING, DEFICIT, PESTICIDE, ECOSYSTEM, GREENHOUSE, LIIKANEN, TUNA, ADDITIVE, BROADBAND, GMO, CONTAMINATE, MULTINATIONALS, PROPORTIONALITY, INTEROPERABILITY, WATERWAY, POLLUTANT, BIOMETRIC, ECO, MACROECONOMIC, LAMFALUSSY, FLEXICURITY, BLUEFIN, INTRA, DOPING, ADDITIVES, HAKE, DESERTIFICATION,

PHTHALATES, WYNN, BIOFUEL, GM, DIOXIN, CONSTITUENT, BIOGAS, CMO, PHARMACOVIGILANCE, BIOCIDAL, MICROFINANCE, CARCINOGENIC, IUU, NANOTECHNOLOGY, NANOMATERIALS, EXCISE

2.3 Group 4

EXTREMISM, LIBERALISE, FUNDAMENTALISM, MILITARISATION, IMPERIALIST, RENATIONALISATION, UNDEMOCRATIC, SEMITISM, SUPERSTATE, DEMONSTRATION, AUSTERITY, INTOLERANCE, TOTALITARIANISM, LEONARDO, RADICALISATION, ISLAMIST

3 Filter 3: Glossaries categorization: summary of words in glossaries except Glossary 1

3.1 Glossary 2

3.1.1 Initial categorization (groups 1-5)

3.1.1.1 Group 1

COMMISSIONER, COMMISSION, REFORM, COLLEAGUE, DEMOCRACY, SOLIDARITY, TERRORISM, EMISSION, DEMOCRAT, HUMANITARIAN, FARMER, GLOBALISATION, INITIATIVES, LIBERALISATION, DISCRIMINATION, SANCTION, PARTNERSHIP, INTERGOVERNMENTAL, SUBSIDIARITY, VISA, DELEGATION, TERRORIST, PACT, MANDATE, PARAGRAPH, TOURISM, INSTITUTIONAL, REFORMS, AUDITOR, PREVENTION, QUOTA, INTERINSTITUTIONAL, REGIME, DISCUSSIONS, PASSENGER, POLITICIAN, BILATERAL, SOCIALIST, ELEMENT, REFERENDUM, INVEST, STRATEGIC, BIODIVERSITY, ADMINISTRATIVE, TECHNOLOGIES, PATENT, MOBILITY, MULTILATERAL, TRANSATLANTIC, CONTINENT, HARMONISED, SOCIO, OPERATOR, MILLENNIUM, EFFICIENT, TOBACCO, AMBITION, TRANSITIONAL, LEGISLATOR, CORRUPTION, PETITION, INNOVATIVE, STABILISATION, TECHNOLOGICAL, XENOPHOBIA, DISTRIBUTE, ENVIRONMENTALLY, GREENS, TEXTILE, EVALUATION, TERRITORIAL, JUDICIAL, RECONSTRUCTION, EUROZONE, QUOTE, HARMONISE, TRANSPOSITION, DICTATORSHIP, COORDINATE, LEGITIMACY, DEFICIT, INDICATOR, DEMOCRATICALLY, AQUACULTURE, GENOCIDE, MEDICINAL, SENSITIVE, RACISM, PREVENTIVE, PESTICIDE, GLOBALISED, CLONE, TRAGIC, ECOSYSTEM, UNILATERAL, MACRO, ECOLOGICAL, INTER, INSPIRE, FORUM, SPHERE, TRAGEDY, FINANCE, DISCRIMINATORY, ALLIANCE, MOBILISE, ASSIST, STIMULATE, TUNA, BUREAUCRACY, PHARMACEUTICAL, BROADBAND, ECONOMICALLY, LOBBYIST, REALISTIC, CRITICISM, CENSURE, SITUATIONS, CONSULTATIONS, EMBARGO, MULTINATIONALS, GEOPOLITICAL, INTERCULTURAL, NEOLIBERAL, PROCEDURAL, UNIVERSAL, BIOMETRIC, PHENOMENON, CHECKS, HISTORIC, IMMUNITY, TOLERATE, SYNERGIES, DICTATOR, PANDEMIC, MACROECONOMIC, MONOPOLY, LIBERALS,

SYNERGY, SCANDALOUS, RECIPROCITY, DISCRIMINATE, TOTALITARIAN, TELECOMMUNICATIONS, GEOGRAPHICAL, LIBERALISED, CONDITIONALITY, RIGOROUS, ORIENT, INHUMANE, DECENTRALISED, CRITICISE, DOPING, PRIORITISE, BIOTECHNOLOGY, EXTREMISM, LIBERALISE, ETHIC, TREND, FUNDAMENTALISM, FEDERALIST, HECTARE, CRIMINALS, INTERNATIONALLY, LOBBY, IMPERIALIST, DIPLOMATIC, DEMONSTRATOR, COLLABORATION, ANTIBIOTIC, SEMITISM, INDUSTRIALISED, BIO, INCLUSIVE, COUNTERPRODUCTIVE, CATASTROPHE, VACCINE, QUALIFIED, PROSTITUTION, INGREDIENT, TSUNAMI, SYSTEMATICALLY, INADEQUATE, JUDICIARY, EUROCONTROL, UNDECLARED, INTERMODAL, INTOLERABLE, MULTINATIONAL, CIVILISATION, PHTHALATES, ELECTRONIC, LICENSE, PORNOGRAPHY, DEMONSTRATION, ORIENTED, DIPLOMACY, BIOFUEL, AUTOMOTIVE, INTERREGIONAL, CAMPAIGNS, DIOXIN, BIOGAS, SEMESTER, COSMETIC, CYBER, LATIN, BIOTECHNOLOGICAL, DIGITAL, MINISTERIAL, BIOCIDES, REPRODUCTIVE, INTOLERANCE, GEOSTRATEGIC, SOYA, REFORMED, EPIDEMIC, TOTALITARIANISM, BIOCIDAL, LOCATE, COORDINATOR, ORIENTATION, ORGANISM, MICROFINANCE, PRODUCTIVE, BARBARIC, INTEGRITY, CENTRALISED, PROBLEMATIC, CARCINOGENIC, PAEDOPHILIA, RADICALISATION, HUMANE, COMMUNICATE, MASSACRE, EUROSCEPTIC, NANOTECHNOLOGY, UNCONTROLLED, MULTIFUNCTIONAL, MEDIATION, THEMATIC, EUROPASS, AGRO, DESTABILISATION, NANOMATERIALS, BRUTAL, BLOCKADE, DECENTRALISATION, INTERMODALITY, BANKING

3.1.1.2 Group 2

PROCEDURE, INSTRUMENT, MONITOR, ASPECT, CRITERION, LEGISLATIVE, COORDINATION, INTEGRATE, UNACCEPTABLE, PLENARY, DECLARATION, HARMONISATION, ASSISTANCE, EXPORT, IMMIGRANT, TRANSPARENT, ALLOCATE, FACILITATE, LABELLING, PARTICIPATION, COMPLICATE, CANDIDATE, SUSPEND, PUBLISH, ASSOCIATE, REGULATORY, REVISION, PENALTY, MIGRATION, COUNCIL, CRISES, DONOR, AUTHORISATION, IGNORE, COHERENT, FISCAL, LEGITIMATE, MODERNISATION, SECTORAL, CONFEDERAL, AIRLINES, ISOLATE, ADEQUATE, PERSPECTIVES, QUALIFICATION, SPECIFY, COEXISTENCE, GLOBALLY, SECRETARIAT, TRANSIT, PROACTIVE, MODERNISE, COLLECTIVE, MULTILINGUALISM, RAPID, TRANSNATIONAL, DYNAMIC, SOVEREIGN, STANDARDISATION, COMPATIBLE, MILITARISATION, INTERPARLIAMENTARY, INFECT, PRIMARILY, COMPENSATE, VACCINATION, CERTIFICATION, CONSISTENCY, TRANSACTION, DELEGATE, ORGANISE, ABSURD, STABILISE, DEGRADE, DYNAMISM, BENEFICIAL, COMPLIMENT, EUROVIGNETTE, PRECEDENCE, EXPORTER, PRECEDENT, VACCINATE, CODIFICATION, ESCALATION, REGIONALISATION, UNPRECEDENTED, ADEQUATELY, AUTHORISE, PARAMETER, DOMINATE

3.1.1.3 Group 3

COOPERATION, PRESIDENCY, RESOLUTION, GUARANTEE, AGRICULTURAL, CONVENTION, AGRICULTURE, MONETARY, ADOPTION, SUBSTANCE, BUDGETS, DIMENSION, ADAPT, COOPERATE, CONSENSUS, CONSOLIDATE, CREDIBILITY, MODIFY, ELIMINATE, RATIFIED, INITIATE, DEADLINE, PROSPERITY, DEMOCRATISATION, RECYCLE, APPLICANT, COHERENCE, CLARIFICATION, PENSION, MOBILISATION, INVESTOR, ALLOCATED, ASPIRATION, PROTECTIONISM, IMPORTS, GENERATE, CONFRONT, EXTREMIST, INSPECTIONS, PERSECUTE, JOURNALIST, ACCELERATE, CONSOLIDATION, CONTAMINATE, COMPENSATION, UPDATE, ADMINISTRATIONS, PERSECUTION, IMPERATIVE, CONCESSION, INTERPRET, FACILITATION, MODIFIED, INTERVENTIONS, LIMITS, MISSIONS, EXPAND, HAZARDOUS, ALLOCATION, DIVERSIFY, EXEMPLARY, PROMPT, SENSIBLE, SCENARIO, MUNICIPALITY, RELOCATE, DESTRUCTION, CORRECTIONS, BENEFITING, EXECUTIONS, CORRECTION, SITUATE, PARALYSE, DIVERSIFICATION, PERSECUTED, SUPERSTATE, GRADUAL, VERIFICATION, SUPPLEMENT, ESCALATE, CONFISCATE, MARGINALISED, TRANSFORM, DEPORT, MARGINALISATION, INCONSISTENCY, PROHIBITION, LEGISLATIONS, CONVENTION'S, COLLABORATE, CULMINATE, RECIPIENT, RELOCATION, CALCULATE, COOPERATIVE, CONTAMINATION, SOLVENCY, DECENTRALIZE, REDISTRIBUTE, INCLINE, INSTALLATIONS

3.1.1.4 Group 4

APPLAUSE, CONSTITUTIONAL, GUARANTEED, OPERATIONAL, INTEROPERABILITY, PROLIFERATION, LIBERALIZATION, ABSTENTION, REDUCTIONS

3.1.1.5 Group 5

MARKETS, MULTIANNUAL, RESPONSIBILITIES, INVESTMENTS, VIOLATION, CONCENTRATE, PILLARS, ACTIVIST, REPRESSION, PRECEDE, REINTEGRATION, TRIBUTE, MODULATION, COGENERATION, BLACKLIST, MIERT, INTERLOCUTOR

3.1.2 Categories 2 and 3 (semantic shifts & false friends) further discussed with Ms. Rejšková

COOPERATION, PRESIDENCY, RESOLUTION, PROCEDURE, GUARANTEE, INSTRUMENT, MONITOR, ASPECT, CRITERION, LEGISLATIVE, AGRICULTURAL, CONVENTION, COORDINATION, INTEGRATE, AGRICULTURE, MONETARY, UNACCEPTABLE, ADOPTION, PLENARY, DECLARATION, HARMONISATION, ASSISTANCE, SUBSTANCE, BUDGETS, EXPORT, IMMIGRANT, TRANSPARENT, ALLOCATE, FACILITATE, DIMENSION, LABELLING, PARTICIPATION, ADAPT, COOPERATE, CONSENSUS, COMPLICATE, CANDIDATE, SUSPEND, CONSOLIDATE, CREDIBILITY, PUBLISH, ASSOCIATE, REGULATORY, MODIFY, ELIMINATE, RATIFIED, REVISION, PENALTY, INITIATE, MIGRATION, DEADLINE,

COUNCIL, CRISES, PROSPERITY, DONOR, AUTHORISATION, DEMOCRATISATION, RECYCLE, IGNORE, COHERENT, FISCAL, LEGITIMATE, APPLICANT, COHERENCE, MODERNISATION, CLARIFICATION, SECTORAL, CONFEDERAL, PENSION, MOBILISATION, INVESTOR, AIRLINES, ALLOCATED, ISOLATE, ASPIRATION, PROTECTIONISM, IMPORTS, ADEQUATE, GENERATE, CONFRONT, PERSPECTIVES, EXTREMIST, INSPECTIONS, PERSECUTE, JOURNALIST, ACCELERATE, QUALIFICATION, CONSOLIDATION, CONTAMINATE, COMPENSATION, SPECIFY, UPDATE, ADMINISTRATIONS, COEXISTENCE, GLOBALLY, PERSECUTION, IMPERATIVE, SECRETARIAT, TRANSIT, CONCESSION, PROACTIVE, INTERPRET, MODERNISE, COLLECTIVE, MULTILINGUALISM, RAPID, TRANSNATIONAL, DYNAMIC, SOVEREIGN, FACILITATION, MODIFIED, INTERVENTIONS, LIMITS, MISSIONS, STANDARDISATION, EXPAND, HAZARDOUS, COMPATIBLE, MILITARISATION, INTERPARLIAMENTARY, ALLOCATION, INFECT, PRIMARILY, DIVERSIFY, COMPENSATE, VACCINATION, EXEMPLARY, PROMPT, CERTIFICATION, CONSISTENCY, SENSIBLE, SCENARIO, TRANSACTION, MUNICIPALITY, RELOCATE, DESTRUCTION, CORRECTIONS, BENEFITING, DELEGATE, ORGANISE, EXECUTIONS, CORRECTION, SITUATE, ABSURD, STABILISE, DEGRADE, DYNAMISM, PARALYSE, BENEFICIAL, COMPLIMENT, EUROVIGNETTE, DIVERSIFICATION, PERSECUTED, SUPERSTATE, GRADUAL, VERIFICATION, PRECEDENCE, SUPPLEMENT, EXPORTER, PRECEDENT, ESCALATE, VACCINATE, CODIFICATION, CONFISCATE, MARGINALISED, TRANSFORM, DEPORT, MARGINALISATION, ESCALATION, REGIONALISATION, INCONSISTENCY, UNPRECEDENTED, PROHIBITION, LEGISLATIONS, ADEQUATELY, CONVENTION'S, COLLABORATE, CULMINATE, RECIPIENT, AUTHORISE, RELOCATION, CALCULATE, COOPERATIVE, CONTAMINATION, SOLVENCY, DECENTRALIZE, REDISTRIBUTE, PARAMETER, INCLINE, INSTALLATIONS, DOMINATE

3.1.3 Third categorization for the final list used for the glossary (category 2 + 3 from filter *internationalisms 2*).

PRESIDENCY, PROCEDURE, GUARANTEE, LEGISLATIVE, AGRICULTURAL, AGRICULTURE, UNACCEPTABLE, SUBSTANCE, BUDGETS, FACILITATE, LABELLING, PARTICIPATION, ASSOCIATE, PENALTY, DEADLINE, COUNCIL, EVALUATION, JUDICIAL, AQUACULTURE, AIRLINES, PERSECUTE, JOURNALIST, UPDATE, ADMINISTRATIONS, INTERPRET, RAPID, TRANSNATIONAL, FACILITATION, CONSUME, HAZARDOUS, INTERPARLIAMENTARY, PROMPT, COLLABORATION, SENSIBLE, MUNICIPALITY, BENEFITING, EXECUTIONS, BENEFICIAL, EUROVIGNETTE, GRADUAL, PRECEDENCE, SUPPLEMENT, VACCINATE, INCONSISTENCY, UNPRECEDENTED, LEGISLATIONS, COLLABORATE, RECIPIENT, CALCULATE, SOLVENCY, REDISTRIBUTE

3.2 Glossary 3

3.2.1 Initial selection

PROPOSAL, AMENDMENT, DIRECTIVE, ADOPT, PROPOSE, IMPLEMENT, INVOLVE, COHESION, DRAFT, COMPROMISE, APPROVE, COMBAT, TABLED, PREPARE, ACCORD, OBJECTIVES, GUIDELINE, AMEND, HIGHLIGHT, COMMIT, RESOLVE, SAFEGUARD, IMPOSE, BIND, ENDORSE, RATIFY, TREATY, SUBSIDY, ENTITLE, REITERATE, ENHANCE, ABSTAIN, ACKNOWLEDGE, INCORPORATE, UNDERMINE, REINFORCE, VIOLATIONS, PROCEEDING, ANNEX, EFFICIENCY, BURDEN, COMPLIANCE, ENVISAGE, WITHDRAW, PROCEED, ENTAIL, OPT, INTERVENE, PARLIAMENTARIAN, CONFINE, PRETEXT, DISPARITY, MOTIONS, DENOUNCE, APPOINT, EXACERBATE, INFRINGEMENTS, STOCKS, PREREQUISITE, MORATORIUM, DISPARITIES, ADVANCE, IMPRISON, IMPUNITY, IMPETUS, INFRINGE, STIPULATES, EMBARK, RETAIN, ADJOURN, DETERIORATE, MERIT, JEOPARDISE, OWE, AFFIRM, OBJECTION, CONTRADICT, FLEXICURITY, BENCHMARK, EXEMPTIONS, PERTAIN, INCUR, WAIVER, PLURALISM, CONVENE, CONSOLIDATING, UNDERLINES, CONTRAVENE, ENTIRETY, WARRANT, ATTAIN, PRECONDITIONS, ACCEDING, TIMEFRAME, IRRESPECTIVE, ALLEVIATE, RELAUNCH, PREMISE, CABOTAGE, AGGRAVATE, CONCUR, EXPEL, ASSENT, ENTAILS, LIVELIHOOD, ALLOT, COMMITTEE, HEARINGS, REPEAL, AUSTERITY, INTERFERE, RECOURSE, CUTBACK, PRECARIOUS, FOSTER, ALLUDE, DIMINISH, CONCEAL, WAIVE, CLEANSE, BORNE, UNEQUIVOCAL, DERIVATIVE, OVERSHADOW, BOTTLENECK, TRAMPLE, FLAGRANT, CIRCUMVENT, REINSTATE, CURTAIL, REDRESS, DISBURSE, RAVAGE, HECKLE, AUTHORIZE, COMMERCIALISATION, GADDAFI, ASTOUND, REFOULEMENT, RENOUNCE, HEARTEN, LEVY, BEFIT

3.2.2 Final selection

PROPOSAL, AMENDMENT, DIRECTIVE, ADOPT, PROPOSE, IMPLEMENT, INVOLVE, COHESION, DRAFT, COMPROMISE, APPROVE, COMBAT, TABLED, PREPARE, ACCORD, OBJECTIVES, GUIDELINE, AMEND, HIGHLIGHT, COMMIT, IMPOSE, BIND, ENDORSE, RATIFY, TREATY, SUBSIDY, ENTITLE, REITERATE, ENHANCE, ABSTAIN, ACKNOWLEDGE, INCORPORATE, UNDERMINE, REINFORCE, PROCEEDING, ANNEX, EFFICIENCY, BURDEN, COMPLIANCE, ENVISAGE, WITHDRAW, PROCEED, ENTAIL, OPT, INTERVENE, PARLIAMENTARIAN, CONFINE, PRETEXT, DISPARITY, MOTIONS, DENOUNCE, APPOINT, EXACERBATE, STOCKS, PREREQUISITE, MORATORIUM, ADVANCE, IMPRISON, IMPUNITY, IMPETUS, INFRINGE, EMBARK, RETAIN, ADJOURN, DETERIORATE, MERIT, JEOPARDISE, OWE, AFFIRM, OBJECTION, CONTRADICT, FLEXICURITY, BENCHMARK, PERTAIN, INCUR, WAIVER, PLURALISM, CONVENE, CONTRAVENE, ENTIRETY, WARRANT, ATTAIN, TIMEFRAME, IRRESPECTIVE, ALLEVIATE, RELAUNCH, PREMISE, CABOTAGE, AGGRAVATE, CONCUR, EXPEL, ASSENT,

LIVELIHOOD, ALLOT, HEARINGS, REPEAL, AUSTERITY, INTERFERE, RECOURSE

3.3 Glossary 4

ENLARGEMENT, TRANSPARENCY, ACCESSION, SUSTAINABLE, SUMMIT, WTO, MEP, ASYLUM, REFUGEE, SMES, SCHENGEN, PPE, ACP, NOS, ECU, EUROPOL, KYOTO, OMBUDSMAN, ACQUIS, ECB, PSE, IGC, ALDE, OLAF, OSCE, NATO, GUE, COMITOLOGY, CFSP, FRONTEX, EIB, EMU, VERTS, EUROSTAT, GDP, MEDA, VIS, PHARE, HAMAS, UEN, EURATOM, CFP, TRIALOGUE, PNR, TACIS, EUROJUST, GALILEO, MERCOSUR, SIS, EDF, LIIKANEN, NATURA, GSP, SME, ILO, GMO, TROIKA, EGF, ERASMUS, PROPORTIONALITY, EURODAC, FYROM, DG, NABUCCO, EPP, INTERREG, ECR, UCLAF, ASEM, LAMFALUSSY, ICT, EEAS, EEC, ESDP, SOCRATES, SOLVIT, EPAS, ASEAN, QUAESTORS, MULTILATERALISM, OECD, ESF, EUROBAROMETER, TRANSNISTRIA, RENATIONALISATION, EUROGROUP, ITER, COMPLEMENTARITY, COREPER, ERDF, EUROMED, GNI, EFD, UKIP, MUNDUS, ACTA, ENP, NAFO, EFSA, ECSC, MONTERREY, EAGGF, TOBIN, EURES, EIT, EMAS, MAGHREB, TACS, DRC, KLA, MFF, DAPHNE, ICCAT, CMO, UCITS, ICAO, PHARMACOVIGILANCE, FTA, HEMICYCLE, ETS, GNSS, SHARIA, EASA, ESMA, AARHUS, IUU, SADC, FIFG, ERTMS, ECOLABEL, EGNOS, ICTY, UNHCR, CSDP, CCCTB, ECJ, SAPARD, ECHR, NEPAD, EXCISE, ECALL, KFOR, HIPC

4 Filter 4: Word omissions

4.1 Group 1: mistakes

CONJURER, Â, THE, Â€, MRÂ, COMMISSIONÂ€™S, FR, PL, POLITIC, Ã, PARLIAMENTÂ€™S, PT, MRSÂ, UNIONÂ€™S, TRANS, EUÂ€™S, EUROPEÂ€™S, EURÂ, SV, PEOPLEÂ€™S, BÃ, DOHA, SOLANA, HU, RO, ES, SK, RÃ, MÃ, SÃ, Å, COUNCILÂ€™S, WOMENÂ€™S, Î, STATESÂ€™, TODAYÂ€™S, JÃ, HÃ, WALLSTRÃ, LOMÃ, PÃ, LÃ, EL, GÃ, Â€™, ARTICLEÂ, CANCÃ N, LIIKANEN, CITIZENSÂ€™, GUANTÃ, GONZÃ, FÃ, MALMSTRÃ, FI, JOSÃ, CÃ, COUNTRYÂ€™S, NDEZ, SCHÃ, GARCÃ, Â€˜THE, BERÃ, Â€˜NOÂ€™, LEZ, MARÃ, RAPPORTEURÂ€™S, WORLDÂ€™S, BG, MARTÃ, INI, GRÃ, CS, PIDLA, NOÂ, MADAMÂ, SCHRÃ, PLATH, DEN, SJÃ, COUNTRIESÂ€™, GROSSETÃ TE, AMENDMENTÂ, MAIJ, GRADIN, AÂ, COMMUNITYÂ€™S, SEPPÃ, WEGGEN, STEDT, ZU, HOUSEÂ€™S, SL, COMMISSIONERÂ, PEIJS, Â€˜YESÂ€™, PRESIDENCYÂ€™S, DELL'ALBA, YEARÂ€™S, VÃ, NAÃ, YEARSÂ€™, CEDERSCHIÃ, LALUMIÃ, LINKOHR, WORKERSÂ€™, VEZ, MIGUÃ, HELMS, NEN, IÂ, Â€˜EUROPEAN, GROUPÂ€™S, PLOOIJ, LVAREZ, MEMBERSÂ€™, ANDRÃ, KOVÃ, PRONK, CHILDRENÂ€™S, COMMITTEEÂ€™S, TOMORROWÂ€™S, SCHWAIGER, COMMISSIONERÂ€™S, BUSHILL, PARAGRAPHÂ, GUTIÃ, EUROPEANÂ, RREZ, KLAÃŸ, EUROPEÂ€™, D'Ã, PRESIDENTÂ, MACCORMICK, GOVERNMENTÂ€™S, GORSEL, PRÃ, RUSSIAÂ€™S, ONEÂ€™S, POTOÄ, ESTÃ, MARTEN, GAUZÃ, TTERING, ORBÃ, VALLÃ, CABROL, RFLER, QUADRAS, IMBENI, ROJO, CHÃ, MARSET, SMET, GHILARDOTTI, REPORTÂ€™S, HERNÃ, ROMANIAÂ€™S, LEGHOLD, BRINKHORST, RY, MCCARTIN, GRAÃ, SARYUSZ, ISRAELÂ€™S, SAKELLARIOU, EMPT, USDÂ, MATHIES, COMMUNIQUÃ, CRÃ, NEYRA, FERNÃ, CHINAÂ€™S, MOLLAR, DEÂ, FRANÃ, JOVÃ, RLING, MAGP, Â€˜WE, THEORIN, CONVENTIONÂ€™S, BNER, TELKÃ, WULF, IRANÂ€™S, DYBKJÃ, KÃ, JIMÃ, TAXPAYERSÂ€™, REZ, BOOGERD, POOS, Ä, PRETS, POMÃ, BONTEMPI, MEMBERÂ, TÃ, SUANZES, TEBORG, FIORI, JANUARYÂ, ZÃ, DESAMA, Â€˜A, PIMENTA, METTEN, PUBLICÂ€™S, MPER, NÃ, RIIS, HULTEN, CALAN, SZÃ, EFÄ, OVIÄ, SANDBÃ, EISMA, FRAHM

4.2 Group 2: general/irrelevant

MADAM, EUR, FIRSTLY, EU'S, SECONDLY, THIRDLY, LASTLY, NL, BSE, VICE, NONETHELESS, COD, EARMARK, EUROS, CNS, BANANA, FOURTHLY, USD, LAUNDER, NOON, TOPIC, TOPICAL, TODAY'S, NEO, FARMING, BEEF, KILOMETRE, BANANAS, GBP, TENS, NSCH, MANOEUVRE, ALTENER, SSEL, VEGETABLE, FACTO, EMPTIVE, THIRDS, GOOGLE, NORD, HOC, OPEL

4.3 Group 3: proper nouns

LISBON, EURO, TURKEY, RUSSIA, KOSOVO, MEDITERRANEAN, PALESTINIAN, BELARUS, BARROSO, BRUSSELS, ROMA, PRODI, BALKANS, UKRAINE, AMSTERDAM, GREECE, CYPRUS, ROMANIA, COPENHAGEN, STRASBOURG, BULGARIA, TURKISH, PORTUGAL, LUXEMBOURG, CZECH,

SWEDISH, AFGHANISTAN, BALTIC, PORTUGUESE, CHECHNYA, SERBIA, HUNGARIAN, IRAN, SWEDEN, DÃ, CROATIA, SLOVAKIA, POLAND, TAMPERE, ISRAELI, FINLAND, FISCHLER, AUSTRIAN, HERZEGOVINA, NETHERLANDS, AUSTRIA, GAZA, MACEDONIA, BELGIAN, GEORGIA, MOLDOVA, SPANISH, ALBANIA, MOROCCO, BURMA, FINNISH, BARCELONA, LAEKEN, TUNISIA, CYPRIOT, DANISH, CUBA, TURKEY'S, CONGO, BOSNIA, DARFUR, CAUCASUS, COTONOU, HELSINKI, LITHUANIA, ALBANIAN, TALIBAN, PUTIN, TIMOR, TIBET, SAKHAROV, SARKOZY, SLOVENIA, SPAIN, PALESTINIANS, BALKAN, STOCKHOLM, ZIMBABWE, ESTONIA, BELGIUM, THESSALONIKI, SUDAN, NAMO, SLOVENIAN, HUNGARY, LATVIA, GOTHENBURG, ITALY, MALTA, BELARUSIAN, SCHREYER, DUTCH, DENMARK, GENEVA, MAASTRICHT, DANUBE, LIBYA, RUSSIAN, TIBETAN, FEIRA, PAKISTAN, AFGHAN, ROMANIAN, ALGERIA, BERLUSCONI, OBAMA, TURKEYÂ€™S, YUGOSLAVIA, FLORENZ, CYPRIOTS, PALESTINE, HAITI, ROTHLEY, AZORES, LAMPEDUSA, LANNOYE, MEDINA, KURDISH, KYRGYZSTAN, BROEK, SLOVAK, CHECHEN, SAHARAN, SYRIA, SAHARA, ROTH, TUNISIAN, BURUNDI, HRKOP, HAGUE, GALICIA, KOREA, KALININGRAD, COLOMBIA, BASEL, CUBAN, ICELAND, BULGARIAN, DAYTON, ANNAN, BURMESE, CANARY, KOSOVO'S, ISRAELIS, RWANDA, MYANMAR, NCHEZ, AZERBAIJAN, GUANTANAMO, CHERNOBYL, BASQUE, MAURITANIA, BIOMASS, FUKUSHIMA, MOROCCAN, CHAD, CONGOLESE, LV, IRANIAN, ASIA, ENVISAGED, ANKARA, BEIJING, ALGERIAN, OSSETIA, GROSCH, LAMA, QAEDA, LEBANON, INDONESIA, UKRAINIAN, MINSK, CUNHA, BISSAU, UZBEKISTAN, RGENSEN, CAMPANIA, FABRA, GUINEA, CROATIA'S, LAHTI, PETERSBERG, HARMONIZED, SAHRAWI, ALBANIANS, SUDANESE, TURKMENISTAN, JOHANNESBURG, KOSOVARS, HLE, NIGERIA, GIL, ESSEN, TIBETANS, FABRE, KIVU, BRENNER, KOSOVAR, GUIN, BALI, TURK, KURDS, ROMANO, MADRID, SAMLAND, LANGENHAGEN, YUGOSLAV, SEVILLE, SRPSKA, SEATTLE, NER, MCKENNA, CARPEGNA, BELGRADE, RUSSIA'S, TAIWAN, STENMARCK, OSLO

4.4 Group 4: proper nouns (politicians)

VERHEUGEN, BROK, NGL, SANTER, SCHULZ, MONTI, BARNIER, FRATTINI, POETTERING, SWOBODA, BOLKESTEIN, PALACIO, ASHTON, POSSELT, READING, OOMEN, VITORINO, JARZEMBOWSKI, BENDIT, CRESPO, LEHNE, RUIJTEN, MILOSEVIC, WALDNER, LAMY, REHN, BARROT, FERRERO, BONINO, LUKASHENKO, DEMONSTRATE, WOGAU, MCCREEVY, RANDZIO, OOSTLANDER, LANGEN, BOURLANGES, BANGEMANN, COHN, KARAS, VERHOFSTADT, ROMPUY, JUNCKER, MUGABE, VELZEN, COELHO, BEHRENDT, MERKEL, HAUTALA, GRAEFE, FONTAINE, BLOKLAND, MULDER, STERCKX, BARINGDORF, DIMAS, FERBER, DAUL, THEATO, ALMUNIA, LEINEN, GEBHARDT, HAUG, TITLEY, IZQUIERDO, SALAFRANCA, NIELSON, DUPUIS, DUISENBERG, TAJANI, HAARDER, HATZIDAKIS, FATUZZO, PATTEN, PIRKER, NASSAUER, KOFI, KALLAS, LUDFORD, FRASSONI, AZNAR, MANDELSON, ELLES, BUITENWEG, SUU, TRAKATELLIS,

SCHROEDTER, NAPOLITANO, TRICHET, KYI, VIGO, PIEBALGS, DIAMANTOPOULOU, STAES, LANCKER, LAMASSOURE, ECHELON, BOWIS, ALAVANOS, CHIRAC, KATIFORIS, WURTZ, LULLING, BUZEK, BERTENS, BREYER, TURMES, KAUPPI, SOLBES, MORILLON, MAATEN, SAVARY, BELDER, LAGENDIJK, HOWITT, CAPPATO, KROES, GALEOTE, BUSQUIN, PAPAYANNAKIS, SACCONI, TANNOCK, CAUDRON, MAES, THORS, FRAGA, FAVA, ROURE, COLOM, TINDEMANS, WIJSENBEEK, ROBLES, BJERREGAARD, PAASILINNA, GARGANI, FIGUEIREDO, ORDEN, ZANA, DIAMANDOUROS, OETTINGER, MARGALLO, KARAMANOU, KRATSA, BONDE, OREJA, BORRELL, GARRIGA, AHERN, MORGANTINI, GOLLNISCH, MONNET, ZAPATERO, BANOTTI, SCHMID, DEPREZ, AUBRESPY, DIMITRAKOPOULOS, BOEL, MUSCARDINI, CASACA, DIRECTIVE'S, NAGORNO, DALAI, LANGE, ROCARD, BLAK, TSATSOS, LIESE, AELVOET, VARELA, DERMAN, FARAGE, SOUCHET, ORTEGA, ROTHE, KREISSL, LEYLA, SCHNELLHARDT, SILGUY, ORTUONDO, MOURA, NEVES, RIBEIRO, WEILER, VOGGENHUBER, BURG, POLLEDO, KEYSER, GOEBBELS, VALLELERSUNDI, ROSSA, WIERSMA, BUTTIGLIONE, MENRAD, VIRRANKOSKI, ANGELILLI, GUCHT, GAHRTON, JOUYET, PIECYK, STUBB, MARFIL, NIEBLER, CAMPOS, BLOTTNITZ, THYSSEN, GIANSILY, JEGGLE, ANDERSSON, RAPKAY, MYLLER, ELMAR, NAPOLETANO, KIRKHOPE, KLAMT, ISLER, PAPOUTSIS, PAULSEN, MARTENS, KABILA, AUNG, LEWANDOWSKI, VELD, TSAGAROPOULOU, HIERONYMI, GOEPEL, DALLI, HOPPENSTEDT, FLAUTRE, WIJKMAN, RADWAN, CORBEY, SPINELLI, KINDERMANN, KUHNE, STAUNER, PITTELLA, ERIKA, HECKE, GAHLER, REINFELDT, RASMUSSEN, CERCAS, CROWLEY, WORTMANN, KOUCHNER, LIENEMANN, QUECEDO, BULLMANN, CORNILLET, ANASTASSOPOULOS, CARNERO, SARNEZ, SHALIT, SCHIERHUBER, STOCKMANN, SADDAM, STIHLER, SCHWAB, FISCHER, MORATINOS, NEYTS, BERTHU, DEM, ARAFAT, NETANYAHU, CUSHNAHAN, SCHUMAN, ROMEVA, RUGOVA

4.5 Group 5: unrecognized lemmas

AMENDMENTS, GENTLEMAN, ADOPTED, NEGOTIATIONS, PARLIAMENT'S, FISHERIES, CONSUMERS, VOTING, RAPPORTEURS, COMBATING, EMISSIONS, EUROPE'S, MEPS, IMPLEMENTING, MONITORING, DIRECTIVES, PARLIAMENTS, DEMOCRATS, FISHING, RELATING, PROMOTING, ENTERPRISES, GUIDELINES, RESOLUTIONS, CONCLUSIONS, NGOS, APPROPRIATION, AMENDING, ADOPTING, GUARANTEES, FARMERS, ESTABLISHING, SANCTIONS, ACHIEVING, PRODUCERS, ENTERPRISE, FISHERMAN, DISASTERS, MINORITIES, RAPPORTEUR'S, PROPOSES, VIOLATIONS, COMMISSIONERS, REPRESENTATIVES, REFUGEES, SUBSIDIES, PEOPLES, COMPETENCES, DEBATING, GUARANTEEING, SUBSTANCES, OBLIGATIONS, PETITIONS, QUOTAS, MECHANISMS, OBSTACLES, FISHERMEN, GMOS, RECOMMENDATIONS, SOCIALISTS, INFRASTRUCTURES, DECLARATIONS, RESPECTING, NEGOTIATING, APPLIES, CONGRATULATING, AIRLINE, UNANIMITY, DEBATED,

PREVENTING, INTENDS, PRESIDENCIES, OPERATORS, SHORTCOMINGS, SAFEGUARDING, UTMOST, UNDERTAKINGS, PROVIDER, GRANTING, VISAS, DRAFTING, TAXPAYERS, PRESIDENCY'S, DESERVES, UNANIMOUS, TABLING, COMPROMISES, HIGHLIGHTED, VOTES, RESTRUCTURING, DELEGATIONS, IMPROVEMENTS, DEALT, DOCUMENTS, DEROGATIONS, DUMPING, PORTS, MAINSTREAMING, MIGRANTS, EMPHASIZE, FOODSTUFFS, INFRINGEMENTS, DEADLINES, ELDR, TACKLING, HARMONIZATION, SIMPLIFYING, PARLIAMENTARIANS, DISPARITIES, HARMONISING, RESOLVING, DEFICITS, LAUNDERING, RESERVATIONS, CONSTITUTES, TERRORISTS, REMARKS, NEIGHBOURHOOD, COMMISSIONER'S, VESSELS, BUREAUCRATIC, COMMUNAUTAIRE, COORDINATING, FINANCED, IMBALANCES, FACILITATING, PASSENGERS, COOPERATING, PREPARATIONS, PEOPLE'S, REFERENDUMS, INCENTIVES, DEMOCRACIES, FLEETS, GOVERNING, STIPULATES, RESTRICTIONS, HARMONIZE, DELAYS, REGIMES, OMBUDSMAN'S, MIGRATORY, DONORS, EXTENDING, APPROVING, PERPETRATORS, ENABLE, PENALTIES, RELOCATIONS, RECOGNISES, WEAPONS, CONTAINS, PARTNERSHIPS, CONVENTIONS, LOBBYISTS, INDICATORS, BURDENS, PREPARING, REFERENDA, REGULATING, ENHANCING, COMBATED, DISTORTIONS, BENEFICIARIES, STRENGTHENS, DELIBERATIONS, EMPHASISES, LANDMINES, ATTACHES, EXEMPTIONS, LIBERALISING, ORGANIZE, CONDOLENCES, PARAGRAPHS, DICTATORSHIPS, CONCLUDES, CONTRIBUTING, REFORMING, PESTICIDES, CONTRIBUTES, ENDORSES, CONSOLIDATING, UNDERLINES, MONOPOLIES, ECOSYSTEMS, SHIPYARDS, ELIMINATING, ADAPTING, INEQUALITIES, UNDERMINING, UNDERMINES, FRONTIERS, PARTICIPATING, MODERNISING, AMBITIONS, AUTHORISATIONS, TERRITORIES, CONDEMNING, PRECONDITIONS, ALE, ACCEDING, DEFENDING, LOOPHOLES, ENTREPRENEURS, RATIFYING, ENLARGEMENTS, FOCUSING, WITNESSING, ABOLISHING, FORA, IMPOSING, CLARIFYING, DIRECTORATES, REINFORCING, JEOPARDISING, ANNEXES, CONCLUDING, FULFILLING, GLOBALIZATION, CONDEMNS, REFER, LEGISLATORS, CLARIFICATIONS, DISMANTLING, MOBILISING, BENCHMARKING, ENTAILS, REGULATORS, RECITALS, DISCARDS, PROSPECTS, TRANSPOSING, VIOLATES, SUBMITTING, BENCHMARKS, TRAGEDIES, CIVILIANS, POSTPONEMENT, DICTATORS, EPLP, BOOSTING, COMPLIES, ECB'S, EUROBONDS, IMO, VACCINES, EMPHASISING, XENOPHOBIC, CONFIRMS, COMMITTEE, ENDEAVOURING, CORNERSTONES, POLLUTING, RESTRICTING, TRAFFICKERS, REJECTING, ERADICATING, EUROSCEPTICS, MASSACRES, CIVILISATIONS, REQUESTING, AUTHORISING, FORUMS, UPHOLDING, TRANSFERS, TARIFFS, ABSTAINING, PRIORITISING, MANDATES, ALZHEIMER'S, LIVELIHOODS, EXTREMISTS, TONNES, EXPECTATIONS, GROWERS, COMBATTING, IMMUNITIES, SMUGGLING, HOUSE'S, OBLIGES, EUROPOL, TEXTILES, ADOPTS, EMERGING, ADVOCATES, RECOMMENDS, COORDINATORS, CROPS, RAISES, MUNICIPALITIES, PREREQUISITES, CLARIFIES, VIOLATING, REAFFIRMS,

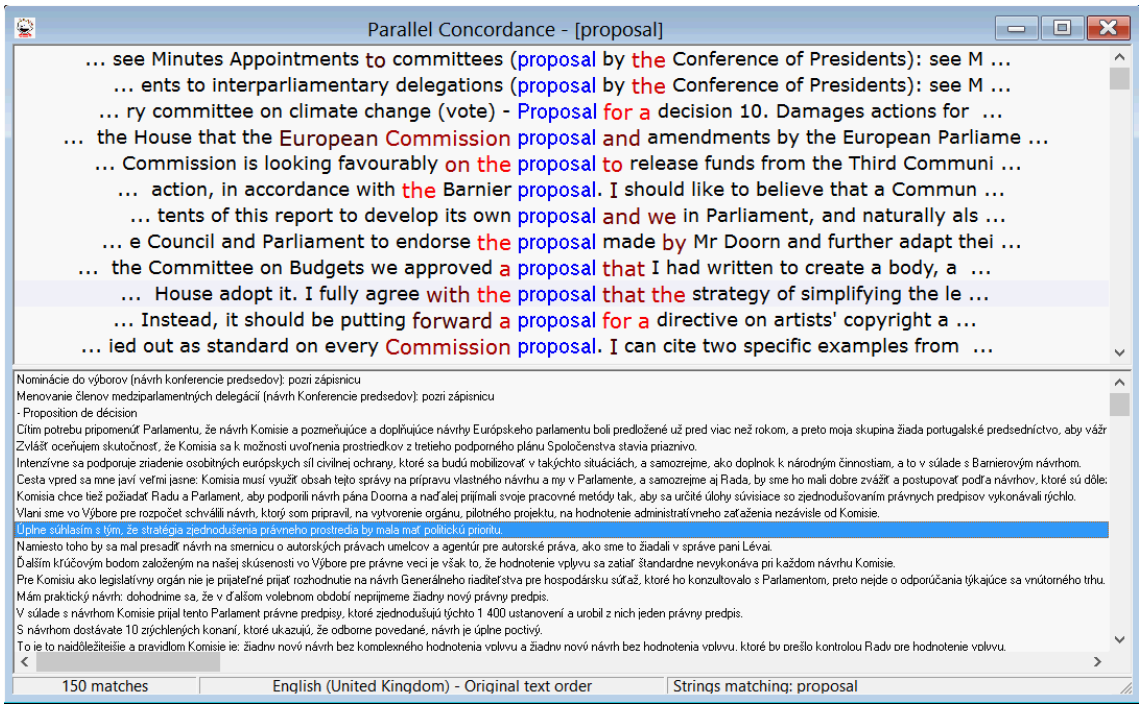
TRIALOGUES, OCCASIONS, EXPRESS, JOURNALISTS, BARROSO'S, OBSERVERS, PRESERVING, AUTHORIZE, ALLOCATING, FULFILS, EXPLANATIONS, POSES, LICENCES, JUSTIFIES, ENVISAGES, REACTORS, RECYCLING, SIGNATURES, REITERATES, SUGGESTIONS, DERIVATIVES, APPROVES, EXCEPTIONS, DESIGNATIONS, ENTITIES, NEGOTIATORS, REGRETS, EARMARKING, PENALISING, NANOTECHNOLOGIES, RELAUNCHING, TRANSACTIONS

Appendix C

a) The EP corpus screen capture (running text in Slovak with tags opened in AntConc through file fiew feature)

<P>
Áno, potrebujeme spoločnú prisťahovaleckú politiku, ale musia sa tým prediskutovať práva a povinnosti občanov a prisťahova
občanstva. Veľmi pekne vám ďakujem.
<SPEAKER ID="184" NAME="Ioannis Varvitsiotis . -" AFFILIATION="PPE-DE">
(EL) Vážený pán predseda, rád by som úradujúcemu predsedovi povedal: Tešil som sa na váš dnešný prejav s osobitným záujm
mesiacov vášho predsedníctva poriadne rozvírite hladinu.
<P>
Celý svet čelí jednej z najvážnejších hospodárskych kríz za posledné desaťročia a okrem toho, Európa čelí aj svojim vlastným kr
<P>
Je zrejmé, že dnes Európa 27 nemôže fungovať podľa pravidiel Európy 15. Európa prechádza inštitucionálnou krízou. Je takistc
Európy, čím nechcú, aby sa naplnila vízia našich predchodcov. Toto je kríza identity. Inštitucionálna kríza sa vyrieši, ale ako sa d
<P>
Obávam sa, že existuje len jedno riešenie: členské štáty, ktoré chcú politickú harmonizáciu by mali pokračovať, a členské štáty,
Skutočne sa obávam, že ak s riešením nepokročíme, Európa prestane byť v určitom bode prestane hráčom na medzinárodnom
<P>
Krízy môžu viesť k veľkým pokrokom, ale len vtedy, keď je dostatok odvahy. Verím, že máte aj víziu aj odvahu. Buďte odvážni!
<SPEAKER ID="185" NAME="Othmar Karas" AFFILIATION="PPE-DE">
(DE) Vážený pán predseda, vážený pán úradujúci predseda Rady, vaše základné stanoviská majú moju plnú podporu. Verím, že
predsedom, s potrebnou citlivosťou na obavy ľudí a odhodlanosťou viesť a zmierovať. Keď ste odpovedali na otázky, neuhýba
Neutekáte, ale čelíte problémom.
<P>
Obzvlášť by som veľmi rád zdôraznil vaše silné zanietenie pre parlamentnú demokraciu. Vaše odhodlanie je v tejto dobe veľmi
Rakúsku - sme boli svedkami zápasu medzi priamou demokraciou proti parlamentnej demokracii, a tým pádom svedkami zbav
parlamentných demokraciách. Postavme sa spolu za parlamentnú demokraciu a proti jej zbaveniu právnej spôsobilosti.
<P>
(potlesk)

b) KWIC in ParaConc of sentenced-aligned text (EN-SK)



c) Manual categorization of the generated keyword list in Microsoft Excel

	A	B	I	J	K	L	M	N	O
1	N	Key word	Lemmas	Set	Set 2	Set 3	Set 4	Filter fo	Filter fo
2	1	CONJURER					1		
3	2	Â					1		
4	3	THE					1		
5	4	PROPOSAL			3	3			
6	5	COMMISSIONER	commissioners		1	2		1	
7	6	AMENDMENT	amendments		1	3			
8	7	CITIZEN			3				
9	8	DIRECTIVE	directives		1	3			
10	9	Â€					1		
11	10	COOPERATION			3	2		3	1
12	11	ADOPT	adopted adopting ado	v	3	3			
13	12	RAPPORTEUR			1				
14	13	PROPOSE	proposes proposing	v	3	3			
15	14	UNITE	unites uniting	v	3				
16	15	AMENDMENTS					5		
17	16	GENTLEMEN			3				
18	17	MADAM					2		
19	18	PRESIDENCY	presidencies		1	2		3	3
20	19	RESOLUTION	resolutions		4	2		3	1
21	20	GENTLEMAN	gentlemen				5		
22	21	IMPLEMENT	implemented impleme	v	1	3		3	

Appendix E

Glossary 1 sample:

English (source)	German	Czech	Slovak	SRC GE	SRC CZ	SRC SVK
PROPOSAL	Vorschlag, r / Antrag, r / Angebot, s	návrh / nabídka	návrh / ponuka	b, f, h	c, h	1, h, c
COMMISSIONER	Kommissar, r / Beauftragte(r), r, e / Mitglied der Kommission, r / Kommissionsmitglied, r	komisář / člen komise / zastupitel	komisár / splnomocnenec / člen komisie	b, f	1, c	c, 1
AMENDMENT	(Ab)änderung, e / Novelle, e / Änderungsantrag, r	změna / dodatok / zlepšení	zmena / novela / dodatok	b, 1,	1, c	1, h, c
CITIZEN	(Staats)bürger, r / Staatsangehörige, r	občan / státní příslušník / obyvatel / civilista /	občan / obyvateľ / civilista	1, b, f, h	c, h	1, c, h
DIRECTIVE	Richtlinie, e / Vereinbarung, e / Rechtvorschrift, e / Anordnung, e	směrnice / pokyn / direktíva	smernica / direktíva / predpis	1, b	1, h	1, c, h
COOPERATION	Zusammenarbeit, e / Kooperation, e / Mitwirkung, e	spolupráce / součinnost / kooperace	spolupráca / kooperácia / súčinnosť	b, h	1, c, h	1, c, h
ADOPT	annehmen / verabschieden / genehmigen	prijmout	prijať	b, 1	b, 1	b, 1
RAPPORTEUR	Berichterstatter, r	zpravodaj / pozorovatel	spravodajca / pozorovateľ	1, h, c, f	1, h, c	1, h, c
PROPOSE	beantragen / Antrag stellen; machen	navrhnout / předložit	navrhnúť / predložiť	c	c	c
UNITE	(ver)einen / (ver)einigen / sich verbinden	spojit (se) / sjednotit se	spojiť (sa) / zjednotiť	b, f, h	c, h	c, h
GENTLEMEN	Herren	pánové	páni	b, f, h	c	c
PRESIDENCY	Vorsitz, r	předsednictví	predsedníctvo	1, b, f, h	c, h	c, h
RESOLUTION	Beschluss, r / Entschließung, e / Resolution, e	usnesení / dohoda	uznesenie / prehlásenie	1, b, f, h	1, c, h	1, c, h

IMPLEMENT	verwirklichen / in Kraft setzen / umsetzen	zrealizovat / uskutečnit / zavádět	zaviesť / uskutočniť / vykonať	1, b, f	c, h	1,
NEGOTIATION	Absprache, e / Verhandlung, e / Begebung, e	vyjednávání / smlouvání / jednání	rokovanie / vyjednávanie	b, f	h, c	c, h
PROCEDURE	Vorgehen(sweise), e / (Arbeits)verfahren, s / Ablauf, r	postup / soudné řízení, proces / jednání	postup / súdny proces, konanie / spôsob	1, b, f, h	c, h	c, h
INVOLVE	einbeziehen / umfassen / mit sich bringen	týkat se	týkať sa	b	c, h	c, h
(ON) BEHALF (OF)	im Auftrag von / im Namen von / im Interesse von / in Vertretung	jménem / ve jménu / za	v mene / za	b, f, h	c	c
CONSUMER	Vebraucher, r / Konsumet, r / Abnehmer, r	spotřebitel, spotřebitelský	spotrebiteľ, spotrebiteľský	1, b, f, h	c, h	c, h
VOTE	Stimme, e; (ab)stimmen / wählen / Stimme abgeben	hlasovat / "the vote" = celkový počet hlasů / volit	hlasovat / "the vote" = celkový počet hlasov / volit	b, f, h	c, h	c, h
COMMISSION	Kommisson, e / Ausschuss, r // Provision, e	komise / výbor / provize, odměna, poplatek	komisia / výbor / poplatok, provízia	1, b, f, h	1, c, h	1, c, h
GUARANTEE	Gewährleistung; gewährleisten / Bürgerschaft; bürgern / Sicherheit; sicherstellen	(za)ručit / zajistit / poskytovat záruku	(za)ručiť / zabezpečiť / poskytovať; dať záruku	b, f, h	c, h	c, h
RELATE	zusammenhängen / sich beziehen auf etw. / zu tun haben	týkat se / mít souvislost; spojitost	mať súvislosť; spojitosť / súvisieť / týkať sa	b, f, h	c, h	c, h
REFORM	reformieren; Reforme, e / (ver)bessern; (Ver)besserung, e / neu gestalten; Neugestaltung, e	reforma, zreformovat / napravit / zlepšit	reformovať, reformovanie / napraviť / zlepšiť	b, f, h	c, h	c, h
ENLARGEMENT	Erweiterung, e	rozšíření / zvětšení / expanze	rozšírenie / zväčšenie / expanzia	1, b, f, h	c, h	c, h
STRENGTHEN	verstärken; Verstärkung, e / bestärken	zpřísnit / posílit / zpevnit	sprísniť / podporiť / posilniť	1, b, f, h	c, h	c, h

Glossary 2 sample:

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***presidency***	KW no.:	18	
GE	preferred	Vorsitz, r			
CZ	preferred	předsednictví	prezidentství, -í s úřad a hodnost prezidenta prezidentství, -í s hodnost, funkce, úřad, činnost prezidenta; doba, po kt. úřad urč. prezidenta trvá Poměr v Google search: *předsednictví* vs. *prezidentství:* 352 000 vs. 1890 (u *prezidenství* se vážou výsledky převážně s konkrétní osobou prezidenta).		1
	admitted				96
	deprecated	prezidentství			
SK	preferred	predsedníctvo	1. volený vrcholný predstaviteľ republiky: p. Slovenskej republiky 2. často nenáležite predseda (významnej) inštitúcie: p. akadémie vied, p. Konfederácie odborových zväzov, p. akciovej spoločnosti (93). Predsedníctvo sa v SK používa viac ako *prezidentstvo* a hoci sa nepodarilo dopátrať, že *prezidentstvo* je jednoznačne nevhodné (v oficiálnych slovníkoch sa uvádza), dôrazne sa odporúča používať slovo *predsedníctvo* vzhľadom na jeho rozšírenie. Pomer výsledkov *predsedníctvo* vs. *prezidentstvo* v Google vyhľadávaní pomocou funkcie *site:.sk*: 157 000 vs. 51).		1
	admitted				93, c
	deprecated	prezidentstvo			

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***procedure***	KW no.:	23	
GE	preferred	(Arbeits)Verfahren, s			
CZ	preferred	postup	a) procedura: procedura, -y ž <l> složitější postup při něj. konání, jednání, řízení ap.: úřední, soudní p.; léčebná p. b) postup: účelný způsob práce n. jednání, metoda, (účelné) jednání.		95, 96
	admitted	procedura			95, 96
	deprecated	N/A			

			Upřednostňuje se spíše používat slovo postup než procedura. Poměr výsledků v Google search: 20 800 000 vs. 359 000	
SK	preferred	postup	a) procedúra: -y ž. ‹l› zložitejší postup pri nejakom konaní, správaní, riadení a pod.: úradná, súdna p.; liečebná p. úkon; výp. tech. postup realizácie algoritmu; uzavretá samostatná časť počítačového programu; b) spôsob konania: schváliť p. vyslanej delegácie (ustálený) spôsob práce, metóda: technologický p. Odporúča sa uprednostniť slovo *postup.* Pomer v Google search medzi slovami *postup* a *procedúra:* 2 170 000 vs. 97 400.	93
	admitted	procedúra, konanie		93, c
	deprecated	N/A		

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***guarantee***	KW no.:	29	
GE	preferred	Garantie, e, garantieren, gewährleisten			b
CZ	preferred	zaručit, zabezpečit	poskytnout záruku, garanci za něco, zaručovat (se), zaručit (se): g. něčí práva; g. kvalitu výrobku (95). Slovo *garantovat* je podle Google search běžně dostupné (454 000 výsledků).		95
	admitted	garantovat			95
	deprecated	N/A			
SK	preferred	zaručiť (sa), zabezpečiť	Podľa Google search je internacionalizmus *garantovať* bežne rozšírený (136 000 výsledkov)		93, c
	admitted	garantovať			93
	deprecated	N/A			

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***		KW no.:	79	
GE	preferred				
CZ	preferred	zákonodárný/ legislativní	obe slova je možné používat jako synonyma, ale je zapotřebí volit správný ekvivalent na základě kolokací se sousedícím slovem. a) *zákonodárný/é* (Google search: 96 300) vydávající zákony, týkající se práva vydávat zákony: z-é Národní shromáždění; z-á moc, funkce legislativní; z. program; práv. z. sbor		95
	admitted	N/A			
	deprecated	N/A			

			b) *legislativní*/é týkající se zákonodárství; zákonodárný; směřující k vydávání právních předpisů vůbec: l. činnost parlamentu; l. moc, právo; l. opatření, úprava; l. období; - l. pracovník (95). zákonodárný: 22 300 výsl.; -á: 33 300 výsl.; -é: 69 400 výsl. vs. legislativní: 561 000 výsl.; -á: 890 výsl.; -e 8050 výsl.	
SK	preferred	zákonodarný/ legislatívny	obe slová sa dajú v slovenčine používať ako synonymá, aj keď používanie slova *legislatívny* je prevažuje častosť výskytu slova *zákonodárny*. a) *zákonodarný*: súvisiaci s vydávaním zákonov: z-á činnosť, moc, iniciatíva. b): *legislatívny:* l. orgán, l-a činnosť, l-a úprava, l-e opatrenie, l-e predpisy; l-e oddelenie ministerstva v ktorom sa pripravujú návrhy zákonov (93). Google search pre legislatívny: 108 000 výsl.; -a 53 000 výsl.; -e: 261 000 výsl. vs. zákonodarný: 14 600 výsl.; -a 71 výsl.; -e 80 výsl..	
	admitted	N/A		
	deprecated	N/A		

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***agricultural***	KW no.:	83	
GE	preferred	landwirtschaftlich			
CZ	preferred	zemědelský	slovo *agrikulturní* se ve slovníku neologizmů nenachází (95) a jako přídavné jméno se sice uvádí (96), no jeho výskyt je značně omezený (978 výsledků v Google search)		c
	admitted				
	deprecated	agrikulturní			
SK	preferred	poľnohospodársky	Slovo *agrikultúrny* sa síce nachádza v slovníkoch JULS (93) a iných cudzojazyčných slovníkoch, ale podľa výsledkov Google search sa v praxi takmer nepoužíva. Čo sa týka predpony „agro“, tu používanie definuje jazykovedný časopis Kultúra slova: agroprodukt gen. -u, muž. (gr. + lat.) poľnohospodársky produkt, poľnohospodársky výrobok: Hlavnými požiadavkami rezortu je podporiť vývoz vlastných agroproduktov. (TLAČ)		
	admitted				
	deprecated	agrikultúrny			

			Slovo agroprodukt zložené z časti agro-, ktorá vyjadruje vzťah k poľnohospodárstvu (z gréckeho slova agros, resp. latinského ager — pole; lat. koreň agri- sa uplatnil v staršom slove agrikultúra — poľnohospodárstvo), a z podstatného mena produkt (z latinského slova productum — výrobok, výtvor) môžeme preložiť ako poľnohospodársky produkt, príp. poľnohospodársky výrobok. Medzi neologizmami s časťou agro- nájdeme také, ktoré odrážajú novú realitu, napr. agropodnikateľ, agroprivatizátor, agrobiznis, agroturistika (vidiecka turistika, ktorá môže byť spojená aj s prácou na poli), ďalej významovo príbuzné slová agrorezort, agrosektor, agrooblasť či príležitostne používané slová ako agrovýdavky, agrodotácie, agropolitika, agroobchod (98).

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***agriculture***	KW no.:	132	
GE	preferred	Landwirtschaft, e			
CZ	preferred	zemědělství	Slovo *agrikultura* se také oficiálně uvádí ve slovnících (96), ale jeho výskyt v Google search je velice omezený (3000 výsledků) a nenachází se také ani mezi neologizmy (95). Proto se toto slovo doporučuje používat jenom v kontextu odborné terminologie.		c
	admitted	N/A			
	deprecated	agrikultura			
SK	preferred	poľnohospodárstvo	Podobne ako v predchádzajúcom prípade je slovo *agrikultúra* oficiálne uvádzané v slovníkoch, ale výsledky Google search nenasvedčujú jeho častému používaniu (80 výsl.). Odporúča sa vyhýbať sa jeho používaniu, príp. ho obmedziť len na odborný kontext.		c
	admitted	N/A			
	deprecated	agrikultúra			
	Categorization	Translation			Source

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***unacceptable***	KW no.:	151	
GE	preferred	inakzeptabel			
CZ	preferred	nepřijatelný	Slovo *neakceptovatelný* se nenachází v kodifikovaných slovnících (96), ale byl nalezen v slovniku neologizmů (95). Přesto se téměř vůbec nepoužívá a proto se doporučuje se mu spíše vyhnout. Google search: neakceptovatelný: 100 výsl.; -á: 80 výsl.; -é: 30 100 výsl. vs. nepřijatelný: 8090 výsl.; -á: 41 výsl.; 521 000 výsl.		c
	admitted	N/A			
	deprecated	neakceptovatelný			

			K zajímavému zjištění ale přijdeme při porovnání antonyma *akceptovatelný*, kde se situace mění. Slovo akceptovatelný by mělo být z hlediska používání upřednostňováno před slovem *přijatelný*. Google search pro akceptovatelný: 375 000 výsl.; -á: 374 000 výsl.; -é: 373 000 výsl. vs. Přijatelný: 30 200 výsl.; -á 9200 výsl.; -é: 15 900 výsl.	
SK	preferred	neprijateľný	Slovo *neakceptovateľný* sa uvádza v kodifikovaných slovníkoch a je pomerne rozšírené, avšak z dôvodu častejšieho používania slova *neprijateľný* sa odporúča uprednostňovať toto slovensky znejúcejšie synonymum. Google search pre neakceptovateľný: 6900 výsl.; -á 4300 výsl.; -é 20 300 výsl. vs. neprijateľný 21 500 výsl.; -á 52 200 výsl.; -é 174 000 výsl. Naopak, antonymum *akceptovateľný* sa používa tak zriedka, že sa odporúča ho vôbec nepoužívať Google search pre akceptovateľný: 86 výsl.; -á: 2950 výsl.; -é: 5780 výsl. vs. prijateľný: 48 200 výsl.; -á: 71 000 výsl.; -é: 158 000 výsl.	1, c
	admitted	neakceptovateľný		93
	deprecated	N/A		

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***substance***	KW no.:	214	
GE	preferred	Substanz, e			
CZ	preferred	látka, hmota	slovo *substance* je spíše abstraktní pojem, proto by se překládá anglické slovo *substance* spše jako *látka* nebo *hmota* substance, -e ž. (z lat.) knuž. 1. základ věcí a jevů; podstata 1: (duše) je poslední s-í jedině jistou (Vrchl.); hmota houbovité s. (Havlasa) 2. hmotný základ; podstata 4, majetek, jmění: žít ze své s.; s. přípravku byly stále měněny látky; ekon. (v kapit.) s. podniku jeho trvalá zařízení; kapitálová s. oběžný i stálý kapitál podniku (96).		
	admitted	N/A			
	not recommended	substance			

SK	preferred	látka, hmota	Anglické slovo *substance* je častejšie významovo bližšie slovenskému ekvivalentu *látka* alebo *hmota*, než slovu *substancia*. a) látka: forma hmoty majúca isté fyzikálne al. chem. zloženie, materiál: stavebná, pohonná l., plastické l-y; b) substancia -ie ž. filoz. al. kniž. podstata (význ. 1), základ (vecí, javov): duchovná, hmotná s.; rozpustné, rastlinné s-ie látky; (93).	1, c
	admitted	N/A		
	not recommended	substancia		

	Categorization	*Translation*	*Notes*		*Source*
EN	***source word***	***budget***	KW no.:	215	
GE	preferred	Budget, s			
CZ	preferred	rozpočet	V kodifikovaných slovnících se slovo *budžet* nenachází, ale je možné najít několik výsledků přes Google search (1490 výsledků), kde se dá rovněž dopátrat i k některým definicím. V kodifikovaném jazyce toto slovo ale nejestvuje.		1, c
	admitted	N/A			
	deprecated				
SK	preferred	rozpočet	Slovo *budget* sa v kodifikovaných slovníkoch nenachádza. Dá sa v nich nájsť slovenský neologizmus *budžet* ale v Google search má len do 100 výsledkov*:* budžet -u m. ‹a < f› ekon. zastar. rozpočet (najmä štátny); (93). Tento anglicizmus je ale vnímaný skôr negatívne a patrí do neformálneho hovorového jazyka, ako informuje jazykovedný časopis *Slovenská reč*. Ide o tzv. xenos alebo profesionalizmus (99).		1, c
	admitted	N/A			
	deprecated	budget			

Glossary 3

TERM	proposal, (n)		Sources:			
GE: Vorschlag, r	**CZ**: návrh		**SK**: návrh		b c	a
syn EN	Suggestion, plan, programme, scheme					
syn DE	Antrag, r; Angebot, s		**Worldlist rank**	116	b, f	
syn CZ	nabídka		**Keyword rank**	4	c, h	
syn SK	ponuka		**Occurance**	77 463	c, h	
Definition	a) A plan or suggestion, especially a formal or written one, put forward for consideration or discussion by others b) The action of proposing a plan or suggestion:					5
Collocations EN	Concerning/related to proposal, formulate p.					c
Collocations GE	Was den Vorschlag anbelangt, V. formulieren,					c
Collocations CZ	Týkající se návrhu, sformulovat/vypracovan N.,					c
Collocations SK	Týkajúci sa návrhu, vypracovať/zformmulovať návrh					c
Further English Collocations	ADJ. concrete | detailed | controversial | compromise | peace, reform, research, etc. QUANT. package, set The government outlined a new set of proposals on human rights. VERB + PROPOSAL formulate | outline | bring forward, make, put forward, submit | accept, back, support, welcome I welcome the proposal to reduce taxes for the poorly paid. | block, oppose, reject, vote against | push through The government could face defeat if it tries to push through the controversial proposals. | drop, withdraw | consider, discuss PREP. ~ concerning/relating to proposals concerning the use of land | ~ for The Ministry submitted a proposal for lower speed limits on motorways. > Special page at MEETING					e
N:grams	/N/ for, to, from, concerning		council's, legislative, commission's /N/			
Context	There is strong support for the establishment of a special European civil protection force to be mobilised in such situations, and always, of course, as a supplement to national action, in accordance with the Barnier proposal. I fully agree with the proposal that the strategy of simplifying the legal environment should have political priority. This is the be-all and end-all and the Commission's rule is this: no new proposal without a comprehensive impact assessment, and no new proposal without an impact assessment which has been scrutinised by the Impact Assessment Board.					

TERM	amendment, (n)		Sources:				
GE: Novelle, e	CZ: změna	SK: zmena			b	c	c
syn EN	Addition, adjustment, revision, adaptation, change, improvement				d		
syn DE	(Ab)Änderung, e; Änderungsantrag, r; Zusatz, r	**Worldlist rank**	219		b, 1, f		
syn CZ	dodatek pozměnovací návrh; novela zákona	**Keyword rank**	6		c, h		
syn SK	dodatok; zlepšenie; doplnok	**Occurance**	31 190		c, h, 1		
Definition	a) A minor change in a document. b) A change or addition to a legal or statutory document:						5
Collocations EN	Put forward amendments, move an A., adopt A.						c
Collocations GE	Änderungsanträge vorlegen, Änderungsvertrag einbringen, Änderungen vornehmen, annehmen, übernehmen						c
Collocations CZ	Předložiť změny/pozměňovací návrh, detto, přijmout Z						c
Collocations SK	Predložiť pozmeňujúce a doplňujúce návrhy, detto, prijať Z a D.						c
Further English Collocations	ADJ. important, major, significant A major amendment was introduced into the legislation. | minor, slight, small | draft, proposed | detailed | constitutional VERB + AMENDMENT introduce, make | draft The committee does not adequately consult others when drafting amendments. | move, propose, put forward, suggest, table He moved an amendment limiting capital punishment to certain very serious crimes. | withdraw She withdrew her amendment and left the meeting. | accept, adopt, approve, carry, pass, ratify, support, vote for Parliament accepted the amendment and the bill was passed. On a free vote, the amendment was carried by 292 votes to 246. | oppose, reject | be subject to The programme is subject to amendment. PREP. without ~ The new clause was accepted without amendment. | ~ to an amendment to the Clean Water Act							e
N:grams							
Context	We tend to agonise for months over this or that amendment, but often put no effort into finding out whether the legislation has had its desired effect. I would like to ask everyone here to support the amendments that have been discussed and introduced, especially Amendment 44, where, at the request of the Council, in addition to vehicles we introduce the concepts of wagons and inland waterway vessels, in order to avoid any possible misunderstanding, and I would ask you to vote in favour of this. Then Amendment 16 could be omitted as unnecessary, or we could vote against it.							

TERM	directive, (n)		Sources:			
GE: Anordnung, e	CZ: směrnice		SK: smernica		b 1 1	
syn EN	Order, instruction, imperative				d	
syn DE	Richtlinie, e; Vereinbarung, e; Rechtvorschrift, e		**Worldlist rank**	159	1	
syn CZ	pokyn; direktíva; nařízení		**Keyword rank**	8	1, h	
syn SK	vyhlásenie, direktíva, nariadenie, predpis		**Occurance**	45 738	1, h	
Definition	legal act which is binding, as to the result to be achieved, upon each Member State to which it is addressed, but leaves to the national authorities the choice of form and methods					1
Collocations EN	Issue a directive, adopt a directive					c
Collocations GE	Eine Richtlinie erlassen/verföffentlichen, R. verabschieden					c
Collocations CZ	Vydát směrnici, přijmout směrnici					c
Collocations SK	Vydať smernicu, prijať smernicu					c
Further English Collocations	ADJ. clear Don't start anything without a clear directive from management. | general | important | draft, proposed | EU, European (Commission/Union), government, ministerial | policy, political | banking, environmental, etc. VERB + DIRECTIVE issue The EU issued a new drinking water directive. | adopt, agree (on/upon), approve, sign | comply with, implement All companies must comply with the new directive. | block, oppose DIRECTIVE + VERB come into force A new EU directive on maternity leave will come into force next month. | require sth The directive requires member states to designate sites of special scientific interest. PREP. in accordance with a/the ~ They acted in accordance with the latest directive from Brussels. | in a/the ~ The proposals are contained in a European directive on wild birds. | under a/the ~ Private health services will be allowed under the directive. | ~ from a directive from the European Commission | ~ on a directive on data protection PHRASES the provisions/terms of a directive						e
N:grams						
Context	We already realised that when we were discussing the directive on the internal market in services. Member of the Commission. - Mr President, the oral question tabled by Mr Gargani, on behalf of the Committee on Legal Affairs, gives me the opportunity to provide you with an update on where the Commission stands regarding the 14th Company Law Directive and the European Private Company (EPC). When can we expect the long-awaited disability directive that will put real legislative weight behind equality for people with disabilities?					

TERM	adopt, (v)		Sources:				
GE: annehmen	CZ: přijmout		SK: prijať		b	c	c
syn EN	Accept, maintain, approve, take up, embrace, decide on				d		
syn DE	verabschieden, genehmigen, einwilligen, erlassen		**Worldlist rank**	684	b, 1		
syn CZ	přejmout; osvojit si (názory metody); zaujmout postoj		**Keyword rank**	11	c		
syn SK	osvojiť si; prebrať; prevziať (názory, stretégie a pod.); zaujať postoj		**Occurance**	39 117	c		
Definition	a) Take up or start to use or follow (an idea, method, or course of action) b) Take on or assume (an attitude or position) c) Formally approve or accept (a report or suggestion) d) (Of a local authority) accept responsibility for the maintenance of (a road).				5		
Collocations EN	Adopt a law, adopt a plan, adopt an attitude				c		
Collocations GE	Ein Gesetz verabschiedenÜannehmen, einen Plan übernehmen, eine Haltung einnehmen				c		
Collocations CZ	Přijmout zákon, prijmout plán, prijmout/zaujat postoj				c		
Collocations SK	Prijať zákon, prijať plán, prijať/zaujať postoj				c		
Further English Collocations	take and use sth ADV. formally, officially, immediately, urgently, subsequently The policy has not yet been formally adopted. VERB + ADOPT tend to | decide to | be forced to PREP. towards the policies employers adopt towards the labour force				E, a		
N:grams							
Context	Finally, the regulations that we wish to adopt, or rather to combine from many different sources, thereby reducing the number, are based on the recommendations drawn up by the UN for the transport of dangerous goods by road, rail and inland waterways, which together account for over 110 billion tonnes/km per annum within the European Union. As a result, the Commission urges the Member States to adopt a more ambitious approach when future proposals are submitted to them in this area. Member of the Commission. - (CS) Madam President, ladies and gentlemen, Council Directive 1999/70/EC concerning the framework agreement on fixed-term work requires the Member States to adopt such measures that would prevent abuse arising from the use of successive fixed-term employment contracts.						

Glossary 4 sample

TERM	**EU enlargement**, (n)	**Occurance:** 14 522	**KW rank:** 35		
GE:,	*CZ:*	*SK:*			
Definition: European Union (EU) enlargement describes the process of admitting new member states to join the EU. To qualify for EU membership, a state must meet the 'Copenhagen criteria (2).		**Notes:** Corpus analysis suggests that EU enlargement is a term used to discuss the process from the point of view of the EU. For referring to individual countries, the word *EU accession* is preferred.			

TERM	**EU accession**, (n)	**Occurance:** 11 202	**KW rank:** 48		
GE:,	*CZ:*	*SK:*	1	2	3
Definition: The process of joining the EU (4)		**Notes:** Current candidate countries are Albania, Macedonia, Montenegro, Serbia, and Turkey (3).			

TERM	**sustainable**, (a.)	**Occurance:** 11 481	**KW rank:** 50		
			1	2	3
Definition: more likely in the EU context: Conserving an ecological balance by avoiding depletion of natural resources (more relevant in political context. Also: b) Able to be maintained at a certain rate or level, c) Able to be upheld or defended (5).		**Context:** sustainable development, mobility, energy, economic, fishing, transport, tourism, agriculture.			

TERM	**summit**, (n)	**Occurance:** 13 252	**KW rank:** 75		
GE: Gesetzgebung, e	*CZ:*	*SK: legislatíva*	1	2	3
Definition: (political context) an important formal meeting between leaders of governments from two or more countries (6)		**Notes:** Summit conference, meeting,			

TERM	**WTO**, (abbbr)	**Occurance:** 6136	**KW rank:** 95		
GE:,	*CZ:*	*SK:*	1	2	3
Definition: The World Trade Organization (WTO) deals with the global rules of trade between nations. Its main function is to ensure that trade flows as smoothly, predictably and freely as possible (7).		**Notes:** The WTO provides a forum for negotiating agreements aimed at reducing obstacles to international trade and ensuring a level playing field for all, thus contributing to economic growth and development. The WTO also provides a legal and institutional framework for the implementation and monitoring of these agreements, as well as for settling disputes arising from their interpretation and application. (7).			

TERM	**MEP**, (abbr)	**Occurance:** 5833	**KW rank:** 105		
GE: Mitglieder des Europäischen Parlaments	**CZ**: Poslanci Evropského parlamentu	**SK**: Poslanci Európskeho parlamentu	16	16	16
Definition: Members of (European) Parliament - Persons elected to the European Parliament		**Notes:** The European Parliament is made up of 751 Members elected in the 28 Member States of the enlarged European Union. Since 1979 MEPs have been elected by direct universal suffrage for a five-year period. Each country decides on the form its election will take, but must guarantee equality of the sexes and a secret ballot. EU elections are by proportional representation (9)).			

TERM	**refugee**, (n)	**Occurance:** 6585	**KW rank:** 136		
GE:,	*CZ:*	*SK:*			
Definition: refugees are persons that receive protection under the Geneva Convention in a Member State (10). Compare: Asylum seekers are persons submitting a claim for refugee status, pending a legal procedure (10).		**Notes:** EU Member States take decisions on how asylum seekers and refugees are treated (10). the Common European Asylum System is a first step towards establishing harmonised standards for asylum seekers and refugees in the EU (10).			

TERM	**Schengen**, (n)	**Occurance:** 4415	**KW rank:** 147		
GE:,	*CZ:*	*SK:*			
Definition: **Shengen agreement:** An intergovernmental agreement on the relaxation of border controls between participatingEuropean countries, first signed in Schengen, Luxembourg, in June 1985 A revised version of the agreement wasincorporated into the European Union in 1999 and widened to include non-EU members of a similar Nordic union (5).		**Notes:** The Schengen Area and the European Union are two completely different zones that shall not be misinterpreted (12). The border-free Schengen Area guarantees free movement to more than 400 million EU citizens, as well as to many non-EU nationals, businessmen, tourists or other persons legally present on the EU territory (13).			

TERM	**Europol** (n)	**Occurance:** 2301	**KW rank:** 315		
GE:,	*CZ:*	*SK:*			
Definition: Europol is the **European Union's law enforcement agency** whose main goal is to help achieve a**safer Europe** for the benefit of all EU citizens. It assists the European Union's Member States in their fight against serious international crime and terrorism (14).		**Notes: Unique services** •support centre for law enforcement operations •hub for criminal information and organisations •centre for law enforcement expertise •one of the largest concentrations of analytical capability in the EU •produces regular assessments and reports •high-security, 24/7 operational centre •central platform for law enforcement experts from the European Union countries (14).			

TERM	**EU enlargement**, (n)	**Occurance:** 14 522	**KW rank:** 35		
GE: Erweiterung	***CZ:*** rozšíření	***SK:*** rozširovanie	1	1	1
Definition: European Union (EU) enlargement describes the process of admitting new member states to join the EU. To qualify for EU membership, a state must meet the 'Copenhagen criteria (2).		**Notes:** Corpus analysis suggests that EU enlargement is a term used to discuss the process from the point of view of the EU. For referring to individual countries, the word *EU accession* is preferred.			

TERM	**(EU) accession**, (n)	**Occurance:** 11 202	**KW rank:** 48		
GE: Beitritt, r	***CZ:*** přistoupení	***SK:*** pristúpenie, prístup	1	2	3
Definition: a) The process of joining the EU (4); b) the act whereby a state accepts the offer or the opportunity to become a party to a treaty already negotiated and signed by other states; it has the same legal effect as ratification and usually occurs after the treaty has entered into force		**Notes:** Current candidate countries are Albania, Macedonia, Montenegro, Serbia, and Turkey (3).			

TERM	**sustainable**, (a.)	**Occurance:** 11 481	**KW rank:** 50		
DE: nachhaltig	***CZ:*** udržitelnost	***SK:*** udržateľný	1	1	1
Definition: more likely in the EU context: Conserving an ecological balance by avoiding depletion of natural resources (more relevant in political context. Also: b) Able to be maintained at a certain rate or level, c) Able to be upheld or defended (5).		**Context:** sustainable development, mobility, energy, economic, fishing, transport, tourism, agriculture.			

TERM	**summit**, (n)	**Occurance:** 13 252	**KW rank:** 75		
GE: Gripfel(treffen)	***CZ:*** summit, schůzka na nejvyšší úrovni	***SK:*** samit, stretnutie na najvyššej úrovni	1	c	c
Definition: (political context) an important formal meeting between leaders of governments from two or more countries (6)		**Notes:** Summit conference, meeting,			

TERM	**WTO**, (abbbr)	**Occurance:** 6136	**KW rank:** 95		
GE: Welthandelsorganisation	***CZ:*** (WTO) světová obchotní organizace	***SK:*** (WTO) svetová obchotná organizácia	1	1	1
Definition: The World Trade Organization (WTO) deals with the global rules of trade between nations. Its main function is to ensure that trade flows as smoothly, predictably and freely as possible (7).		**Notes:** The WTO provides a forum for negotiating agreements aimed at reducing obstacles to international trade and ensuring a level playing field for all, thus contributing to economic growth and development. The WTO also provides a legal and institutional framework for the implementation and monitoring of these agreements, as well as for settling disputes arising from their interpretation and application. (7).			

TERM	**MEP**, (abbr)	**Occurance:** 5833	**KW rank:** 105		
GE: Mitglieder des Europäischen Parlaments	**CZ**: Poslanci Evropského parlamentu	**SK**: Poslanci Európskeho parlamentu	16	16	16
Definition: Members of (European) Parliament - Persons elected to the European Parliament		**Notes:** The European Parliament is made up of 751 Members elected in the 28 Member States of the enlarged European Union. Since 1979 MEPs have been elected by direct universal suffrage for a five-year period. Each country decides on the form its election will take, but must guarantee equality of the sexes and a secret ballot. EU elections are by proportional representation (9)).			

TERM	**refugee**, (n)	**Occurance:** 6585	**KW rank:** 136		
GE: Flüchtling, r	***CZ:*** uprchlík	***SK:*** utečenec			
Definition: refugees are persons that receive protection under the Geneva Convention in a Member State (10). Compare: Asylum seekers are persons submitting a claim for refugee status, pending a legal procedure (10).		**Notes:** EU Member States take decisions on how asylum seekers and refugees are treated (10). the Common European Asylum System is a first step towards establishing harmonised standards for asylum seekers and refugees in the EU (10).			

TERM	**Schengen**, (n)	**Occurance:** 4415	**KW rank:** 147		
GE: Schengen	***CZ:*** Schengen	***SK:*** Schengen	1	1	1
Definition: **Schengen agreement:** An intergovernmental agreement on the relaxation of border controls between participatingEuropean countries, first signed in Schengen, Luxembourg, in June 1985 A revised version of the agreement wasincorporated into the European Union in 1999 and widened to include non-EU members of a similar Nordic union (5).		**Notes:** The Schengen Area and the European Union are two completely different zones that shall not be misinterpreted (12). The border-free Schengen Area guarantees free movement to more than 400 million EU citizens, as well as to many non-EU nationals, businessmen, tourists or other persons legally present on the EU territory (13).			

TERM	**Europol** (n)	**Occurance:** 2301	**KW rank:** 315		
GE: (Europol) Europäisches Polizeiamt	***CZ:*** (Europol) Evropský policejní úřad	***SK:*** (Europol) Európsky policajný úrad	1	1	1
Definition: Europol or European Police Office is the **European Union's law enforcement agency** whose main goal is to help achieve a**safer Europe** for the benefit of all EU citizens. It assists the European Union's Member States in their fight against serious international crime and terrorism (14).		**Notes: Unique services** •support centre for law enforcement operations •hub for criminal information and organisations •centre for law enforcement expertise •one of the largest concentrations of analytical capability in the EU •produces regular assessments and reports •high-security, 24/7 operational centre •central platform for law enforcement experts from the European Union countries (14).			

TERM	**Kyoto**, (n)	**Occurance:** 2405	**KW rank:**		328
GE: Kyoto-Protokoll, Protokoll von Kyoto zum Rahmenübereinkommen der Vereinten Nationen über Klimaänderungen	***CZ:*** Kjótský protokol, Kjótský protokol k Rámcové úmluvě Organizace spojených národů o změně klimatu	***SK:*** Kjótsky protokol, Kjótsky protokol k Rámcovému dohovoru Organizácie Spojených národov o zmene klímy	1	1	1
Definition: The Kyoto Protocol is an international agreement linked to the United Nations Framework Convention on Climate Change, which **commits** its Parties by setting internationally binding emission reduction targets (15). The Kyoto Protocol was adopted in Kyoto, Japan, on 11 December 1997In Doha, Qatar, on 8 December 2012, the "Doha Amendment to the Kyoto Protocol" was adopted (15).		**Notes:** Recognizing that developed countries are principally responsible for the current high levels of GHG emissions in the atmosphere as a result of more than 150 years of industrial activity, the Protocol places a heavier burden on developed nations under the principle of "common but differentiated responsibilities." (15).			

TERM	**(European) ombudsman**, (n)	**Occurance:** 2889	**KW rank:**		337
GE: (Europäischer) Bürgerbeauftragter	*CZ:* (evropský) veřejný ochránce práv	*SK: európsky* ombudsman	1	1	1
Definition: The term 'ombudsman' originates from Sweden meaning an official appointed to receive and investigate complaints (Marriam-Webster, 2011). Introduced in the Treaty of Maastricht (1stNovember 1993) under the 'People of Europe' initiative, the European Ombudsman is the institution for citizens of the European Union (EU) to report to if they believe any EU institution has been guilty of maladministration. The European Parliament will then ask the Ombudsman to review the complaint and, if the complaint is founded, carry out an investigation.		**Notes:** The general protocol for investigating complaints is that the European Ombudsman approach the institution or body of the EU which a complaint has been filed against, the Ombudsman looks for a peaceful solution to put right the case of maladministration. If this fails the European Ombudsman makes a special report to the European Parliament who will then investigate further. In the view of the citizens the European Ombudsman has helped improve the quality of EU administration, receiving over one thousand complaints annually, however the organisations main problem has been a lack of resources and actually defining the term maladministration.			

TERM	**Acquis communitaire**, (n)	**Occurance:** 1910	**KW rank:**		373
GE: Besitzstand der Gemeinschaft	**CZ**: acquis Společenství	**SK**: acquis Spoločenstva	1	1	1
Definition: . The *acquis* is the body of common rights and obligations that is binding on all the EU member states (42). Also: Community acquis (1)		Notes: It is constantly evolving and comprises: - The content, principles and political objectives of the Treaties; - Legislation adopted pursuant to the Treaties and the case law of the Court of Justice; - Declarations and resolutions adopted by the Union;			

	- Instruments under the Common Foreign and Security Policy; - International agreements concluded by the Union and those entered into by the member states among themselves within the sphere of the Union's activities (42).

TERM	**ECB**, (abbr)	**Occurance:** 1875	**KW rank:** 375		
GE: EZB, Europäische Zentralbank	**CZ**: ECB, Evropská centrální banka	**SK**: ECB, Európska centrálna banka	1	1	1
Definition: The European Central Bank (ECB) is the central bank for Europe's single currency, the euro. The ECB's main task is to maintain the euro's purchasing power and price stability in the euro area. The euro area comprises the 19 European Union countries that have introduced the euro since 1999 (43).		Notes: ECB's basic tasks: - The definition and implementation of monetary policy for the euro area; - The conduct of foreign exchange operations; - The holding and management of the official foreign reserves of the euro area countries (portfolio management); - The promotion of the smooth operation of payment systems (43).			

TERM	PES, (PSE) (abbr)	**Occurance:** 1811	**KW rank:** 402		
GE:, Sozialdemokratische Partei Europas	**CZ**: (PES) Strana evropských socialistů	**SK**: (PES) Strana európskych socialistov	1	47	46
Definition: The Party of European Socialists (PES) (formerly PSE (45)) brings together the Socialist, Social Democratic and Labour Parties of the European Union (EU). There are 33 full member parties from the 28 EU member States and Norway. In addition, there are 13 associate and 12 observer parties (44).		**Notes:** - The strengthening of the socialist and social democratic movement; - Contributing to forming a European awareness and to expressing the political will of the citizens; - Defining common policies for the European Union and to influence the decisions of the European institutions; - Leading the European election campaign with a common strategy and visibility (44),			

TERM	**ALDE** (ELDR), (abbr)	**Occurance:** 1353 (674)	**KW rank:** 488 (901)		
GE: (ALDE), Fraktion der Allianz der Liberalen und Demokraten für Europa	**CZ**: (ALDE), Skupina Aliance liberálů a demokratů pro Evropu	**SK**: Skupina Aliancie liberálov a demokratov za Európu	1	1	1
Definition: Political party active in the EU. Alliance of Liberals and Democrats for Europe Party (ALDE) European Liberal Democrats (ELDR) (41).		Notes: Formerly the European Liberal Democrat and Reform (ELDR) party, on 10 November 2012, European Liberal Democrat delegates voted overwhelmingly to change the name of the party to Alliance of Liberals and Democrats for Europe Party (ALDE) (41).			

Appendix E

Additional sources used during the glossary compilation:

a	EP Corpus
b	Linguee.com
c	Lingea Dictionaries
d	Collins dictionary
e	http://oxforddictionary.so8848.com/
f	Leo
g	Google search
h	Glosbe.com
1	http://iate.europa.eu/
2	http://www.politics.co.uk/reference/eu-enlargement
3	http://ec.europa.eu/enlargement/countries/check-current-status/index_en.htm
4	http://ec.europa.eu/enlargement/policy/steps-towards-joining/index_en.htm
5	http://www.oxforddictionaries.com/
6	http://dictionary.cambridge.org/dictionary/english/
7	https://www.wto.org/
8	http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+RULES-EP+20060703+RULE-001+DOC+XML+V0//EN&language=EN&navigationBar=YES
9	http://www.europarl.europa.eu/meps/en/about-meps.html
10	http://www.ecre.org/refugees/refugees/refugees-in-the-eu.html
11	http://www.allacronyms.com/EUROMED
12	http://www.schengenvisainfo.com/schengen-visa-countries-list/
13	http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/schengen/index_en.htm
14	https://www.europol.europa.eu/content/page/about-us

15	http://unfccc.int/kyoto_protocol/items/2830.php
16	http://www.icnl.org/research/monitor/osce.html
17	http://www.businessdictionary.com/definition/North-Atlantic-Treaty-Organization-NATO.html
18	http://www.nato.int/nato-welcome/index.html
19	http://www.guengl.eu/group/about
20	http://ec.europa.eu/transparency/regcomitology/index.cfm?do=FAQ.FAQ#1
21	https://www.fas.org/sgp/crs/row/R41959.pdf
22	http://frontex.europa.eu/about-frontex/mission-and-tasks/
23	http://www.investopedia.com/
24	http://www.greens-efa.eu/staff/press-webcommunications-and-multimedia/about-us/48-who-we-are.html
25	http://ec.europa.eu/eurostat/about/overview
26	http://www.merriam-webster.com/dictionary/parliamentarian
27	http://www.pravda.sk/trendove-temy/zostatnenie/
28	https://edis.ifas.ufl.edu/wc137
29	http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/visa-information-system/index_en.htm
30	http://www.europarl.europa.eu/enlargement/briefings/33a1_en.htm#summary
31	http://www.mdcr.cz/cs/Evropska_unie/Fondy_EU/PHARE.htm
32	http://archiv.vlada.gov.sk/phare/
33	http://www.nctc.gov/site/groups/hamas.html
34	http://www.britannica.com/topic/Hamas
35	http://en.euabc.com/word/931
36	http://eur-lex.europa.eu/legal-content/CS/TXT/?uri=uriserv:xy0024
37	https://ec.europa.eu/energy/en/topics/nuclear-energy
38	http://ec.europa.eu/fisheries/cfp/index_en.htm
39	http://slovnik-cizich-slov.abz.cz/web.php/slovo/trialog

40	http://www.europarl.europa.eu/news/en/news-room/content/20150217STO24619/html/Much-Ado-About-PNR
41	http://www.aldeparty.eu/en/about/the-alde-party
42	http://ec.europa.eu/enlargement/policy/glossary/terms/acquis_en.htm
43	https://www.ecb.europa.eu/ecb/tasks/html/index.en.html
44	http://www.pes.eu/about_us
45	http://tilastokeskus.fi/til/euvaa/2009/euvaa_2009_2009-06-12_tie_001_en.html
46	http://www.europskaunia.sk/socialisticka_skupina0
47	https://www.euroskop.cz/8626/sekce/strana-evropskych-socialistu/
48	http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:r17003
49	http://www.welcomeurope.com/european-funds/tacis-270+170.html#tab=onglet_details
50	http://www.esa.int/Our_Activities/Navigation/The_future_-_Galileo/What_is_Galileo
51	http://www.cfr.org/trade/mercosur-south-americas-fractious-trade-bloc/p12762
52	http://www.schengen.euroiuris.sk/index.php?link=schengensky_informacny_system
53	https://ec.europa.eu/europeaid/funding/funding-instruments-programming/funding-instruments/european-development-fund_en
54	https://www.google.sk/search?q=natura+2000+program&oq=natura+2000+program&aqs=chrome..69i57j0l5.2810j0j9&sourceid=chrome&es_sm=93&ie=UTF-8
55	http://www.nature.cz/natura2000-design3/sub-text.php?id=2102
56	http://ec.europa.eu/trade/policy/countries-and-regions/development/generalised-scheme-of-preferences/
57	http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32003H0361&from=EN
58	http://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition/index_en.htm
59	http://www.gmo-compass.org/eng/news/country_reports/
60	http://ec.europa.eu/social/main.jsp?catId=326
61	http://www.europeanlawmonitor.org/eu-legal-principles/eu-law-what-is-the-principle-of-proportionality-a-subsidiarity.html
62	https://www.wsws.org/en/articles/2013/07/13/nabu-j13.html
63	http://www.etrend.sk/trend-archiv/rok-2013/cislo-26/projekt-nabucco-padol.html

64	http://www.finance.cz/zpravy/finance/217880-rwe-eu-musi-vice-podporit-projekt-nabucco-hlavne-financne/
65	http://www.epp.eu/about-us/history/
66	http://ec.europa.eu/regional_policy/en/policy/cooperation/european-territorial/
67	http://ecrgroup.eu/about-us/the-ecr-in-the-european-parliament/
68	http://www.mfcr.cz/cs/zahranicni-sektor/ochrana-financnich-zajmu/financni-zajmy-evropskych-spolecenstvi/urad-pro-boj-proti-podvodum-olaf/zakladni-informace-olaf
69	http://www.infoplease.com/encyclopedia/history/european-economic-community.html
70	http://www.euractiv.cz/vzdelavani0/link-dossier/vzdelavaci-programy-eu
71	http://www.asean.org/asean/about-asean
72	http://www.europarl.europa.eu/aboutparliament/en/20150201PVL00010/Organisation-and-rules
73	http://slovnik-cizich-slov.abz.cz/web.php/hledat?cizi_slovo=multilateralni&typ_hledani=prefix
74	http://www.oecd.org/about/
75	http://ec.europa.eu/esf/main.jsp?catId=35&langId=en
76	http://www.mzv.cz/chisinau/cz/informace_pro_cesty_a_pobyt_konzularni/pro_obcany_cr/o_cestovani_do_podnestri_doporuceni_a.html
77	http://sk.sciencegraph.net/wiki/Vyvlastnenie
78	http://business.center.cz/business/pojmy/p1752-etatizace.aspx
79	https://managementmania.com/sk/hruby-narodny-dochodok-hnd
80	http://eeas.europa.eu/enp/
81	http://www.efsa.europa.eu/en/aboutefsa
82	http://thelawdictionary.org/
83	http://siteresources.worldbank.org/KFDLP/Resources/461197-1122319506554/What_is_the_Monterrey_Consensus.pdf
84	http://www.rozvojovka.cz/humanitarni-pomoc-a-rozvojova-spoluprace
85	http://www.mvro.sk/sk/rozvojova-spolupraca/rozvojova-pomoc-vo-svete
86	http://www.eurolabour.org.uk/about
87	http://www.euractiv.sk/danova-politika/analyza/tobinova-dan-trojsky-kon-eu-018490

88	https://www.euroskop.cz/9008/20178/clanek/financni-dan-72-z-kapes-londynske-city/
89	http://www.newworldencyclopedia.org/entry/Maghreb
90	http://www.collinsdictionary.com/
91	http://www.merriam-webster.com/dictionary/
92	http://www.diffen.com/difference/Effectiveness_vs_Efficiency
93	http://slovnik.juls.savba.sk/
94	http://ec.europa.eu/social/main.jsp?catId=102
95	http://www.neologismy.cz/
96	http://prirucka.ujc.cas.cz/
97	http://www.juls.savba.sk/ediela/sr/1987/5/sr1987-5-lq.pdf
98	http://www.juls.savba.sk/ediela/ks/2001/1/ks2001-1.html
99	http://www.juls.savba.sk/ediela/sr/2013/5/SR2013-5.pdf

Appendix F

Linguistic comparison of the EP corpus and other corpora

Source: Bick, E.: Degrees of orality in speech-like corpora: Comparative annotation of chat and e- mail corpora. In: Proc. of the 24th Pacific Asia Conference on Language, Information and Computation, pp. 721—729. Waseda University, Sendai (2010)

	Chat	E-mail	Euro-parl	BNC spoken	BNC written
function words	20.0 M	82.5 M	24.8 M	18.9 M	48.1 M
av. sentence length	8.74	19.71	***21.61***	17.27	18.12
av. word length	**4.4**	5.07	***5.27***	4.92	4.97
finite subclauses	4.32	**3.28**	4.29	***4.43***	4.09
relative	***1.96***	1.72	1.84	1.65	**1.57**
accusative	0.78	**0.64**	1.12	***1.28***	1.01
adverbial	1.25	**0.63**	0.93	***1.18***	1.12
gerund subclauses	***2.61***	1.43	**1.1**	1.2	1.3
infinitive subclauses	**1.57**	2.45	***2.48***	1.86	1.86
past part. subclauses	**0.21**	***0.42***	***0.37***	**0.21**	**0.22**
auxiliaries	**2.71**	5.06	***5.13***	4.10	3.79
active pcp2	**0.27**	0.55	0.72	***0.79***	***0.76***
passive pcp2	**0.33**	1.28	***1.48***	1.26	1.22
coordinating conj.	**3.14**	**3.36**	**3.52**	**3.56**	***3.76***
subordinating conj.	**1.33**	1.65	***2.04***	1.81	1.6
vocative	0.01	0	0.01	0.01	0.01
imperative	***0.35***	***0.5***	**0.05**	0.27	0.28
would, should, could	**0.41**	0.64	***0.8***	0.54	0.49
interjections	***0.92***	0.03	**0.01**	***0.56***	**0.1**
demonstrative	**1.04**	1.36	***2.23***	1.21	1.06
attributive	**5.15**	**5.51**	**7.51**	**7.74**	***8.42***
common nouns	25.61	***28.54***	**20.81**	21.71	22.62
proper nouns	**2.28**	**2.25**	3.89	4.18	***4.76***
finite verbs	***10.48***	10.21	9.36	10.92	10.47
personal & possessive pronouns	***12.36***	***3.32***	5.55	7.06	5.86