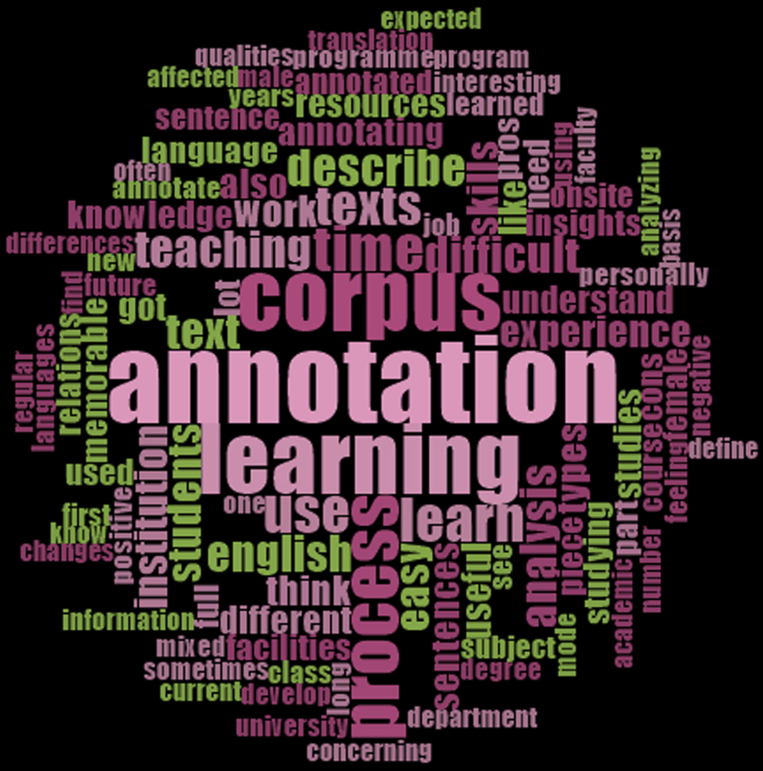
Corpus Analysis for Language Studies at the University Level



Giedrė Valūnaitė Oleškevičienė,
Liudmila Mockienė and Nadežda Stojković

Corpus Analysis for Language Studies at the University Level

Corpus Analysis for Language Studies at the University Level

By

Giedrė Valūnaitė Oleškevičienė,
Liudmila Mockienė
and Nadežda Stojković

Cambridge
Scholars
Publishing

Corpus Analysis for Language Studies at the University Level

By Giedrė Valūnaitė Oleškevičienė, Liudmila Mockienė
and Nadežda Stojković

This book first published 2021

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-6470-3
ISBN (13): 978-1-5275-6470-1

Contents

Acknowledgements

We are grateful to Mykolas Romeris University, the Institute of Humanities, and the University of Niš, Faculty of Electronic Engineering, for the environment conducive to research that inspired and equipped us with the means necessary for the research.

Concepts Used in the Monograph

Communicative language competence could be defined as comprising four competence areas, namely, linguistic, sociolinguistic, discourse and strategic; and each component in its own turn comprises knowledge and skills and know-how. (Council of Europe 2011)

Corpus is a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject. (Merriam-Webster Online Dictionary)

Corpus annotation is defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. (Leech 1997)

Corpus linguistics is the study of language based on the samples of corpora containing real-world texts. (Sinclair 1992)

Comparable corpus is one which selects similar texts in more than one language or variety. (Sinclair, 1996)

Discourse could be defined as written or spoken communication, or a mode of organizing knowledge, ideas or experience that is rooted in language and its concrete contexts. (Merriam-Webster Online Dictionary)

Higher education is education beyond the secondary level, especially education provided by a college or university. Institutions of higher education include not only colleges and universities but also professional schools in such fields as law, theology, medicine, business, music and art. They also include teacher-training schools, community colleges and institutes of technology. At the end of a prescribed course of study, a degree, diploma or certificate is awarded. (Kraujutytė 2002)

Inductive qualitative research is often referred to as a "bottom-up" approach to knowing, in which the researcher uses observations to build an abstraction or to describe a picture of the phenomenon that is being studied. The inductive approach enables researchers to identify key themes in the area of interest by reducing the material to a set of themes or categories. (Lodico, Spaulding, and Voegtle 2010)

Parallel corpus is a collection of texts, each of which is translated into one or more other languages than the original. (Sinclair, 1996)

Translation competence could be defined as including an array of knowledge, skills and abilities, so-called translation skills, which are exhibited through a translator's ability to juggle the forms of the languages in order to produce the translation requested by the contemporary language norms. (PACTE 2000)

References

Council of Europe (Council of Europe Council for Cultural Co-operation, Education Committee, Modern Languages Division). 2011. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Strasbourg: Cambridge University Press.

Kraujutytė, L. 2002. *Aukštojo mokslo demokratiškumo pagrindai*. Vilnius: LTU.

Leech, G. 1997. "Introducing corpus annotation." *Corpus annotation: Linguistic information from computer text corpora*, 11–28. Routledge.

Lodico, M., D. Spaulding, and K. Voegtle. 2010. *Methods in educational research: From theory to practice*. San Francisco, CA: John Wiley & Sons.

Merriam-Webster Online Dictionary. n.d. Accessed April 20, 2019. http://www.merriam-webster.com.

PACTE (Action Plan for Business Growth and Transformation). 2000. "Acquiring translation competence: hypotheses and methodological problems of a research project." In *Investigating Translation*, edited by Allison Beeby, Doris Ensinger and Marisa Presas, 99–106. Amsterdam: John Benjamins.

Sinclair, J. 1992. "The automatic analysis of corpora." In *Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82)*, edited by J. Svartvik. Vol. 65. Berlin: Mouton de Gruyter.

Sinclair, J. 1996. Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P. Available at: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html. Accessed March 3, 2020.

INTRODUCTION

Corpora development has stimulated the ongoing progress in the advance of knowledge concerning lexis, grammar, semantics, pragmatics and textual features (Sinclair 1991; Stubbs 2004). Its increasing relevance is related to the fact that corpus linguistics focuses on sources of naturally occurring, spontaneous, uncensored, real-life data regarding language use. Since context is crucial in researching and describing language use, this aspect is also related to corpus linguistics tools and analyses presenting extensive contextual information about sociolinguistic metadata. Therefore, the approach to teaching foreign languages is now changing due to the impact of technology which allows the use of current crucial linguistic data, empirically obtained and thus trustworthy, regarding actual language use in context.

There is a need for cross-fertilization between corpus research and its application in language teaching settings (Mukherjee 2004; Römer 2009; Widdowson 1990, 2000). According to recent studies, corpus analysis has been applied to carry out research on vocabulary quite extensively as corpus analysis tools can provide great amounts of information on such aspects of lexical items as their frequency, semantic and syntactic environment (Rundell 2008). Different types of corpus software comprise a variety of tools which could be used to analyze lexis, including frequency wordlists, concordance lines, key words in context (KWIC), term extraction, collocates, colligates, taggers and lemmatizers. The extracted information could be used for all kinds of lexicographic research activities, such as compiling term banks, glossaries, dictionaries, terminology databases and translation memory databases. As Zanettin (2002) observes, there is value not only in using specialized corpora but also in their creation per se. Laurence Anthony, the developer of AntConc freeware—a well-known corpus toolkit—states that corpora and corpus tools are of great value not only for researchers of languages but also for teachers and learners (Anthony 2009). The studies by Cobb and Boulton (2015) reveal that the innovative idea of using corpora in teaching and learning appears to be effective and efficient. According to Boulton and Tyne (2014), data-driven learning (DDL) comprises a number of crucial concepts in the existing approaches of language learning, such as authenticity, autonomy, cognitive depth, consciousness-raising, constructivism, context, critical thinking, discovery learning, heuristics, ICT, individualization, induction, learner-centeredness, learning to learn, lifelong learning, (meta-) cognition, motivation, noticing, sensitization and transferability. Therefore, the authors support DDT (data-driven teaching) as it can provide the necessary exposure to authentic language.

The current study focuses on corpora use in teaching foreign languages in university education, which comprises teaching foreign languages in both non-linguistic and linguistic departments. Corpus analysis tools can be employed in teaching English at university level for corpus compilation, data extraction, and further contrastive and linguistic (especially lexical) analysis. It can be given as an assignment in the form of a project or case study to students who study philology (linguistics) or even those who study English for Specific Purposes (ESP) as a part of their course assessment. Corpus analysis tools can also be used by students of philology (linguistics) who write their course papers or bachelor's or master's theses.

The problematic areas for advanced language learners seem to be coherence, cohesion and textual rhetorical features. Thus, cohesive devices and discourse markers get the researcher's attention as the tools for ensuring textual and discourse management. Research on proper discourse use is looking for answers as to what could be taught (and how) at more advanced levels concerning the matters of textual features. The suggestions offered by the recent research lead to the idea of direct corpus use by language learners and teachers. The studies by Cobb and Boulton (2015) show that the application of such an advanced idea of using corpora in teaching and learning appears to be really effective. Fawcett (1987) observes that corpus-based teaching and learning could be a promising means of translator preparation because the purpose of translator education is to equip trainees with skills applicable to any texts related to any subjects, and corpus-based teaching can provide trainees with such skills. The author stresses that corpus-based translation classes enable students to learn about corpora, corpus analysis tools and their applications for translation. The current research focuses on the process of teaching and learning a foreign language at more advanced levels while applying corpus analysis and building tools for corpus annotation. It envisions looking deeper at the experience of students and teachers in the study environments enriched with corpus analysis and building tools, and at how the research participants perceive their experience of the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies. Additional research questions embrace such matters as the following: what features does the meaning of the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies consist of; and what dimensions emerge in the perceived meaning of the use of corpus analysis and building tools for corpus annotation and use in teaching and learning a foreign language at more advanced levels in university studies by teachers and students.

Research object. The research object is the meaning of using corpus analysis and building tools for data extraction and annotation in teaching and learning a foreign language at more advanced levels in university studies. The research investigates the phenomenon of corpus design and annotation use in teaching and learning a foreign language at more advanced levels in university studies with the particular focus on the meaning of the "lived experience" of the research participants.

Research aim and objectives. This investigation belongs to the qualitative research paradigm, which contributes to the broad research field with multiple approaches to the use of corpora in university studies. The aim of the present research is to investigate the phenomenon of the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies based on its participants' lived experience. The meaning is revealed through exploration of teachers' and students' personal stories of the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies. Pursuing the research aim, the following research objectives have been set:

1. To present the discourse on the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies.
2. To describe in a structural way the lived experience of the research participants—teachers and students—while using corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies.
3. To disclose the recommendations for the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies.

The research field is comparatively new and developing, still embracing many unanswered questions. The question of the human factor seems to be important in researching the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies as human factor in the study environments saturated with technologies of corpus analysis and building tools cannot be easily counted. In this context the research of the phenomenon of the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies is absolutely relevant and new as it is directed to look deeper into the phenomenon and find out how the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies could have some enhancing effect.

Corpora use is penetrating into the university studies arena. Thus, the research on the phenomenon of the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies is a scientific research input into the vast field of the research on corpora educational use. The research creates better understanding of the use of corpus analysis and building tools in teaching and learning a foreign language at more advanced levels in university studies by revealing how university study participants—teachers and students—make sense of the use of corpus analysis and building tools for corpus annotation in teaching and learning a foreign language at more advanced levels in university studies through their own lived experience. The results of the research enable us to provide recommendations for the use of corpus analysis and building tools in university studies and also envision areas for future research.

Methodology of the research (methods and implementation). The qualitative research paradigm was applied as it helped us to understand human experience in a specific context (Creswell 2007) and thus is suitable for researching the human experience in the study environments while applying corpora tools. Qualitative inductive content analysis by Elo and Kyngas (2007) was chosen as a core method for the current research depending on the research question, as the current research is intended to investigate how the participants make sense of teaching and learning while applying corpus analysis and building tools for analyzing textual cohesion using discourse connectives through their own lived experience. The authors analyzed the research participants' experience in a structural way by aiming to formulate certain conclusions and recommendations for using corpus analysis and building tools while teaching and learning a foreign language at more advanced levels. Qualitative inductive content analysis by Elo and Kyngas (2007) enables structural analysis of teaching and learning experiences while applying corpus analysis and building tools for analyzing textual cohesion through discourse connectives. The structural analysis of the meaning which research participants ascribe to shared lived experience helps us to examine the real situation (how things really are) and make certain conclusions and recommendations. In education it could theoretically be known how matters should be, but it is a sensitive area where regulations and instructions may clash with human realities, and research may reveal certain areas for improvement.

Students and teachers were included in the interview series to ensure well-rounded understanding, and semi-structured interviews (Ghiglione and Matalon 2001) were performed. The inductive qualitative content analysis was carried out applying NVivo, which is a well-established and efficient software product widely used for organizing and managing data. The authors instantaneously analyzed the interviews just after the interviews by constantly comparing the structuralized data material. The data have undergone several coding stages, starting with initial open coding and followed by axial coding and selective coding.

Limitations. The choice of the qualitative research paradigm involving qualitative inductive content analysis might be considered as strength of the research. Qualitative research does not imply making any assumptions before the research starts. Qualitative research methods facilitate capturing stories of participants' own experience; what is more, qualitative research has the power of sensitively registering human realities in education environments and revealing the real state of the situation. On the other hand, the application of an exclusively qualitative approach might be perceived as a limitation since the current research is only focused on the subjective perspective of corpus annotation use in teaching and learning a foreign language at more advanced levels in university studies. The research would have been enriched if different perspectives—e.g., technology enhanced learning and teaching based on a constructivist approach and objective measurement had been added to the research; then a more comprehensive understanding of the use of corpus analysis and building tools in language studies at university level could have been obtained. However, it should also be acknowledged that research based on objective measurement would have been a separate additional study.

Having interview material as the only empirical data source could be considered another limitation as, for example, Silverman (2005) suggests using multiple sources to obtain a more extensive understanding of a phenomenon. However, Ghiglione and Matalon (2001) advocate for using a method of semi-structured interviews as the most suitable means for obtaining empirical data. The authors argue that interviews provide a perfect opportunity to deepen the understanding of a phenomenon through the subjective perspective of the research participants, to register the subtleties which are seldom explored.

The structure of the monograph. The monograph is organized into three chapters. Chapter One provides a brief review of teaching foreign languages in the settings of non-linguistic departments. It presents generic attributes, the importance of communication and social skills, teaching and learning foreign languages for employability, and the relevance of translation and corpus linguistics for learning material design in the discussed settings. Chapter Two focuses on the application of corpus analysis and building tools in teaching English at university level for corpus compilation, data extraction, and further contrastive and linguistic (especially lexical) analysis. It provides a detailed case study of analyzing terminology of constitutional law in English and Lithuanian as an example to illustrate the possibility of integrating corpus analysis tools into the process of teaching and learning languages at more advanced levels. Chapter Three provides a brief theoretical background focusing on corpora application in language studies, followed by a discussion of certain issues in discourse management and organization, and closes with insights on principles of teaching and learning with technology and the role of the initial knowledge. The authors also explain the methodological approach of the research by providing the grounds for the methodological choices of the qualitative research and describing the research procedures. Finally, the results of the research are presented and the authors provide recommendations for teaching and learning a foreign language at more advanced levels while applying corpus analysis and building tools.

References

Anthony, L. 2009. “Issues in the Design and Development of Software Tools for Corpus Studies: The Case for Collaboration.” In *Contemporary Corpus Linguistics*, edited by P. Baker. Continuum.

Boulton, A., and H. Tyne. 2014. *Corpus-based study of language and teacher education*. New York: Routledge.

Cobb, Th., and A. Boulton. 2015. Classroom Applications of Corpus Analysis.” In *The Cambridge Handbook of English Corpus Linguistics*, edited by D. Biber-Reppen, 478–97. Cambridge: Cambridge University Press.

Elo, S., and H. Kyngas. 2007. "The qualitative content analysis process." *Journal of Advanced Nursing* 62 (1): 107–15.

Fawcett, P. 1987. "Putting translation theory to good use." In *Translation into Modern Languages Degree*, edited by Hugh Keith and Ian Mason, 31–38. London: CILT.

Ghiglione, R., and B. Matalon. 2001. *O Inquérito: Teoria e Prįtica*. Oeiras: Celta Editora.

Mukherjee, J. 2004. "Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany." In *Applied Corpus Linguistics: A multidimensional perspective*, edited by U. Connor and T. Upton, 239–50. Amsterdam: Rodopi.

Römer, U. 2009. "Corpus research and practice: What help do teachers need and what can we offer?" In *Corpora and Language Teaching*, edited by K. Aijmer, 83–98. Amsterdam: John Benjamins.

Rundell, M. 2008. "The corpus revolution revisited." *English Today* 24 (1): 23–27.

Silverman, D. 2005. *Doing qualitative research: A practical handbook*. London: SAGE.

Sinclair, J. M. 1991. *Corpus, concordance collocation*. Oxford: Oxford University Press.

Stubbs, M. 2004. "Language corpora." In *The handbook of applied Linguistics*, edited by A. Davies and C. Elder, 106-113. London: Blackwell Publishing.

Widdowson, H. 1990. *Aspects of language teaching*. Oxford: Oxford University Press.

Widdowson, H. 2000. "The limitations of linguistics applied." *Applied Linguistics* 21 (1): 3–25.

Zanettin, F. 2002. "DIY Corpora: The WWW and the Translator." In *Training the Language Services Provider for the New Millennium*, edited by Belinda Maia, Jonathan Haller and Margherita Urlrych, 239–48. Porto: Facultade de Letras, Universidade do Porto.

CHAPTER ONE

THEORETICAL CONSIDERATIONS FOR TEACHING FOREIGN LANGUAGES IN UNIVERSITY EDUCATION SETTINGS

NADEŽDA STOJKOVIĆ

Teaching foreign languages (FLs) in university studies comprises teaching foreign languages in both non-linguistic and linguistic departments. The current chapter provides an overview of teaching foreign languages in the settings of non-linguistic departments. It presents generic attributes, the importance of communication and social skills, teaching and learning foreign languages for employability, and the relevance of translation and corpus linguistics for learning material design in the discussed settings.

1.1 Generic attributes

Teaching foreign languages in university settings, in non-linguistic departments, is present throughout European settings and beyond. It follows instruction at previous formal educational levels, and the preconditions for course entry most often imply that complete grammar, syntax and vocabulary have been covered, all up to B2 level according to the *Common European Framework of Reference for Languages* (Council of Europe). Language instruction at university level is the final stage before students enter the job market, which today is highly mobile and inherently international in character, and therefore requires language skills that enable successful, immediate and precise conveying of expertise. For these reasons, universities have in their curricula incorporated mission statements on institutional objectives and graduate attributes that include language skills. Those are interchangeably referred to as generic attributes of graduates, described with generic terms of the intended learning outcomes, such as: specialist knowledge, general intellectual skills and capacities, and particular personal qualities, which are developed through university education with the aim of enhancing students' cognitive and affective attributes and abilities.

These objective statements and graduate attributes, have become vital in assessing whether the university curriculum is the direct and foreseen response to, and the accommodation for society's changing directions and aspirations (Barnett 1990). There are various societal requirements that influence the formulation of those attributes, central among which is the call for universities to educate more employable graduates, in alliance with the employable skills agenda of industry and governments, and in that way the call forms a vital intersection, a focal point of convergent forces shaping the society. Here it is obvious that the contemporary university setting, speaking in worldwide terms, is directly shaped by linking national educational policies and economic growth agendas (Woodhouse 1999), at the same time producing new quality assurance standards for HE institutions internationally. These requirements are at present increasingly more difficult to conceptualize, meet and formulate in curricula regarding the information explosion and the consequent proliferation of accessing knowledge (Barnett 2000).

"Generic graduate attributes" is the most widely accepted term denoting that the targeted educational results encompass more than personal skills and attitudes; rather, new personal characteristics reach out beyond mere disciplinary content knowledge and are applicable in a range of social contexts, including international ones. For these reasons they are also termed core, key or transferable (Bowden et al. 2000). These attributes are considered—rather than domain knowledge, which they transcend—central achievements of university studies, applicable to a range of contexts, because it is through them that a person is prepared to successfully enter the world of work, to be a global citizen and an effective member of contemporary society.

This all reflects the fact that university settings are changing under the influence of neoliberal societal attitudes that align the goals of (governmental, university) educational policies, business and scientific development (Giroux 2010; Olssen and Peters 2005) in the contemporary, international, supranational knowledge economy, yet taking care not to commodify teaching and learning (Cribb and Gewirtz 2013). This is why Barnett (2000) summarizes university studies goals as educating students to be able to independently cope with dynamic employment perspectives, and teaching them how they can provide positive contributions to the current heterogeneous communities, not only of practice but of their entire lives. In this way, it is clear graduate attributes reach significantly beyond mere employability.

They reflect university studies creating career competencies, and academic citizenship as well. These competencies, subsumed as graduate attributes, include development of personal qualities such as ethical, moral and social responsibility, intercultural awareness and personal integrity, and at the same time multiple and diverse skills, some of those being critical thinking, intellectual curiosity, problem solving, reflective judgment, leadership and team work, information literacy, digital literacy and effective communication skills.

1.2 Communication and social skills

Communication in this setting implies native language skills as improved through domain subjects taught in it, and foreign language skills, which in university studies is a foreign language for specific purposes (LSP). Communication is referred to together with social skills to emphasize their mutual interdependence; this reflects citizenship characteristics necessarily intertwined with employability, as these two come to be inseparable. Communication that is to be developed in university studies refers to oral, written and effective listening skills in national, international and cross-generational environment, contributing to productive and harmonious relations in business settings. Communication and social skills are therefore the ability to communicate and collaborate independently and/or in teams across professional and social settings. This ability is seen as critical for sustained and successful employment. Perfected communication and social skills incorporate careful listening, clear, appropriate formulation, and conveying of ideas, information and responses in various formats.

In some universities’ goals statements, communication and social skills are referred to as “social communication skills” or “communicative language competence”, reflecting the inseparableness of the two, and including teaching students how to use language for a range of functions, like asking for or providing information, negotiating, arguing or clarifying issues; conversational skills, such as introducing a topic, maintaining it through the smooth flow of conversation, being appropriate and politely taking turns in conversation; understanding assumed knowledge and implied meanings of the listener(s); non-verbal communication, such as significance and meaning of eye contact, facial expressions, gestures, and culturally modeled physical proximity and distance.

Many of these skills are perfected indirectly through students being taught major subjects in their native language. It is the very way professors speak and act that conveys their personal mastery of these skills to students, who are passive recipients while listening to lectures and active when they need to reproduce the knowledge. This is the induction model of transferring these necessary generic attributes, and the transfer happens without much conscious or reflective awareness of it on the part of both lecturers and students. Therefore, regarding communication skills, both verbal and non-verbal ones, those of the student's mother tongue are transferred thus through domain subject professors and associates working on them, while foreign language communication skills are dealt with in specialized foreign language courses.

1.3 Teaching and learning a foreign language for employability at university

In post-secondary education, teaching a foreign language in most cases[1] represents the continuation of the language instruction, building upon already acquired language proficiency towards higher levels. Then, on the basis of language content covered in previous educational stages, it is assumed that students possess sound knowledge of general English (GE), possibly with some elements of the target science they are commencing to study, up to the upper intermediate level of proficiency. Very often the requirements for FL course entry at university state precisely that this has been achieved previously. Then, the focus of FL instruction shifts from GE to language needed for professional and scientific settings that students are preparing for, in line with their major. This means that the format of FL instruction at university is that of languages for specific purposes (LSP) and academic FL.

Teaching LSP is the most common form of FL instruction in academia, it being in accord with the profile of the major studies and, at the same time, with prospective job positions in that field. *Instruction in LSP provides for multiple goals: it teaches communicative, social, transferable employability skills*. In what follows, this claim will be elaborated on and supported.

The LSP syllabus is conceptualized according to the curriculum of the faculty/university where the course is taught. A long while prior to LSP course commencement, lecturers conduct various types of research regarding the profile of the institution. They inquire into the content of the curriculum and subjects related to the major individually. This is only the

[1] When a second foreign language is introduced at university, then the instruction begins from the beginner level.

first instance demonstrating the particularly demanding position of LSP lecturers. Since they have not been educated in the science they are to conceptualize a language course for, it is certainly difficult for them to ponder, comprehend, analyze and segment such content; moreover, they must possess abstract linguistic characteristics that will allow for creating a syllabus that simulates communication in real scientific and professional situations. In this they resort to analyzing the curriculum, interviewing major subject lecturers and doing the research on their own. In the early days of LSP consolidated theory, two of its major and still most referred-to experts, Hutchinson and Waters (1987), said LSP practitioners are “solitary travelers into uncharted lands”, subsuming the challenges and real difficulties within this vivid and potent metaphor.

Upon collecting necessary material on the content of the curriculum, lecturers are then truly left on their own to design the syllabus of an LSP course. This is when yet another difficulty is encountered. As an inherent characteristic of LSP courses, and given that the justification for their existence is to linguistically “serve” the major subjects of the given, particular institutional profile, and future professional profile of the students, the availability of ready-made teaching and learning material is questionable. Big international publishing houses that offer books on LSP (though most often it is ESP), produce material that is of a specific purpose, yet far too general at the same time. Even as such, two characteristics are striking. First, such books almost never reach beyond intermediate level. This in itself contradicts the premise explained earlier in the text here, namely that LSP instruction at university is the continuation and upgrading of the foreign language proficiency already gained in the previous stages of education, and that the entrance requirement for an LSP course is having acquired intermediate-level skills. Another striking characteristic is the segmentation of texts and exercises in those books. Students “study”—in the original, Latin meaning of that word, as in thorough devotion, adherence, diligence and industriousness, which in themselves are transferable skills. Thus, batches of short exercises, common in LSP textbooks, comprising most often up to ten exemplary sentences, or very short texts for reading and analysis, are all inherently incompatible with the overall aim of university studies—to study thoroughly.

Another peculiarity tightly connected to LSP material design is the position of LSP lecturers and the contemporary fast-changing nature of sciences. First, frequently there is just one lecturer at the institution. The task of comprehending and navigating through the content of major studies would be a meaningful task for a team. On top of all that, lecturers can rarely harvest previously designed and used material, as the very curriculum changes to include the advances in the sciences studied, and the content of foreign language courses is to follow them.

Despite these significant challenges to the post of an LSP lecturer, this approach to language study has matured over the past decades to become a professional lingua franca, with needs analysis and discourse analysis its most prominent aspects that serve students for successfully entering the work community they are preparing for in their studies. The number of ways of producing and designing teaching material have recently been on the rise, most notably due to the resourcefulness and availability of technologies that support individual, original coverage of relevant texts and practices, as well as their dissemination and so further use and upgrading.

1.4 Relevance of active use and practice of translation in LSP

After the period of the communicative approach in language teaching methodology that functioned almost to the complete exclusion of translation, active fostering of this skill in students is now emphasized for the benefits it brings to their understanding of the two languages in question, but equally so for their comprehension and internalization of the content knowledge, particularly in the fields where accuracy is vital in communicating rigorous information through a reliable linguistic medium. The methodology of teaching translation relies on the use of authentic materials; it is interactive, learner-centered and promotes learner autonomy, all in particular valid for LSP teaching and learning at university as a preparation for a prospective entry into a job post. Teaching translation at university studies language instruction has become relevant for the numerous outstanding advantages it offers, most broadly listed as heightened awareness of the language(s) use, enhancement of cognitive and receptive skills, and certainly instruction in necessary pragmatic and stylistic approaches to target language use (Fernández-Guerra 2014, 155; Dagiliene 2012, 124). Translation practice forces students to actively ponder semantic meaning, not mechanically substitute words in two languages, and so to think comparatively between them. Through this process they can comprehend the non-parallel nature of languages which compensates for the absence of perfect, one-to-one correspondence, all to their own advantage when using either. In addition, students become aware of the often-characteristic positive and negative transfers, and so better understand the target language. This shift of the emphasis, the revival of interest in translation, was partly caused by findings that the use of a native language does promote language learning, and that through translation qualities like accuracy, clarity and flexibility, that are essential to any language learning, and which are generic in nature, are further promoted (Duff 1994). Also, translation in the higher stages of language learning, as in university studies, is observed as the fifth language skill along with the four basic ones (listening, speaking, reading, writing). "Translation holds a special importance at an intermediate and advanced level: in the advanced or final stage of language teaching, translation from L1 to L2, and L2 to L1, is recognized as the fifth skill and the most important social skill since it promotes communication and understanding" (Ross 2000).

In line with inherent LSP characteristics, the best-suited approach to teaching translation within university studies is found to be functionalist, in which a text is seen as an "offer of information", a segment of the overall communication action within a specific discipline. Students are to be instructed to conceive of themselves when translating as choosing information elements they consider necessary to achieve the purpose of the original text and transfer it by constructing a new text in the target language. For this, they need to take into account the communicative framework of the particular discipline and conform to it (Berkenkotter and Huckin 1995, 1). This implies that LSP translation fosters the interdisciplinary concept of specialized communication, transgressing far beyond only relevant linguistic approaches to include cognitive, knowledge-oriented semiotic approaches. To illustrate this, in practice it often means directing students to actually "retell" the source text in the target language, taking all the care to transfer precisely the whole information load, and not focus on linguistic correspondence.

The benefits of practicing translation are numerous. Through translation practice, LSP students at once exhibit acquired specialized domain knowledge and in turn foster it further by interiorizing specialized knowledge systems through texts on which they work. In LSP instruction, translation is often crucial as often accurate equivalence is needed; at the same time the work on authentic texts is a necessary requirement in a syllabus to cater for the students' needs. Further along this line, as regards certain specialized texts, at present they are primarily characterized by the highly frequent appearance of new terminology, as a result of social, cultural, scientific and economic alterations. It is therefore true that original texts are most often the most reliable and most representative sources for learning domain language in its natural, vivid, accurate form. Translation theory and practice of the twentieth and twenty-first centuries developed strategies that are useful when students face unknown terminology or that which can be characterized as barely translatable. Those strategies have proven only to heighten the efficiency of LSP instruction. This work on such texts is valuable as it further enhances the skills of students to search for relevant information on their own, an important aspect of LSP education and acquisition of domain knowledge. The required meeting of students' needs is carried out through this work on the text of specific contexts as they make "use of underlying methodology and activities of the discipline" (Fortanet-Gomez and Räisänen 2008, 61). Thus, LSP in this way, too, proves to be eclectic as combining linguistic and domain-specific methodology, making the knowledge aspect central for the success of the teaching process.

Translation practice leads students to gain valuable insights into characteristics of both languages by necessarily having to compare the given texts. When exploiting this, language learners themselves indicate language areas in which they need to improve, those findings being highly valuable for lecturers as well. This is the side of translation showing how it assists students in developing primary communicative skills. Unlike students of philology departments, when embarking on translation practice ESP students do not need translation theory instruction; their needs are different and therefore they benefit from smaller-scale directions regarding techniques of translation: "It is not essential to be an expert in translation and translation theory to use translation in class" (Witte and Harden 2009, 176). Through exposure to various disciplinary texts, students also practice intercultural communication. Commenting on the relation between translation and intercultural generic communicative skills, Pym (1996, 337) states: "I tend to see the purpose of translation as a privileged index of wider intercultural phenomena and translation theory as a source of interesting models for such relations." That translation is a practice in language teaching that has multiple benefits, including learning the foreign language, intercultural communication, domain knowledge and generic competencies throughout, is summarized by Leonardi (2009, 141) who stated: "The role of translation is thus fundamental in teaching and showing students mediation strategies and both linguistic and cultural differences through employing a contrastive approach to language. Through translation, students can learn more about problem-solving strategies, improve their analytic skills and strengthen their grammatical and lexical competence and performance."

1.5 Benefits of corpus linguistics for LSP teaching and learning

With the advent of powerful and available computers, various language learning software and software tools keep appearing that now strongly influence the development—and moreover determine the further directions—of foreign language learning research and practice, particularly in university studies. The most prominent tool is the emerging field of corpus linguistics, primarily seen as direct access to actual discourse patterns in both spoken and written language in target social settings of GE or LSP. Corpus linguistics is criteria-determined analysis of principled collections of language, of particular discourse, in an electronic format, called "corpora". This new approach to the study of language was initiated with the newly discovered ability of computers to store large amounts of data, and consequently the era of mega-corpora such as the Collins Corpus and Bank of English (each approx. 2.5 billion words), and the Oxford English Corpus and the Cambridge English Corpus (each approx. 2 billion words in size), compiled for lexicographical purposes. At the same time, corpus linguistic methodology started to be exploited for research by other linguistic frameworks, smaller in size and dedicated to a certain segment of pragmatic use of language, such as conversation analysis and spoken discourse analysis. A particular relevance of such smaller corpora that keep emerging is the fact that they facilitate a "constant interpretive dialectic between features of texts and the contexts in which they are produced" (Vaughan and Clancy 2013, 70), which makes them directly useful for actual work in Foreign Language Teaching (FLT). For these reasons, here only briefly sketched (to be elaborated on in further chapters), it is clear that the use of corpora, the authentic linguistic data—even called a "corpus revolution" (Rundell and Stock 1992)—informed a whole new output of reference and pedagogical materials in FLT, thus now having a decisive influence on second/foreign language teaching. Corpus analysis is now indispensable "in virtually all branches of linguistics or language learning" (Leech 1997, 9), as its strength is its empirical nature, making linguistic analysis more objective (McEnery and Wilson 2001, 103).

Its growing relevance is due to the fact that corpus linguistics offers sources of naturally occurring, spontaneous, uncensored, real-life data on language use. As context is crucial in describing language use, this aspect is also included in corpus linguistics tools and analyses, providing extensive contextual information in the form of sociolinguistic metadata. Therefore, the impact of technology allowed for current crucial linguistic data, empirically obtained and thus trustworthy regarding actual language use in context, that is now changing the approach to and execution of GE and LSP teaching. Corpus linguistics allows for compiling frequency lists, particular necessary specifications of textual features, text types and genres, grammatical patters, collostructions and much more, all leading to creation of data-driven learning activities. Those are crucial for FL development in learners, as they incite development of pragmatic competence as "the ability to use language effectively in order to achieve a specific purpose and to understand language in context" (Thomas 1983, 92).

Such characteristics of corpus linguistics make its findings particularly relevant for teaching LSP in university settings, where mastering genres and specialized registers is essential and the empirical material which is provided in a corpus-informed approach becomes indispensable. Corpus analysis is a foundation for an empirically based understanding of discourse and language for specific purposes. This outstanding relevance of corpus linguistics calls for its larger inclusion in actual teaching practice; there is a need for a cross-fertilization between corpus research and its application in language teaching settings (Mukherjee 2004; Römer 2009; Widdowson 1990, 2000). In LSP, corpora and corpus-driven learning are particularly useful for the lexico-grammar of its contextualized, domain language varieties. Those varieties that need to be taught in LSP instruction, while obviously conformant with the overall syntax and semantics of the language in question, are characterized by the selective occurrence of certain structures and the prevalence of domain-specific, conventionalized phraseologies and patterns (e.g., collocations, lexical bundles), as well as the present-day extremely fast evolution of new scientific and professional terminology. For these, corpora become crucial, as traditional reference books and dictionaries now cannot compete with web-based corpora with regard to lexical and terminological evolution record.

As there are opinions that corpora use is not exploited in classroom teaching to its fullest extent, this monograph is also dedicated to exemplifying how this situation can be changed for the benefit of both students and lecturers, offering to the former the real-life language examples, and to the latter an invaluable resource to assist them in material design. Corpora help lecturers indirectly, in deciding what to teach, but also in their direct use, regarding how to teach. The reason some theorists argue that the majority of the existing, publicly available corpora are not widely used in teaching practice can be summarized as the fact that they have been developed "as tools for linguistic research and not with pedagogical goals in mind" (Braun 2007). This calls for development of pedagogically motivated corpora that need to be "complementary to school curricula, to facilitate both the contextualisation process and the practical problems of integration" (Braun 2007, 310). The potential of corpora is such that Conrad (2000) spoke of them as a means that will thoroughly change the teaching of foreign languages and the overall language education, to include both what is taught and how it is taught. Moreover, well developed corpora, as Gavioli and Aston (2001) claim, are also viewed as resources for students' autonomous study, which is one of crucial goals of LSP teaching methodology. An independent, self-study capable LSP learner profile can be more successfully attained through learner-centered, individualized methods of learning, harvesting the benefits of corpora use (Johns 1990).

In the following chapters there will be further both theoretical, more detailed and in depth elaboration of the theoretical stances here summarized, as well as the empirical research on the use of corpora in the practice of language studies at university level that proves its direct benefit for the teaching/learning outcomes of foreign languages university courses.

References

Barnett, R. 1990.*The Idea of Higher Education*. Buckingham: Society for Research into Higher Education & Open University Press.

Barnett, R. 2000.*Realizing the university in an age of supercomplexity*. Buckingham: Society for Research into Higher Education & Open University Press.

Berkenkotter, C., and T. N. Huckin. 1995. *Genre Knowledge in Disciplinary Communication: Cognition / Culture / Power.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bowden, J., G. Hart, B. King, K. Trigwell, and O. Watts. 2000. Generic Capabilities of ATN University Graduates. http:/www.clt.uts.edu.au/ATN.grad.cap.project.index.html.

Braun, S. 2007. "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora."*ReCALL*19(3): 307–28.

Conrad, S. 2000. Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, *34*, 548–560.

Council of Europe. The CEFRL levels. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions. Accessed May, 15th 2019

Cribb, A., and S. Gewirtz. 2013."The hollowed-out university? A critical analysis of changing institutional and academic norms in UK higher education."*Discourse: Studies in the Cultural Politics of Education* 34, 338–50.

Dagiliene, I. 2012. "Translation as a Learning Method in English Language Teaching."*Studies About Languages*, no. 21, 124–28.

Duff, A. 1994.*Translation: Resource Books for Teachers*. Edited by A. Maley. Oxford: Oxford University Press.

Fernández-Guerra, A. B. 2014. "The Usefulness of Translation in Foreign Language Learning: Students' Attitudes." *International Journal of English Language & Translation Studies* 2, no.1 (January–March): 153–70.

Fortanet-Gomez, I., and Ch. Räisänen. 2008. *ESP in European Higher Education: Integration Language and Content.* Amsterdam: John Benjamins.

Gavioli, L., and G. Aston. 2001. "Enriching reality: language corpora in language pedagogy." *ELT Journal* 55(3): 238–46.

Giroux, H. A. 2010. "Bare pedagogy and the scourge of neoliberalism: Rethinking higher education as a democratic public sphere." *The Educational Forum* 74, 184–96.

Johns, T. 1990. "From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning."*CALL Austria* 10, 14–34.

Leech, G. 1997. "Teaching and language corpora: A convergence." In *Teaching and language corpora*, edited by A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles, 1–23. London: Longman.

Leonardi, V. 2009. "Teaching Business English Through Translation." *Journal of Language & Translation* 10-1 (March 2009): 139–53.

McEnery, T., and A. Wilson. 2001. *Corpus linguistics*. 2nd ed. Edinburgh: Edinburgh University Press.

Mukherjee, J. 2004. "Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany." In *Applied Corpus Linguistics: A multidimensional perspective*, edited by U. Connor and T. Upton, 239–50. Amsterdam: Rodopi.

Olssen, M., and M. A. Peters. 2005."Neoliberalism, higher education and the knowledge economy: From the free market to knowledge capitalism."*Journal of Education Policy* 20, 313–45.

Pym, A. 1996. "Material Text Transfer as a key to the purposes of translation." In *Basic Issues in Translation Studies* edited by A. Neubert, G. Shreve, and K. Gommlich, 337–46. Kent, OH and Leipzig: Kent State University Institute for Applied Linguistics.

Römer, U. 2009. “Corpus research and practice: What help do teachers need and what can we offer?” In *Corpora and Language Teaching*, edited by K. Aijmer, 83–98. Amsterdam: John Benjamins.

Ross, N. J. 2000. “Interference and Intervention: Using Translation in the EFL Classroom.” *Modern English Teacher* 9(3): 61–66.

Rundell, M., and P. Stock. 1992.“The Corpus Revolution.”*English Today* 8, 9–14.

Thomas, J. 1983. “Cross-cultural pragmatic failure.” *Applied Linguistics* 4, 91–112.

Vaughan, Elaine, and Brian Clancy. 2013. “Small corpora and pragmatics.” *The Yearbook of Corpus Linguistics and Pragmatics*1, 53–73.

Widdowson, H. 1990. *Aspects of language teaching*. Oxford: Oxford University Press.

Widdowson, H. 2000. “The limitations of linguistics applied.”*Applied Linguistics* 21(1): 3–25.

Witte, A., and T. Harden. 2009. *Translation in Second Language Learning and Teaching*, Bern: Peter Lang.

Woodhouse, D. 1999. “Quality and quality assurance.” In *Quality and Internationalisation in Higher Education, Programme on Institutional Management in Higher Education*. Paris: OECD.

Chapter Two

Application of Corpus Analysis of Terminology in Language Studies at University Level

Liudmila Mockienė

Special lexis, or more specifically, terminology, is the aspect which to the largest extent makes various specific domains distinct. As LSP studies focus on the language of specific areas or domains, terminology is at its core. Thus, terminology is part of the focus of teaching a foreign language at more advanced levels.

Corpus analysis has been applied to conduct research on vocabulary quite extensively. Indeed, corpus analysis tools can provide a great amount of information on such aspects of lexical items as their frequency and their semantic and syntactic environment. As Rundell (2008, 23) claims, though “human beings tend to notice the unusual”, one should concentrate more on “the most usual, the most frequent, the most typical”. These could include the behaviour of commonly used words, their typical usages, collocations and syntactic patterns—i.e., their “lexical profiles”. Different types of corpus software include a variety of tools which can be used to analyze lexis, such as frequency wordlists, concordance lines, keywords in context (KWIC), term extraction, collocates, colligates, taggers and lemmatizers. The extracted information can be used for all kinds of lexicographic activity, such as to compile term banks, glossaries, dictionaries, terminology databases and translation memory databases. As Zanettin (2002b) notes, there is value not only in using specialized corpora but also in their creation per se. Laurence Anthony, the developer of AntConc freeware—a well-known corpus toolkit—claims that corpora and corpus tools are of great value not only for researchers of languages but also for teachers and learners (Anthony 2009).

Thus, application of corpus analysis tools can be employed in teaching English at university level for corpus compilation, data extraction and further contrastive and linguistic (especially lexical) analysis. It can be given as an assignment in the form of a project or a case study to students who study philology (linguistics) or even those who study English for Specific Purposes as a part of their course assessment. Corpus analysis tools can also be used by students of philology (linguistics) who write their course papers or bachelor or master thesis.

To be able to perform the assignment successfully, students first need to be introduced to the main principles of the application of corpus building and analysis tools for corpus design and data extraction—in this particular case, terminology (in English and their native language), focusing on such aspects as what counts as a corpus, types of corpora, principles of compiling an ad hoc (DIY, disposable, specialized) corpus, corpus balance and representativeness, advantages of corpus creation and analysis tools, and the compilation protocol. Next, they need to be acquainted with the main principles of contrastive analysis of terminology, with attention drawn to such issues as linguistic means of term formation in the analyzed languages and methodology of contrastive analysis of terminology, including principles of selection of terms for the analysis with regard to a particular domain, distinction between a word and a term, distinction between a multi-word term and a free word phrase, and classification of linguistic means of term formation as a common ground for contrastive analysis in two languages. Before students can carry out the research on their chosen domain autonomously, they are given an example of a case study on the application of corpus analysis tools for term extraction and a model for further linguistic analysis of terminology in a particular area.

A case study of analyzing the terminology of constitutional law in English and Lithuanian is given as an example to illustrate the possibility of integrating corpus analysis tools into the process of teaching and learning languages at more advanced levels.

2.1. Corpus building and analysis tools in terminological analysis

What counts as ‘corpus’?

There have been several discussions on what ‘corpus’ is and what actually counts as ‘corpus’, including classification of corpora into certain types according to certain explicit criteria (Atkins, Clear and Ostler 1991; Sinclair 1996; McEnery 2003, 449; Hunston 2008, 154; Anthony 2009).

What can be considered a corpus? Could a single text count as a corpus? What should the number of texts included in a corpus be? What should their lengths be? Should those be whole texts or parts thereof? These major questions are actually not that easy to answer. Apparently, not every collection of texts can be called a corpus. In most cases "corpus" has been defined in terms of composition and/or the functions it performs.

As Sinclair (1996) suggests, there should be certain minimum requirements for any collection of language samples to be referred to as a corpus, and one should distinguish "corpora which record a language in ordinary use from corpora which record more specialised kinds of language behaviour". Here Sinclair drew the line between corpora used for reference of the general language use and special corpora, which reflect only certain features of the language.

The issue of defining a corpus is closely related with the issue of corpus design as such. Sinclair (1996) introduces several principles applied in designing a corpus. First, the size of the corpus should be as large as possible, provided the technology allows that. Next, the corpus should be representative and thus consist of samples from a vast range of material. The corpus as a whole and its individual samples should be divided into genres. The sizes of the samples should be similar. The latter principle is in fact controversial. Finally, the corpus should have a clear origin and source indicated.

Thus, Sinclair (1996) provided his definition of a corpus as follows: "A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language." In this definition Sinclair used a general word "pieces" instead of "texts" because if the corpus compiler uses samples of equal size there is little possibility that they will be whole texts; rather, "most of them will be fragments of texts, arbitrarily detached from their contents". Further on, Sinclair distinguishes between collections and archives as sets of texts which, unlike corpora, are not selected or ordered based on linguistic criteria. Thus, there should be certain minimum criteria, or characteristics, for a set of pieces of language to be called a corpus.

Sinclair (1996) assumes the following four characteristics attributed to a corpus with their default values: quantity, quality, simplicity and documented status. The first feature assumes that the corpus is large enough. The size of the corpus is related to "the ease or difficulty of acquiring the material". This characteristic is rather subjective and depends to a large extent on the decision of the corpus compiler and the overall aim of designing the corpus. The second characteristic—i.e., quality—requires the material to be authentic, without intervention by the linguist. This default value can be altered in the case of compiling a special corpus, which as such does not focus on description of the ordinary language. The third characteristic is termed "simplicity" and is defined by the default value of the text being plain—i.e., without imposing "any additional linguistic information on the text". Finally, by "documented" Sinclair means that "full details about the constituents of a component are kept separately from the component itself".

Considerable attention should be paid to sampling techniques as well. Samples are not texts as such, and they are usually of the same size and small. Thus, Sinclair (1996) proposed to distinguish between a *text corpus* (or a *whole text corpus*) and a *sample corpus*. He argues that "whole text" value should be a default condition. Thus, almost a decade later he defines "corpus" as a "carefully selected collection of texts, involving a great deal of human judgement" (Sinclair 2008, 30).

McEnery (2003, 449) incorporated the notion of sampling into the definition of "corpus" as well:

> "The term corpus should properly only be applied to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of a certain linguistic feature (or a set of features) via the data collected."

Notably, the term "collection of data" is used in the definition rather that the word "text".

Types of corpora

First, as Sinclair (1996) suggests, corpora can be divided into subcorpora, whereas subcorpora can be divided into components, or *sublanguages*. A component "is a collection of pieces of language that are selected and ordered according to a set of linguistic criteria that serve to characterise its linguistic homogeneity" (Sinclair 1996). Thus, sublanguages (or components) provide for a more detailed analysis of such linguistic phenomena as genre and LSPs (languages for special purposes) in particular. There is an increasing need for "access to corpora containing sublanguage material, in order to build systems capable of handling specialised texts" (Sinclair

1996). Subcorpora in this sense can be opposed to reference corpora, which are "designed to provide comprehensive information about a language" (Sinclair 1996). The goal of reference corpora is "to represent all the relevant varieties of a language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials" (Sinclair 1996). As Ostler (2008, 458) states, "In brief, text corpora give evidence in extenso about a language, and about the content that has been expressed in that language." Hunston (2008, 154) also distinguishes between these two types of corpora—"general, reference corpora designed to investigate a given language as a whole, to specialised corpora designed to answer more specific research questions". Zanettin (2011, 15) gives a more detailed explanation of what a specialized corpus can focus on, namely,

> "a specific text type/genre (e.g., fiction, news, academic prose), domain/topic (e.g., biology, social sciences), production method (e.g., learners' language, translated language), or a combination of various defining features (e.g., translated academic medical language)".

One more distinction made by scholars is into monolingual, bilingual and multilingual corpora depending on the number of languages involved. Moreover, bilingual and multilingual corpora can be further divided into parallel and comparable depending on the nature of texts used to compile the corpora. Sinclair (1996) defines a parallel corpus as "a collection of texts, each of which is translated into one or more other languages than the original". Such corpora can include from two to several dozens of languages. According to Sinclair (1996) "a comparable corpus is one which selects similar texts in more than one language or variety". McEnery and Xiao (2007, p. 20) provide a more specific definition of a comparable corpus, which "can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness". However, extensive discussion among scholars focuses on what 'similarity' (or 'comparability') of texts is. This issue will be reviewed in more detail in the next part of this chapter on the principle of compiling corpora.

An interesting subtype of comparable corpora that has not been extensively used apparently due to the time-consuming and complicated process of compilation, but which presents a number of advantages for language analysis, is a comparable multimodal corpus (Ignatova 2018; Navarretta et al. 2011).

Principles of compiling an ad hoc corpus

After students are introduced to the principles of analyzing general (reference) corpora in their linguistics or LSP classes, the task of compiling a specialized corpus seems plausible and expedient for a more specific analysis of particular linguistic features of various specific discourses. To be able to cope with such an assignment, students need to analyze the main principles of compiling a corpus and, in particular, a specialized corpus (*ad hoc* corpus or *DIY* corpus).

As King (2009) notes, there are certain well-established guidelines for designing a corpus; however, they are still developing. The aspects of corpus design usually addressed by scholars at large are size, content, representativeness, balancing, sampling techniques, accessibility of data and the ethical issues of material collection (e.g., Sinclair 1996; Biber, Conrad and Reppen 1998; Kennedy 1998; McEnery and Wilson 2001; McEnery and Gabrielatos 2006; Meyer 2002; Zanettin 2002a, 2000b; Baker 2006; Rundell 2008; Anthony 2009). Some authors have presented and discussed additional or more specific design criteria (Meyer 2002; Losey-Leon 2015).

The quality of the research and the validity of the results will depend to a large extent on the quality and design of the corpus. As Rundell (2008, 23) suggests, the quality of the data one can obtain and use is “dependent upon the corpus itself being large enough and sufficiently well-balanced to be reliable”. Zanettin (2011, 16) also argues that “the composition of a corpus will affect the findings derived from the analysis”. Likewise, “a poorly constructed corpus will inevitably lead to poor results” (Anthony 2009, 90). The decision regarding inclusion or exclusion of a document from a corpus is not that easy as it can affect the composition of the corpus and the distribution of linguistic features (Schäfer, Barbaresi, and Bildhauer 2013).

The indispensable parameters of a corpus such as size (how large the corpus should be) and content (what and how many texts to include) have been discussed by numerous scholars (Sinclair 1996; Biber, Conrad and Reppen 1998; Kennedy 1998; McEnery and Wilson 2001; Meyer 2002; Baker 2006; Rundell 2008; King 2009; Anthony 2009). A corpus compiler should answer the questions “what is the optimum extent of a corpus?” and “how can it be ‘representative’ of the language of which it is a sample?” (Rundell 2008, 25).

First, the parameter of ***size*** has been addressed by nearly every scholar discussing corpus design issues. Even this major and most obvious parameter is not clear-cut but rather controversial. Rundell (2008, 25) suggests that “the arguments for collecting large corpora remain compelling”. King (2009, 302) supports this view and claims that “one must strive to make the corpus as large as possible (in relation to the resources and time available)”. Having the advancement of technologies and the ease of access to most data sources in mind, this task seems quite doable. However, Anthony (2009, 91) surprisingly states the opposite and argues that “even a single text can be considered a corpus if we observe it using the same procedures and tools that we would use to observe more traditional large-scale bodies of texts”. Although even a single text can be considered a corpus, the title of the corpus should be informative and precise enough to reflect the nature of the text used to compile the corpus. Anthony (2009) concludes that the functionality of a corpus does not as such depend on its size but on the software tools applied for the analysis and thus the kind of data that can be extracted from it for the analysis.

Thus, one might infer that the quality of a corpus does not depend on its size per se but on other aspects as well, such as the aim and the users of the corpus. According to Meyer (2002, 30), in cases of general reference corpora, “decisions concerning the composition of a corpus will be determined by the planned uses of the corpus”. Moreover, a corpus compiler should first take into account the available resources to be used, and the amount of time for searching and collecting the materials, making decisions on inclusion of certain texts, computerizing them, etc. (Meyer 2002). After having evaluated the resources at his or her disposal, the researcher should consider the length of the corpus which would enable him or her to carry out the analysis. As Meyer (2002, 33) notes, if the aim of using the created corpus is compiling a dictionary, the researcher would require “a much larger database than is available in shorter corpora”, and Meyer supports the idea that “the lengthier the corpus, the better”. According to Meyer (2002, 33), it is even possible to determine the necessary minimum size of a corpus on the basis of statistical calculations which rely on “frequency with which linguistic constructions are likely to occur in text samples and calculate how large the corpus will have to be to study the distribution of the constructions validly”. However, there is a risk that less frequently occurring linguistic features will not be reflected in the results. Ultimately, the corpus should be large enough to produce reliable results of the most infrequently occurring linguistic features as well. As Biber (1993) suggests, it is also necessary to identify the linguistic features that the researcher will analyze in the corpus, as the distribution of particular linguistic features may vary across texts. The author also suggests compiling a pilot corpus of texts to see whether it is representative of the analyzed linguistic feature.

With specialized corpora one might be interested in analyzing more specific linguistic features of a particular area or domain, thus "a study of a small corpus can be compared with looking at a small group of stars, from which we can gain more detailed information about their unique properties" (Anthony 2009, 91). As Losey-Leon (2015, 296) maintains, at least with regard to specialized corpora there is a general agreement that the final corpus size is defined by the aim of the corpus and "no minimum or maximum extension is particularly required". In the case of specialized corpora the number of specialized relevant texts can be limited, or, as Anthony (2009) suggests, one might be interested in analyzing a single text only. Zanettin (2011) also agrees that the size of a very specialized corpus can be smaller than that of a general one. In the case of ad hoc corpora it is rather easy to add new texts; thus in this sense DIY corpora are disposable.

As the parameter of size is not the most crucial in compiling ad hoc corpora and should only be in conformity with the aim of the corpus, there are other important criteria for a corpus to be well-designed, informative and representative. Representativeness of a corpus can be achieved by two parameters, namely, corpus balance and text sampling.

One of the most important methodological issues which must be addressed by a corpus compiler is ***text sampling***. Some scholars believe that the decision on the length of text samples must be made prior to the stage of collecting material for a corpus (Meyer 2002). Moreover, Meyer (2002, 38) agrees that it is "desirable to include complete texts in corpora". Sinclair (1996) named "whole text" as a default condition for a corpus. For a general reference corpus it might be more beneficial to include a greater number of shorter text samples from a greater variety of sources, genres, text types, and diversity of speakers and writers. For instance, if the aim of the research is to determine the frequency of grammatical constructions, text fragments would be more preferable for inclusion in a corpus than whole texts (Meyer 2002). However, for a specialized corpus it would be more efficient to include full texts rather than fragments because such inclusion "enhances language usage recovery and knowledge from a real collection of texts and increases the range of terminological research aspects as its ultimate ends" (Losey-Leon 2015). Moreover, having in mind that a DIY corpus is usually restricted in scope and addresses a particular narrower area of knowledge, it would be essential to rely on whole texts for the analysis. Only then would it be possible to ensure that the corpus is fully representative of the area analyzed with regard to most frequently and less frequently occurring linguistic features.

Corpus balance is another important consideration to be taken into account while selecting texts for inclusion in a corpus. Corpus balance refers to types of genres included in the corpus. A corpus compiler should decide what types of genres to include and in what proportions. In the case of special-purpose corpora, the focus is usually on a particular genre as such. What the corpus compiler should decide is the type of texts to be included—e.g., text categories or types of documents.

One more aspect related to corpus balance (and representativeness) is the ***number of texts*** to be included in a corpus. Meyer (2002) suggests that there are two perspectives from which to approach this issue: a purely linguistic perspective and the perspective of sampling methodology. The sampling methodology, applied in social sciences, relies on statistical values "to enable researchers to determine how many 'elements' from a 'population' need to be selected to provide a valid representation of the population being studied" (Meyer 2002, 40). In corpus linguistics this technique would be used to establish the minimum number of texts to be included in a corpus to provide an adequate reflection of linguistic features and maintain the necessary level of representativeness. Another approach is analyzing how much internal linguistic variation exists in a genre. Meyer (2002) concludes that, "the number of samples of a genre needed to ensure valid studies of the genre is best determined by how much internal variation exists in the genre: the more variation, the more samples needed". Some scholars (Corpas and Seghiri 2009; Seghiri 2011) have used and validated the computer program ReCor to count the minimum number of texts and even words that have to be included in a specialized corpus for it to be deemed representative. The ReCor program helps to ensure precise evaluation of how representative a corpus is (Corpas and Seghiri 2009). As Losey-Leon (2015, 298) notes, when discussing representativeness of ad hoc corpora which are aimed at terminological and lexicographic analysis, "the corpus representativeness is achieved if the *lexical density* does not alter when more texts are added". The same rule can be applied to representativeness of a corpus with regard to the analysis of other linguistic features: corpus representativeness is achieved if the *occurrence of linguistic features* does not alter when more texts are added.

Apart from these major considerations of compiling a corpus, there are a number of additional or more specific corpus design criteria. The criteria for selecting documents for inclusion in a corpus should also focus on

period coverage, or ***time frame***, when the texts were produced. This parameter depends to a large extent on the aim and the type of the corpus. Thus, synchronic corpora should include text samples produced within a comparatively short time frame to ensure a precise representation of the contemporary language state without being affected by change, though it is arguable whether it is possible to have purely synchronic studies (Meyer 2002). The most changeable level of the language is its lexis. New words constantly appear to name new phenomena; thus, language change is never ending (Meyer 2002). To be able to build a really synchronic corpus, a corpus compiler has to include text samples from a very narrow time span, and Meyer (2002, 46) suggests that "a time-frame of five to ten years seems reasonable". If one aims to compile a diachronic corpus meant to analyze linguistic features from a historic perspective, the time frame is usually determined by the established historical periods or stages of development of a language. The situation might be quite different in the case of special corpora which are created to analyze and extract such data as terminology of a particular area of knowledge, which is the level of language most prone to rapid change and development. New terms are introduced constantly to name new concepts, which appear with the development of a particular area of science or specific domain. Thus, one should take into consideration the stages of development of the analyzed domain in order to determine the most relevant time frame of texts for inclusion in the corpus.

Another issue related to selection of texts for inclusion into a corpus is ***authorship***—i.e., text samples produced by native vs. non-native speakers. In this respect, the issue of using translated texts becomes most debatable. It is quite natural to assume that translation texts are used in translational corpora (i.e., corpora used to analyze the language of translations). Thus, Sánchez-Gijón, Inés and Lonsdale (2009) state that ad hoc corpora designed to be used as translation resources are not only useful as training resources but also can be used by professional translators. However, Zanettin (2011, 14) takes a strong stand on the position that "translations should be included in most corpora, be they used in monolingual corpus linguistics or in corpus-based contrastive linguistics". He argues that translated texts should be included not only in corpora which are designed to analyze the language of translated texts as such but also in general reference corpora. The most common arguments for exclusion of translated texts from corpora rely on "a more or less implicit assumption that they 'corrupt' the reference norm for a language"; however, the author claims that there is actually a lack of justification for this assumption from any theoretical considerations and translated texts actually may "represent a sometimes substantial proportion of all linguistic production in a given culture" (Zanettin 2011, 21). Zanettin

(2011) agrees that this proportion might be different from language to language, or rather culture to culture. If we explicate Zanettin's thought and take as an example international legal documents, we will find out that in most cases either they may be translations from one source language they were drafted in or they may have been drawn up by a team whose members are not necessarily native speakers. As translated texts represent a particular variety of language, they should be included in corpora to ensure representativeness of what is found in the language. So, monolingual corpora should not exclude translated texts because such texts are part of a language produced and received by speakers of that language. Besides, "translations contribute to the creation of the standard norm for a language and should therefore be part of the sampling frame for a corpus aiming to represent that language" (Zanettin 2011, 17). Moreover, as the author claims, translation texts could comprise a translational subcorpus within a larger reference corpus and could be used for comparative analysis with other subcorpora of the same language. Next, both monolingual general reference and monolingual specialized corpora in several languages can be combined to form a bi- or multi-lingual comparable corpus. Inclusion of translation texts into bilingual comparable corpora may provide a practical advantage, as such corpora would include parallel subcorpora which, in turn, reflect links between the languages involved and enable the contrastive and comparative analysis of linguistic features of the languages (Zanettin 2011). By all means, translation texts comprise the basis of corpora which are designed for corpus-based translation studies in order to analyze translation aspects between several languages. To sum up, Zanettin (2011) strongly believes that translated texts should be included in most general reference and specialized corpora, which can be mono-, bi- and multi-lingual. His argument seems logical and convincing and indeed applicable in compiling ad hoc corpora with the aim to analyze linguistic features in specific domains, such as the language of legal acts, court rulings, contracts and agreements.

One more issue related to selection of texts is deciding on the ***source*** of the material to be included in the corpus. At this point the corpus compiler might be restricted by copyright laws and would need to obtain permission to use copyrighted material. Again, this limitation depends on the type of texts and the type of the domain that the intended corpus will cover. For instance, there should be no serious obstacles if one wants to build a corpus of the EU legislation on a particular matter, as it is available online in several languages, which would make it possible to create not only a monolingual but also a bi- or multi-lingual corpus. It is also quite easy to get access to national legislation online or court decisions. Other areas are more restricted, such as getting access to agreements and contracts in the private sector. However, it is possible to cooperate with companies and create a specialized corpus for the internal use of the company, such as compiling glossaries of the most frequently used terminological units and their translation to unify and harmonize the use of terminology throughout the company. With the increasing trend of open access scientific material, it is also quite easy to get access to academic research papers published in scholarly journals.

Text format (or "mark-up") is another necessary consideration for a corpus compiler. In this respect, texts can be simple (non-annotated and untagged) or annotated and tagged. The most common type of annotation is part of speech (POS) tagging, which involves labelling each word in a corpus with a grammatical category (part of speech). POS annotation can be extremely useful in cases of word polysemy (Reppen 2010). Parsing is another type of mark-up, and it is the analysis of syntactic structure. Anthony (2009) sees great value in annotation, claiming that "annotated data provide us with essential information for understanding how linguistic objects operate within texts, and ultimately help us to refine our models of language itself". It should be noted that the process of annotation and parsing is usually time-consuming and requires great human (expert) and material resources.

Other additional considerations to be taken into account while compiling a corpus, which depend entirely on the aim and type of the corpus, are the degree of specialization and sociolinguistic variables, such as gender balance, age, level of education, dialect variation, social contexts and social relationships (Meyer 2002).

On balance, an ad hoc corpus should be long enough to provide reliable results of the analyzed linguistic feature and consist of texts (preferably whole) which are representative of the domain to be analyzed by means of the corpus. Nonetheless, Anthony (2009, 90) maintains that a researcher should not focus on corpus design too heavily and forget that for a corpus analysis to be successful and productive one needs two more components, namely, "(1) human intuition (to interpret the data derived from corpora, and more importantly perhaps, (2) software tools to extract the data in the first place".

Finally, one more aspect of corpus compilation to be discussed is compiling parallel and comparative corpora. As Fantinuoli (2018) notes, domain-specific corpora are very rarely available for use, thus they must be built. At first sight the process of compiling a parallel corpus seems relatively easy. As a parallel corpus by definition is a corpus which consists of texts in the original (or source) language and its translation into other languages (one or more), all what one needs to do is to collect the texts and their translations. However, the problem that researchers often have to deal with in this case is the scarcity of the translated resources, especially for language combination which include languages other than English (Alonso et al. 2012; Delpech et al. 2012; Goeuriot et al. 2009; Morin et al. 2011; Morin & Prochasson 2011; Rivera et al. 2018; Rigouts Terrin et al. 2020). In case of compiling a comparable corpus one needs to select similar texts in two or more languages. The main problem here is to define what makes texts similar or comparable. Thus, researchers often distinguish levels or degrees of comparability, which affect the quality of data extraction. Skadiņa et al. (2010a) suggest four levels of comparability, namely, corpora which are parallel (i.e. composed of accurate or approximate translations), strongly comparable (i.e. closely related texts on the same events or subject), weakly comparable (i.e. texts of the same domain and genre, but on different events or subject), and non-comparable at all (i.e. random texts extracted from large collection of texts). Skadiņa et al. (2010a) also distinguish between preferred, suitable and minimally acceptable comparability.

As the main characteristics for text comparability many researchers first of all mention ‘similar criteria of composition, genre, and topic’ or, in other words, ‘similarity of content, domain and communicative function’ (Zanettin 1998). As Zanettin (1998) points out, the idea of collecting texts in several languages on grounds of similarity of these characteristics for research and training in translation was widespread quite before the era of electronic corpora. The accessibility of resources in electronic format has eased this process. The Web provides almost unlimited and easy access to huge amounts of documents that can be used for compilation of comparable corpora (Deléger & Zweigenbaum 2009). One more aspect mentioned by researchers as a characteristic for comparability is the period the texts were created. However, it is never used alone, but is combined with other characteristics such as domain (Goeuriot et al. 2009; Baroni & Bernardini 2003). These would usually include general language works. For specialized language works, Goeuriot et al. (2009) emphasize the importance of combining the criteria of the same discourse (the domain and the topic) and genre of the texts, which increases the level of their comparability. Deléger & Zweigenbaum (2009) and Bernardini & Ferraresi (2013) also insist that the domain and the genre of the text are crucial for the relevance of the retrieved texts. Alongside such characteristics as domain, genre and time, Morin & Prochasson (2011) also highlight the importance of a more precise criterion such as the register. Besides, the domains should be very specific, otherwise addition of texts which are out of the domain will decrease the quality of the corpus (Delpech et al. 2012). Delpech et al. (2012) also highlight the importance of including texts created in comparable situations of communication. Snover et al. (2008) also add that comparability of documents may be ensured by focusing not only on the same genre, but also on the same event or story, or even a related story. (Biel (2016) makes an observation that it is not possible to have the full comparability of corpora, at least in Translation studies because the genres of texts are culture-specific. Kilgarriff (2010) suggests that word frequency lists (top ten, or top twenty most or least frequent words) can help assess comparability of several corpora.

As for the size of comparable corpora, Zanettin (1998) points out that even several texts could be sufficient if they are written by specialists of the domain and have a high degree of language precision and technicality, though he agrees that ultimately the size of the corpus depends on the criteria of comparability applied to create the corpus. Fantinuoli (2018) notes that a domain-specific corpus should consist of around 80-100 texts.

The process of compiling a comparable corpus does not necessarily have to be manual. This is possible due to introduction of clear criteria for text search and the availability of multilingual documents on the Web. Bernardini & Ferraresi (2013) distinguish manual small ad hoc corpora, automatic large web-derived corpora and semi-automatic corpora. According to Bernardini & Ferraresi (2013) manual small ad hoc corpora are usually more reliable and tailored to a specific goal. However, when the size of a corpus, either semi-automatically or automatically compiled, increases, its reliability decreases. In terms of time and effort, compilation of small ad hoc corpora can be less time-consuming and demanding if the compiler is a domain specialist. Likewise large automatically web-derived corpora, though created in minutes, can demand a considerable amount of time and effort for revision of the results and discarding the irrelevant sources. Moreover, Snover et al. (2008) emphasize that shorter and fully comparable texts should be preferred to longer, but only partially comparable texts. Researchers and language users will start compiling and using comparable corpora only if the balance of time and effort required to compile the corpus will be positive (Bernardini & Ferraresi 2013; Fantinuoli 2018). For instance, Goeuriot et al. (2009) have developed a tool for compiling specialized comparable corpora, which maintains quality comparable to a manually compiled corpus. Alonso et al. (2012) have developed automatically-built specialized comparable corpora by using a special method of comparability statistics. Fantinuoli (2018) discusses a number of tools that can be successfully used for semi-automatic and automatic extraction of specialized texts from the Web, such as BootCat, AntCorGen and SketchEngine and presents a software that provides for compilation of specialized corpora, which is little-demanding in terms of effort. To improve the quality of a bilingual comparable corpus used for lexicon extraction, Li & Gaussier (2010, p. 651) based the comparability measure "on the expectation of finding translation word pairs in the corpus".

Advantages of ad hoc corpora

The advantages of using ad hoc (DIY, specialized) corpora to analyze specific linguistic features are more than evident. As the size of an ad hoc corpus is usually relatively small, the researcher is able to carry out not only a quantitative analysis but also a deeper qualitative analysis of all tokens found. As Koester (2010, 66) notes, "In a small corpus, on the other hand, all occurrences, and not just a random sample, of high frequency items can be examined." Smaller specialized corpora provide the researcher with more insights into the use of lexico-grammatical patterns in particular contexts. Koester (2010, 67) maintains that the relation "between the corpus and the contexts of use is particularly relevant in the fields of English for Specific Purposes (ESP) and English for Academic Purposes (EAP)".

Moreover, depending on the nature of the texts used to compile the corpus, parallel and comparable corpora can provide different benefits for the language analysis. Comparable corpora actually have more advantages than parallel corpora. The main advantage of comparable corpora is the authentic data that they include. Comparable corpora tend to contain "more idiomatic expression that parallel corpora do because the target texts do not bear the influence of the source language" (Delpech et el. 2012). In contrast, parallel corpora can be inconsistent and less representative in this respect, though they do provide insight into equivalence. Moreover, as parallel corpora are almost unavailable for under-resourced languages, compilation and use of comparable corpora is vital for the analysis of such languages (Skadiņa et al. 2010a). Besides, large amounts of sources are available online for many languages and comparable corpora can be built based on much richer and diverse sources produced on a daily basis (Skadiņa et al. 2010b; Skadiņa et al. 2012). Thus, the scarcity of parallel corpora can be compensated by compiling comparable corpora.

The most immediate use of comparable corpora is apparent in the area of lexis. Word frequency lists obtained from comparable corpora can be used in language teaching for a number purposes, such as syllabus design, decisions on certain lexis to be included in textbooks and dictionaries as well as designing tests for non-native speakers (Kilgarriff 2010).

As texts used to compile comparable corpora represent the target culture, the use of comparable corpora can reveal differences in grammatical usage of certain patterns, which in spite of their grammatical correctness can produce unnaturalness of the translated text (Loock 2015).

A great number of research has been conducted on the use of comparable corpora in translation studies and training with a wide range of aspects covered, such as mining for term translations in comparable corpora including analysis of corpora in the translation classroom with the focus of using the corpus to translate, to learn about terminology and content, and to explore texts (Zanettin 1998), collocational differences in monolingual comparable corpora (Baroni & Nernardini 2003), extraction of collocations and their translation equivalents (Sharoff et al. 2006a), solution of problems that human translators find difficult including translation of polysemous lexical items (Sharoff et al. 2006b), extraction of multi-word expressions and their translation equivalents (Sharoff et al. 2009), extraction of lay paraphrases of specialized expression with the focus on paraphrases of nominalizations and neo-classical compounds (Deléger & Zweigenbaum 2009), use of in-domain terms for bilingual lexicon extraction (Ismail & Manandhar 2010), domain-specific bilingual lexicon extraction including application of a two-way translation of context vectors to improve the quality of the retrieved translation variants (Fišer et al. 2011), single-word and multiple-word terminology extraction and terminology mapping (Ştefănescu 2012), development of multilingual lexicon based on collocational networks (Alonso et al. 2012), identifying highly confident word translations from comparable corpora without any prior knowledge (Vulic et al. 2012), application of corpus methodology for descriptive translation studies (Zanettin 2013), extraction of bilingual terminologies in new technical domains and for less widely-spoken languages (Aker et al. 2013), the use of corpora in institutional legal translation (Biel, 2016), extraction of phrasal verb from the comparable English corpus of legal texts (Bilić & Gaspar 2018), cluster equivalence and translator education (Lewandowska-Tomaszczyk & Pęzik 2018), the use of comparable corpora in interpreting practice and training focusing on speech corpora in interpreter training and using domain-specific corpora for confirmation of hypothesis about the linguistic phenomena (Fantinuoli, 2018), a cross-linguistic study of phraseology across specialized genres (Roldán-Riejos & Grabowski 2019), multiword terms and machine extraction focusing on the pointwise mutual information (Potemkin 2019), monolingual and multilingual automatic term extraction based on the gold standard (Rigouts Terrin et al. 2020), multiword term variation such as omissions, changes, and inaccuracies in Eco-Lexicon (León-Araúz, 2020).

Not only comparable corpora can be used in translation studies and research. Postolea & Ghivirigă (2016) analysed using small parallel corpora to develop collocation-centred activities in specialized translation classes focusing on specialized collocations, which are fixed word combinations that do not refer to a concept, but are nonetheless frequently used.

The major difference in parallel and comparable corpora is that the former consist of texts in the source language and in the translation language, whereas the latter consist of texts in the source languages only, thus one should be aware of the possibility of "inconsistencies in parallel corpora, which are then replicated by translators" (Postolea & Ghivirigă 2016, 69). This is why some researchers suggest combining the use comparable and parallel corpora to make the most of the advantages offered by both types (Bernardini 2007, 2011; Biel 2016; Giampieri 2018; Morin & Prochasson 2011).

Corpora have also been extensively used for teaching and research not only in translation studies, but also in the area of English for Academic Purposes. Corpora analysis equip students with general research skills, corpus compilation and analysis skills (Krishnamurthy & Kosem 2007).

Finally, corpora have been widely used in contrastive studies in general. Comparable corpora in particular are viewed "as a useful source for creating translation memories, and bilingual or multilingual terminologies" (Alonso et al. 2012).

In this respect, the use of corpus building and analysis tools is indispensable. Having completed the task of selecting the pieces of language to be included in the corpus, the process of compiling a corpus takes very little time and usually does not cost anything (depending on the software used[2]). As Anthony (2009) notes, ideally corpus linguists should be able to create their own software tools for corpus analysis, or at the very least they have to be involved in this process. He also argues that these tools should be the focus of all corpus research. A corpus as such is merely a reference point comparable to a library. It is the software tools that make

[2] A comprehensive list of tools used in corpus linguistics with descriptions and fees, compiled by Kristin Berberich, Ingo Kleiber et al., is available at https://corpus-analysis.com/.

the corpus analysis possible. Anthony (2009, 91) suggests that we should see “the corpus itself as a tool that reveals to us the mysteries of the universe of language” and thus be more proactive in tools development. It is important to develop new corpus software tools to get more insights into the language. Unfortunately, in reality not many corpus linguists are specialists in computer programming. As Anthony (2009, 104) maintains, the developer of DIY tools can tailor them to his or her specific needs; thus, he concludes that corpus linguists have to “strive to improve our existing tools, and actively push for the creation of new ones”. As he admits himself, a more realistic scenario would be involvement of both corpus linguists and software developers in order to develop new tools.

Compilation protocol

Several authors have suggested following a specific protocol while creating a corpus (Corpas and Seghiri 2009; Seghiri 2011). A protocolized compilation methodology allows the corpus compiler to ensure the representativeness of the corpus. Such a protocol actually indicates the stages of compiling a corpus, namely, searching for material and getting access to it, data download, conversion of the text format and storage of the data.

The first stage focuses on searching for material and getting access to it. Seghiri (2011) suggests that there are two ways to search for information, namely, institutional search and search by keyword. The author claims that institutional search is most productive for extraction of legal documents for several reasons. First, there are a great number of accessible documents on the official websites of most international and national institutions and organizations. Second, the quality of the documents and the degree of reliability are usually very high due to the fact that drafters and editors of the documents are experts in the domains. An institutional search can be carried out in several search sources, such as institutional per se, regulatory and legislative. Legislative information can be found in a number of sources—mainly official websites of institutions and international organizations, databases of legislation, legislation adopted by law-making bodies, governmental agencies, educational institutions, professional networks and associations. For instance, the best source for the European Union legislation is the European Union law portal Eur-Lex. For other types of documents, a keyword search can be more fruitful. A keyword search is performed via generic search engines. Its advantages include ease and speed of searching; however, one needs to define the key words precisely in order to get relevant information and discard the irrelevant search results (i.e., the "noise").

The next stage involves downloading the obtained and selected data. Here one should pay attention to the fact that information on the websites can be presented in different formats; thus, very often downloading of the documents has to be done manually.

After having downloaded the required data, one might need to change the text format to be able to work with the material later on. The most common conversions are from HTML or PDF to Word. Very often the texts have to be converted to plain text (.txt) format.

Finally, the selected and downloaded data has to be organized and stored in the main folder clearly indicating the subject of the corpus and subfolders with respective divisions.

After students have completed the task of compiling their ad hoc corpora in English and their native language and extracted the relevant lexical units for further analysis—a list of terms from a particular domain, they can perform a lexical analysis of the terms. Thus, they need to be introduced to the main principles of contrastive analysis of terminology.

2.2. Background to contrastive analysis of terminology

This part focuses on the analysis of contrastive research of terminology carried out by most prominent scholars worldwide and presents a review of typical linguistic means that are used to form terms analysed by linguists in English and Lithuanian in particular.

2.2.1. An overview of research on contrastive analysis of terminology

The most prominent scholars who analyze issues of terminology science at large, such as principles of term formation, their typology, sources, development of terminology, and specific features of terms, in English, are: Cabré Castellví and Sager (1999); Cabré Castellví, Condamines, and Ibekwe-SanJuan (2007); Kageura (2002, 2012); Sager (1990, 1997, 2004); Rey and Sager (1995); Temmerman (2000); and Daille (2017), among others. The most prominent scholars in this field in Lithuanian include: Gaivenis (2002); Keinys (1980, 2005a, 2012); and Jakaitienė (2010).

Contrastive analysis of terminology of several languages, especially research that focuses on legal terminology in particular, is not extensive.

Lithuanian legal terminology has been analyzed mainly from the standpoint of correct usage of the language and from the comparative historical perspective. For instance, the usage, norms and correctness of legal terms have been analyzed by Paulauskienė (2004), Pečkuvienė (2009, 2013) and Rudaitienė (2008, 2012, 2013). Umbrasas (2010) analyzed Lithuanian legal terminology and its status in Lithuania from 1918–1940, the change of terminology in translations of the civil code and the criminal statutes which were in force during that period.

Some aspects of contrastive analysis of terminology of Lithuanian and other languages, such as equivalence, were analyzed by Marina (2006) and Kontutytė (2008).

Synchronic contrastive analysis of criminal law terms has been conducted by Rackevičienė (2006, 2008) and by Janulevičienė and Rackevičienė (2009, 2010, 2014).

Recently there has been quite extensive research conducted on the influence of translations of the European Union legislation on the Lithuanian legal and administrative language, as Lithuanian legislators directly rely on such legislation when drafting legal acts (Auksoriūtė 2009).

The most prominent foreign scholars who conduct contrastive analysis of terminology of different languages (English, French and German) and problems of translation of terms, especially legal terms, are Sandrini (1996, 1999), who focuses on the issue of equivalence of legal systems and translation of legal terms; Mattila (2006, 2012), who analyzes the concept of legal language, legal terminology and legal English, legal French, and legal German; and de Groot and van Laer (2006, 2011), who pay a lot of attention to the issue of semantic analysis of legal terms, translation and equivalence.

2.2.2. Research on term formation in English and Lithuanian

The research on term formation in English and Lithuanian was already discussed by the author in her PhD thesis Mockienė, L. 2016, *Formation of Terminology of Constitutional Law in English, Lithuanian and Russian*.

This section focuses on a review of typical linguistic means that are used to form terms analysed by linguists in English and Lithuanian. It will present an overview of different classifications of ways and means of term formation used in English and Lithuanian. General principles and methods of term formation discussed in works in the area of terminology in English, such as in works by Sager and the International Standard ISO 704, and in Lithuanian, such as in works by Keinys and Gaivenis, have already been briefly presented by Mockienė and Rackevičienė (2016). However, this chapter presents several approaches to classification of ways and means of term formation used in special languages in English and Lithuanian in more detail, with the emphasis on the legal and administrative language.

The process of term formation is closely related to the process of word formation in the language in general. However, the process of term formation, according to Sager, is a deliberate, "conscious human activity and differs from the arbitrariness of general word formation processes by its greater awareness of pre-existing patterns and models…" (Sager 1997, 25). This means that not only does this process rely on existing lexical elements and combine them in particular ways, but it can also be described in terms of patterns according to which these elements are combined, which in turn can be used for subsequent term formation (ibid).

Keinys also shares the point of view that terms are created and standardized consciously (Keinys 1980, 60). He also admits that terms comprise a distinct part of the literary language. This is why the literary language is characterized by both general trends in the language development and by peculiar features and specific requirements and development (Keinys 2005g, 231).

Linguistic means of term formation in English

This part discusses several approaches to classification of ways and means of term formation used in special languages in English. First is a) a classification proposed by Sager, a well-known terminologist; second is b) a universal classification presented in the International Standard ISO 704 (Terminology work—Principles and Methods) of the International Organization of Standardization (ISO 2000), which was applied by Valeontis and Mantzari in their contrastive research of English and Greek terminology; and third is c) a classification of ways and means of term formation in legal English, in particular discussed by Mattila, Professor Emeritus of Legal Linguistics at the University of Lapland, Finland, who conducts studies on legal languages, comparative law and comparative legal linguistics.

a) Sager (Sager 2004, 1924) claims that terms used in special languages are made up of the same range of morphological structures as words of the general language. However, the specialized vocabulary "exhibits far greater regularity as a result of the deliberate and often systematic techniques of term creation" (ibid).

So let us have a closer look at the ways and means used to form terms in English. First of all, it should be said that although there are several classifications of ways and means of term formation used in English, all of them are primarily based on the distinction between the use of the existing forms and creating new forms (relying on internal sources) and the use of new resources (relying on external sources).

Sager proposed a classification of the main ways and means of term formation in 1990 in his book *A Practical Course in Terminology Processing* (Sager 1990, 71–80) and in 1997 in the chapter on "Term formation" in the *Handbook of Terminology Management* (Sager 1997), which was later slightly modified by him and discussed in the chapter on "Terminology in special languages" in *An International Handbook on Inflection and Word Formation* (Sager 2004, 1924–28), where he refers to means of term formation as linguistic methods of designation and uses slightly different terminology when referring to those means. The means of term formation discussed by Sager in these works apply to special English languages in general—i.e., a variety of specialised subject domains. Besides, Sager claims that the description of the means of term formation he discusses "is not intended to be exhaustive, but is rather indicative of the range of possibilities".

The three methods of term formation proposed by Sager are:

- the use of existing sources;
- the modification of existing sources;
- the use of new resources (to create new lexical entities).

By "the use of existing resources" Sager means extension of the meaning of a word which already exists in the general language. This can be achieved by means of a simile (naming a concept in analogy to another, familiar one), a metaphor (naming a concept by referring to the thing it most resembles) or the use of a proper name.

Another method of term formation is the modification of existing sources, which, according to Sager, includes such means as affixation (or derivation), backformation, compounding, creating phrasal terms, conversion and compression.

He states that affixation (i.e., suffixation and/or prefixation) is a very important means of term formation as it contributes significantly to the systematic structuring of terms due to the precise expression and systematic reference of affixes. He also claims that the variety of affixes used for term formation in special languages is far greater than in general English because English has borrowed and assimilated a lot of words, word elements and affixes from neoclassical languages, such as Greek and Latin, especially in the domain of science and technology.

Another means of term formation, which is also significant for systematicity of specialized vocabulary, is compounding. A compound is formed by means of combining two or more words into a new syntagmatic unit, which has a new meaning independent of the constituent parts and as a term represents a concept relevant to a particular subject field (Sager 1997, 34). If a compound consists of two elements, the first element, the determinant, usually modifies the second element, the nucleus. However, compounds, as he claims, can be made up of not only two, three or four elements but also five and six, although these compounds are rather rare. He also mentions that there are *phrasal compounds*, which are linked by prepositions (Sager 2004, 1927) and *compounds of phrases* containing prepositions, articles, conjunctions and adverbs (Sager 1990, 74). In another source (Sager 1997, 30, 36), he uses the term *phrasal terms* and discusses them as a separate category, though he admits that their formation is closely related to compounding. However, the distinction between compounds and phrasal terms is not clear-cut.

Conversion occurs when a word changes its category without morphological alteration of the word inflection. As Sager claims, nouns are frequently formed by conversion from verbs and adjectives and vice versa; however, it is not always possible to determine the direction of this process. Additionally, in scientific English the productivity of this means of term formation is reduced due to the fact that a high proportion of terms are derived from Latin and Greek word elements, which have noun endings that are unsuitable for conversion (Sager 2004, 1927).

Special languages are also characterized by terms created by various forms of compression of existing long terms. The most frequent and highly productive means of compression are acronymy, abbreviation and clipping.

The next means of term formation, based on the modification of existing sources, also discussed by Sager (2004), is backformation. He claims that backformation is used mainly in the domain of technology rather than science to form complex verbs which refer to nominal concepts of processes and is often combined together with compounding.

The last term formation method is the use of new resources or creation of new lexical entities (neologisms), which can be of two types: creation of totally new entities and borrowing from other languages (direct borrowing and loan translation). In science and technology, this process results from the need for the unique naming of new concepts. Creation of totally new entities is very rare in special language because new terms should reflect the relationships between new concepts and existing ones. In other words, creation of new terms should be systematic, which can be perfectly achieved by means of affixation and compounding, as mentioned above. Sager also claims that in English it is often difficult to distinguish between the creation of genuine neologisms by means of derivation and borrowing of terms from Latin, Greek and French directly. Besides, the source of borrowings is not always clear because English has had "such a long tradition of borrowing from all three languages that it is very often impossible to say whether a word has come into English via French or whether it has been taken directly from one of the classical languages" (Sager 1990, 38). Moreover, in modern English borrowing from other languages is rather infrequent. It is usually other languages that borrow new technology and new terminology from English. The other form of borrowing, loan translation or calque, is the result of literal translation—word-for-word substitution of the lexical components of compounds. According to Sager, "loan translation is preferred to direct borrowing, but neither form of term creation is acceptable if it violates the natural word formation techniques of a linguistic community" (Sager 1990, 87). This means that borrowings, either as direct ones or loan translations, have to be adapted to the requirements of the receiving language, and this process is rather smooth in English. He also notes that in time loan translations might be "replaced by more appropriate autochthonous forms in order to exploit the creative potential of the language" (Sager 1997, 39).

Ultimately, based on Sager's description of the main ways and means of term formation in special languages in English, the following characteristic features of typical linguistic means used for term formation in technical and scientific English can be distinguished:

- affixation is an extremely important means of term formation in English;
- the variety of affixes used in term formation is greater than in word formation in general English;
- the main source of borrowing affixes and stems for English special vocabulary is Latin and Greek;
- the first element of a compound is usually a determinant, while the second is the nucleus;
- although most term formation means are characteristic of special languages in general, some of them are more characteristic of a particular subject area, such as technology or science;
- conversion, which is strongly characteristic of general English, is reduced in scientific English;
- the distinction between genuine neologisms formed by means of derivation and direct borrowing often is not clear;
- in the case of borrowing from Latin, Greek and French into English, it is not always clear whether the term was borrowed directly from a neoclassical language or came into English via French.

When talking about the use of existing sources, no distinction is made as to whether a term is simple or a formation—i.e., whether the word which was terminologized or transterminologized was simple in its structure or was a derivative. The problem here lies in the fact that in word formation cases it is very difficult to say whether words were created on the basis of existing sources as new formations or were terminologized or transterminologized (i.e., transferred from general language or another subject domain).

b) Another classification of term formation ways and means is presented in International Standard ISO 704 (Terminology work—Principles and Methods) of the International Organization of Standardization (ISO 2000). This classification is relied upon by the *Guidelines for Terminology Policies: formulating and implementing terminology policy in language communities*, prepared by Infoterm, the International Information Centre for Terminology (Infoterm 2005, 9–11), and it is discussed at length by Valeontis and Mantzari, who applied this classification not only to the analysis of English but also to the Greek language and for contrastive study of English and Greek terms. As they claim, these means of term formation are applied in English and are also appropriate to be used in other languages (Valeontis and Mantzari 2006).

The ISO 704 standard specifies that the guidelines it provides neither cover all possible means used for English term formation nor are they intended to be universal and applied to all languages, because means of term formation "depend on the lexical, morphosyntactic, and phonological structures of individual languages, language-specific principles of term formation" and should be "described in national and regional standards dealing with a particular language rather than in International Standards" (ISO 2000).

The main methods of term formation discussed in ISO 704 are:

- creating new forms;
- using existing forms;
- translingual borrowing.

It is noteworthy that the above classifications apply to special languages in general—that is, they describe ways and means of term formation that can be followed when creating terms in any area of science and technology. By all means, they can differ from one area to another and from one language to another.

c) There is a classification of means of term formation used in legal language in particular. Mattila, in his book *Comparative legal linguistics: Language of law, Latin and modern lingua francas*, discusses the following methods of formation of vocabulary of legal language (Mattila 2012, 145–47):

- a word already in existence in ordinary language, or in the language of another specialism, obtains a specialized or broader meaning;
- a neologism of national origin is created;
- a word is borrowed from a foreign language (or from another national language).

It is evident that the first two ways of forming terms, namely using existing forms and creating new forms on the basis of existing forms, are based on internal sources, whereas the last, borrowing from a foreign language, is based on external sources.

Mattila (2012, 145) states that in comparison with other languages for special purposes, legal language contains more words which are used in ordinary language. However, such words have a precisely defined or even distinct meaning which distinguishes them from words of ordinary language.

The next method of term formation is called “neologisms of national origin”, which includes creating entirely new words, derivation of new words on the basis of words already in existence, and forming compound words and phrases.

Creating entirely new words is not characteristic of legal language, which is consistent with the same statement about other languages for special purposes. However, according to Mattila, acronyms (or initializations) are quite common in legal language.

Term formation on the basis of words already in existence is very frequent, and the number of abstract derivatives is particularly high in legal language because law and legal science are complex, abstract phenomena.

Next, formation of compound words and phrases is a very productive means of creating legal neologisms, although the author emphasizes the difference between languages: compound words are more typical of such languages as the Scandinavian group of languages, Finnish and German, whereas phrases are more characteristic of such languages as English and French.

Mattila distinguishes loanwords as a separate category and emphasizes not only the technical aspect of this complex phenomenon but also the ideological (Mattila 2006, 147).

The comparison between the three classifications presented above reveals that categorization of the main ways and means of term formation in English is very similar. The main criterion for classification is the opposition of internal and external sources of terms: the use or modification of existing (internal) sources to create new terms is opposed to the use of external sources (translingual borrowing).

Yet, there are some differences in these classifications. First of all, Sager views conversion as a modification of existing sources, alongside such means of term formation as affixation and compounding; however, ISO 740:2000(E) standards attribute conversion to the use of existing forms, alongside terminologization and transterminologization. This difference is quite consistent with the different conceptions of conversion in the theory of word-building. Some linguists consider conversion to be a subtype of derivation (Plag 2003, 17); others discuss it as a separate category distinct from derivation and compounding (Jackson and Zé Amvela 2012, 100; Šeškauskienė 2013, 123).

Additionally, there is also some difference in the interpretation of neologisms and their place in the classification. Sager attributes borrowings from other languages to the category of neologisms (Sager 1997, 38), whereas Valeontis and Mantzari, who rely on ISO standards, claim that according to the new definition of the term "neologism" by ISO/TC37 [1087-1:2000], only newly coined terms, either simple or complex, which appear in a language for the first time and have been created by means of linguistic mechanisms such as derivation, compounding or blending, can be considered neologisms. Thus, borrowing from foreign languages cannot be attributed to the category of means of creating neologisms (Valeontis and Mantzari 2006). Mattila also specifies the national origin of neologisms, which indicates that borrowings are not attributed to this category. The distinction between national and foreign origin of the sources of term formation is very important.

Linguistic means of term formation in Lithuanian

The following part presents approaches to the classification of ways and means of term formation used in special languages in Lithuanian: a) first, a universal classification of ways and means of formation of special terminology proposed by such famous scholars as Gaivenis and Keinys is discussed; b) second, a classification presented by Akelaitis, who analyzed types and sources of terms of administrative language, is presented; and c) finally, a classification of sources and means of term formation in legal Lithuanian in particular used by the prominent terminologist Umbrasas, who analyzed legal terminology used in the period from 1918 to 1940, is reviewed.

a) In the Lithuanian terminology science, three means of term formation are usually distinguished: 1) using existing vocabulary of standard language and dialects; 2) creating new words (neologisms); and 3) borrowing terms from other languages (Gaivenis 2002; Keinys 2005a). The first two means of term formation rely on internal sources, whereas the third is based on the use of external sources.

The first means of term formation is based on the use of vocabulary of native origin. Using vocabulary of native origin means that either a word, which has already existed in the language or its dialect, is terminologized, or a term, which has already existed in another terminological field, is transterminologized (Keinys 2005g, 231). A word from the general language or its dialect can be terminologized in two ways: its lexical meaning is either extended or narrowed (Gaivenis 2002, 52–53). The content, valency and use of terms created by this means of formation usually changes. Words of a general language or its dialect used as terms usually acquire new semes which they do not have in the general language (ibid). Most one-word terms, which are simple in structure, are terminologized words of a general language or its dialect. However, as Gaivenis (2002, 56) claims, there are cases when terms whose structures are complex (i.e., they are formations) cannot be attributed to the category of terms created by means of word formation because in fact they were terminologized—i.e., the word as such was formed in the general language (or its dialect) and then it was used as a term.

Very often, it is extremely difficult to establish whether a term is a result of the process of word formation or terminologization. This might depend on the field of science and the history of the development of terminology of that particular field. However, the criterion of "newness" of a term is not that reliable, as new words can be formed by the same means at different periods and in different locations where the language is spoken. Even the fact that a word existed in the past and was used in some historic document does not signify that the word was taken from that old source and not created as a new entity (Keinys 2005d). Keinys (Keinys 2005d) believes that all terms that can be synchronically viewed as formations should be treated as a result of word formation rather than terminologization, although not all of them were created for special purposes as terms. In most cases it is impossible to establish whether a term was created as a neologism or was taken from the general language. Besides, terms which were formed on the basis of ordinary words nevertheless reflect certain word formation types; thus, it would be inaccurate not to consider them as representatives of those types of word formation (Keinys 2005d, 22; Keinys 2005e, 113). Umbrasas also supports this approach and does not distinguish between terms formed for specific purposes and terminologized words; he treats them formally and attributes all terms of complex structure (derivatives and compounds) to one category (Umbrasas 2010, 67–68). The general requirement for the terminologization process is to form a term that has only one meaning, even if it is formed on the basis of a polysemous word of the general language (Akelaitis 2009, 57–58).

The next method of term formation in Lithuanian is creating new terms in accordance with all main types of word formation on the basis of the existing words. The four main means of term formation in Lithuanian are prefixation, suffixation, inflection and word compounding (Gaivenis 2002, 54; Keinys 2005g, 232; EC 2006).

Keinys, who analyzed term formation means in different fields of science, claims that suffixation is the main means of term formation and that most such terms are derived on the basis of verbs. Other characteristic features of terms formed by means of suffixation include a large number of abstract nouns, a lack of emotional connotation, and a large number of hybrids (Keinys 2005d). Keinys analyzed hybrids—i.e., terms which consist of a foreign base and native suffix—as derivatives and attributed them to the same group as derived terms of native origin (ibid). However, not all linguists treat hybrids in this way; Umbrasas distinguishes hybrids, as a separate category, distinct from terms of native origin formed by means of suffixation, and discusses such means of term formation as a subtype of term formation based on elements of foreign origin (Umbrasas 2010, 132).

Another very productive means of term formation in Lithuanian is compounding. In fact, compounding as a means of term formation is quite popular in certain areas of professional language and is the second most productive means of term formation after suffixation (Keinys 2005c, 129). Its productivity in creating terminology can be explained by the necessity of naming a complex concept. Additionally, such compounds conform to the vitally important requirements of accuracy, clarity and conciseness. In many cases this can be ensured by using a multi-word term; however, one-word terms are more convenient and allow terminologists to create terms that are relatively short generic names that help avoid using attributives. The majority of compounds in terminology are formed on the basis of two nouns, a noun and a verb, and an adjective and a noun (Gaivenis 2002; Keinys 2005c).

One more quite productive means of term formation in Lithuanian is inflection. It is noteworthy that unlike in English, in Lithuanian inflection serves not only as a functional affix but also as a derivational affix. Inflection as a means of word formation is very close to suffixation in nature, derivational meaning and form (Keinys 2005b). Due to their simple structure, terms formed by means of inflection are very convenient. This means of term formation could be used more extensively (ibid). The majority of terms formed by inflection are derived from verbs and adjectives, and very rarely from nouns, pronouns and numerals.

Prefixation as a means of term formation is used relatively seldom in comparison with other means of derivation because nouns can be formed by means of prefixation only on the basis of other nouns (Keinys 2005e), whereas suffixation and inflection are used to form nouns primarily on the basis of verbs, adjectives and nouns. Nevertheless, the role of prefixation as a means of term formation is quite significant (ibid).

Finally, the last method of term formation, which is usually distinguished in Lithuanian terminology science, is borrowing terms from other languages. Borrowing of a concept together with the term which signifies that concept is quite common in terminology science and language in general (Keinys 2005g). Gaivenis states that it is difficult to avoid borrowings in terminology, and that there is no necessity to do that (Gaivenis 2002, 57). What is important is that borrowings should not supersede the existing terms of Lithuanian origin and should comply with the rules of phonology, morphology and spelling. According to the degree of assimilation, borrowed terms are usually divided into three groups: 1) old borrowings; 2) international words; and 3) barbarisms. Old borrowings are words that have been completely assimilated and adapted to the language system. They are not considered borrowings in terminology. Next, international words for the most part are terms of different fields of science. They have come into Lithuanian mainly from Greek and Latin directly and through intermediary languages. Lastly, barbarisms are words which do not conform to the norms of the language (ibid). They do not become part of the vocabulary. Finally, it should be noted that the number of borrowings is distributed unevenly in different fields of terminology. It is believed that there are a lot more borrowings in the latest scientific and technical branches of terminology and fewer in fields which have long traditions and rely largely on the vernacular vocabulary (Keinys 2005f).

In essence, Lithuanian scholars, such as Gaivenis and Keinys, believe that the main source of term formation should be internal—i.e., the native language, either its general vocabulary or its dialects, or the native means of word-building. Thus, the main means of term formation are terminologization of words of general language or its dialect and word formation. These semantic and morphological means are the basis of term formation in Lithuanian.

b) Akelaitis analyzes terms of the administrative language (Akelaitis, 2008, 2009). He notes that sources of administrative terms are the same as sources of Lithuanian terms in general (Akelaitis 2009). Akelaitis explicitly bases his classification on the sources of terminology: internal and external, and then he discusses their subtypes in more detail.

Internal sources of terms are used for formation of terms by means of one language; in this case, Lithuanian. Folk (or inherited) terms do not automatically become terms of a particular branch of science. They, as well as words of the standard language, are terminologized—i.e., their meaning is expanded, narrowed or changed in some other way. Terminization is quite a complicated process. A terminologized word might acquire an absolutely new meaning. Next, the meaning of a word might be expanded (i.e., the word retains the main components of its meaning and acquires some additional components). Third, a word might acquire a meaning which is narrowed (it does not acquire any new components of meaning but becomes more abstract). Finally, a word might acquire a figurative meaning. Such cases are rare in the administrative language because a word which becomes a term by means of getting a figurative meaning acquires expressive components, which is not recommended in term formation. Akelaitis also notes that several methods of terminologization can be combined at the same time. Term formation includes neologisms (which are completely new words) and formations (which are the result of derivation and compounding).

Means of term formation based on external sources include borrowing of terms (international terms, old borrowings and new borrowings) and translation (either morphological or semantic).

Akelaitis discusses different criteria for classification of terms, such as their form, scope, content and grammatical relation. According to the form, terms are classified into one-word and multi-word terms. Akelaitis pays a lot of attention to the structure of multi-word terms. He claims that classification of terms according to the structural models is more informative than classification according to the type of syntactic relation. Besides, classification according to the type of syntactic relation is possible with two-word terms but is very complicated, very hard to apply and not informative with terms of more than three words (Akelaitis 2009, 54). Within the structure of multi-word terms, he suggests distinguishing the position of the head of the multi-word term (which can be a single word or a phrase) and its dependents—i.e., the pre-head position and the postposition.

c) Umbrasas (2010), who analyzed legal Lithuanian terminology used in the period from 1918 to 1940, classified terms based on their structure, origin and means of word formation. Umbrasas applied the usual classification of terms proposed by Gaivenis to terms formed on the basis of words which exist in the native general vocabulary, creating new words on the basis of elements of native origin, and borrowing of terms from other languages.

First of all, Umbrasas classifies terms according to their structure into one-word and multi-word terms. Then he classifies one-word terms on the basis of their origin into terms of native origin (terminologized terms simple in structure and terms formed by means of derivation, compounding and conversion), and terms of foreign origin (borrowings, hybrids and barbarisms). He attributes all terms of native origin which are simple in structure to the category of terms based on terminologization of the existing words of general vocabulary, whereas all terms which are formed by means of word-building, irrespective of whether they have been terminologized or transterminologized, are attributed to the category of terms formed by means of word formation (derivation, compounding and conversion). He admits that it is almost impossible to establish whether a term has been formed by means of word formation specifically in a particular field of science or in the general language and then was terminologized. He uses a fairly formal approach and classifies terms into simple terms and formations on the basis of the means of linguistic expression rather than the way a word appeared in the language. Additionally, Umbrasas discusses several cases of conversion which were found in legal Lithuanian as a separate type of word formation, alongside derivation and compounding.

Multi-word terms are classified according to the number of constituent words (two-word terms, three-word terms and terms formed of four and more words). Next, Umbrasas analyzes the origin of the constituent words and their syntactic relations within the term structure. Within the groups of terms which consist of three and more constituent words he distinguishes between terms which contain a preposition and those which do not. His analysis also includes the aspect of the position of the nucleus (the determined word) within a multi-word term.

On the basis of Umbrasas' description of the means of term formation used in civil and criminal codes of Lithuania from 1918–1940, the following characteristic features of typical linguistic means used for term formation in legal Lithuanian can be distinguished:

- multi-word terms in legal terminology prevail over one-word terms;
- the majority of one-word terms come from internal sources, most of which are terminologized or new formations and only some of which are terminologized simple words;
- more than three fourths of formations are formed by means of suffixation;
- prefixation and compounding are used quite infrequently to form legal terms;

- legal terms formed by means of inflexion comprise a very small number of the total terms;
- only about one fifth of legal terms are borrowings, most of which are international words of Latin origin;
- hybrids are quite rare among legal terms;
- the majority of multi-word terms consist of two words;
- in most cases two-word terms are composed of words of native origin, in some cases they are composed of one word of native origin and a borrowing, and cases where terms are composed of two borrowings or hybrids are extremely rare;
- three-word terms constitute only about one fifth of multi-word terms.

The comparison between the three classifications presented above reveals that Lithuanian terminologists classify the main ways and means of term formation in a similar way. The main criterion for classification is the opposition of internal and external sources of terms: the use of existing vocabulary of the standard language or dialects and creating new words (internal sources), as opposed to the use of borrowings (external sources). Additionally, in Lithuanian terminology science it is normal to classify terms according to their structure into one-word and multi-word terms. However, there are a few minor differences in the presented classifications, such as different attitudes towards the position of old borrowings. For instance, Akelaitis considers old borrowings to be among the terms which come from external sources; however, Umbrasas maintains that old borrowings can be attributed to internal sources on the grounds that they have been totally assimilated into Lithuanian and are not perceived as borrowings by native speakers.

2.3. Research methodology of contrastive analysis of terminology

The first section of the part on the research methodology presents the principles of selection of terms for the analysis, including such aspects as principles of selection of terms for the analysis with regard to a particular domain, the distinction between a word and a term, and the distinction between a multi-word term and a free word phrase. The second section, meanwhile, focuses on the principles of the analysis of the collected data, such as classification of linguistic means of term formation as a common ground for comparative analysis in two languages.

2.3.1. Principles of selection of terms for the analysis

During the process of selection of terms for the analysis, two major types of problems can be encountered.

The first type of problem is related to the linguistic aspect—i.e., the distinction between a word and a term. Constitutional law is closely interrelated with many spheres of social life; thus, there are cases when the same word can be used as a term in particular domain or a word of a general vocabulary (e.g., family). So, one of the tasks related to this issue is to define whether a particular word is a term or belongs to the general vocabulary. At this point it is expedient to discuss what a term is and the difference between a term and a word. There are numerous definitions of "term" proposed by different scholars. Keinys defines a term as a word or a word phrase of a particular area which has a definite meaning; terms are names of concepts of science, technology, art, production and other special areas (Keinys 1980, 13). Keinys states that a term differs from a word due to the fact that it has a definite meaning and strictly defined area of use (Keinys 1980, 14). Additionally, most terms are terms only when used in a specific area; moreover, they form a system of terms in that particular area (Keinys 1980, 22–23). Gaivenis notes that a term differs from a word not because of one feature but because of a whole set of features (2002, 13). None of the features of a term—such as the nature of the concept, clearly defined meaning, specificity of the concept, having no synonyms or having only one meaning—taken separately mean that a word is a term (Gaivenis 2002, 13Gaivenis also notes that terms cannot be separated from their definitions (Gaivenis 2002, 14).

In the present research, during the selection process the following major aspects were taken into account: a term expresses a concept of an area of law, a term has a definition, a term belongs to a system of terms, or a term is fixed in a dictionary. Thus, monolingual and bilingual dictionaries of different kinds were used to establish whether a word or a combination of words is a legal term.

Another aspect which complicated the selection of data is related to the distinction between a multi-word term and a free word phrase. There is no doubt that a term can be expressed through either one word or a word phrase. As Keinys claims, terms of most areas are multi-word terms (Keinys 1980, 17). It is not always clear whether a word phrase is a multi-word term or a free word phrase, or a combination of distinct terms, especially if it consists of numerous constituent words. Umbrasas (Umbrasas 2010, 6) states that the longer the term is, the higher the chance that it is a combination of several distinct terms. One of the methods that can be used to establish whether a word phrase is a multi-word term, as suggested by Gaivenis, is to apply a statistical criterion, which can be used to establish whether word phrases are constant, which is an important distinguishing feature of multi-word terms (Gaivenis 2002, 14). For instance, the word phrase *teisė dalyvauti valdant savo šalį* ("a right to participate in governing one's own state") is used in the Constitution of the Republic of Lithuania and is also found in the Lithuanian translation of the Universal Declaration of Human Rights. By all means, this is not the only criterion that can indicate whether a word phrase is a term; however, in some cases it is quite handy.

Another aspect related to the problem of distinguishing terms is the relationship between terminology and nomenclature. In English, lists of terms about a particular subject are often referred to as both "terminology" and "nomenclature", without making a distinction between the two. Lithuanian linguists, such as Auksoriūtė and Umbrasas, consider nomenclature to be a subsystem of terminology and analyze it together with terms (Umbrasas 2010, 14).

Yet another aspect related to the problem of distinguishing terms is deciding what parts of speech can be considered terms. In Lithuanian terminology science, traditionally only nouns and nominal phrases are viewed as terms. However, alongside nouns Umbrasas includes a small number of verbs in his research on legal terminology as well. *A Glossary of Constitutional Terms* (Cottrell and Dhungel, 2007) in English contains not only nouns but quite a number of adjectives and verbs. Only nouns were included in the current research as they perform a nominative function.

After the principles of defining a term are established, the next problem to deal with in the process of selection of data for the analysis, which is "subject" related, is classification of law into areas and branches and attributing a term to the category related to constitutional law. The problem of the subjectivity of selection of legal terms and classification thereof according to branches of law has been already discussed by Umbrasas (Umbrasas 2010, 6–7).

It is not easy to select terms related to constitutional law because it is not always clear whether a term belongs exclusively to the area of constitutional law or to another branch of law as well (for instance, the term *family* can belong to the area of constitutional law or family law; the term *government* can be used extensively in the area of constitutional law as well as administrative law). Similar to the classification of terms into general scientific ones and ones specific to a particular area of science, it is possible to classify legal terms into general legal terms, terms specific to a particular area of law or non-legal technical terminology (Mattila 2012, 4–5). Thus, a term might belong to several branches of law or even all branches of law (e.g., *a law*). As far as constitutional law is concerned, according to the different aspects of classification of branches of law discussed by Vansevičius, constitutional law is integrating and fundamental; it is the core of the legal system, which means that this area of law encompasses other branches of law and is closely related to them (Vansevičius 2000, 151–52). Thus, there might be no strict limits when attributing a term to constitutional law or to other branches of law.

Beinoravičius, Pogožilskaja and Vainiutė note that recently there have been changes in the perception of constitutional law as such. They claim that, "The previously prevalent conception of constitutional law as one of the branches of law has been gradually replaced by the perception of constitutional law as not merely a branch of law, but rather the law of the Constitution, whereas the Constitution is viewed not just as an act (or one of the most important acts), but as a specific area of law, which comes into foreground among other laws and differs from them in many aspects" (Beinoravičius, Pogožilskaja, and Vainiutė 2013). This means that the constitution is now perceived as the primary and ultimate law, which is the nucleus of the whole legal system. This view of the constitution as the central part of the legal system means that it is an act that "integrates the whole legal system, directs the legal regulation and determines its content" (ibid). During the process of selection of the data, with the aim to analyze as many terms as possible related to the system of government and state and government structure, its institutions and main institutes, the relations between the citizens and the state, constitutional law can be interpreted as a wide area of law which integrates other branches of law as well. Thus, all legal terms that are found in the constitutional legal acts have to be included in the analysis.

Dictionaries of various kinds and other sources should be used where possible to establish whether a selected term is a legal term or, even more precisely, a term of constitutional law. In English, dictionaries of law were used, such as *Collins Dictionary of Law* (Stewart 2001) and *Dictionary of Law* (Collin 2004), as were legal writings and other sources. In Lithuanian, the sources used were a dictionary of concepts used in the legislation of the Republic of Lithuania by Mockevičius (Mockevičius 2002) and English-

Lithuanian law dictionaries by Armalytė and Pažūsis (Armalytė and Pažūsis 1998) and Bitinaitė (Bitinaitė 2008).

Finally, one more aspect has to be taken into account when selecting terms, which is determining the legal system to which the term belongs. Lithuanian is used within one country and one legal system; thus, it has no varieties of its legal language as such. However, English is used throughout the world in a number of English-speaking countries with distinct legal systems, such as the United Kingdom, Ireland, the United States of America, Australia, New Zealand and other Commonwealth countries. Although the legal systems of these countries are to a large extent based on the Anglo-Saxon law and share similar features, there are also significant differences, which reflect the unique legal traditions developed in particular English-speaking countries and manifest themselves in different legal concepts or different terminology used to express the same concepts. Mattila discusses in detail this idea of different legal Englishes (2006, 240–54). He gives examples of how American legal English differs from the legal English used in the United Kingdom: first, there are examples of differences in the system of concepts, such as a different court system, which, as a result, produces different court names; second, there are instances of expressing the same concept by means of different terms, such as *corporate law* in the United States and *company law* in England used to refer to the law of companies (Mattila 2006, 243–44). Thus, it is necessary to define the variant of legal English that is analyzed in this research. Terms used in the British legal tradition were chosen for analysis as they represent the primary original Anglo-Saxon legal system.

2.3.2. Classification of linguistic means of term formation for comparative analysis

The comparison between term formation means in English and Lithuanian reveals that scholars use similar criteria for classification—i.e., origin (using or modifying the existing forms as opposed to using external sources) and structure (one-word terms as opposed to multi-word terms); however, the sequence of applying those criteria might produce different classifications. The English scholars discussed above base their classifications on the combination of several criteria (e.g., opposing native and foreign origin, semantic and morphological means of term formation) and try to produce one classification. The usual problem of classifying terms is that it is not always clear which category a term should be attributed to (e.g., a compound or a phrasal term, a borrowing or a neologism derived on the basis of foreign elements).

The idea that terms should be first classified according to their structure into one-word and multi-word terms, and only then should each category be further classified into other types based on the origin of the terms, is applicable in Lithuanian. Additionally, in Lithuanian it is also common to analyze one-word terms according to their structure into simple terms and formations, whereas multi-word terms are analyzed according to the number of constituent words, their origin and their syntactic relation within the structure of the term.

For the purposes of this study we suggested classification of ways and means of term formation on the basis of several criteria. First of all, the terms were classified according to the number of constituent words and then these groups were further classified according to their sources and structure. Thus, a distinction is made between one-word and multi-word terms.

One-word terms were first classified into terms formed on the basis of internal, external or a combination of both sources. One-word terms which come from internal sources were grouped into terminologized simple words (terminologized words of general vocabulary) and terminologized or newly created formations (terms created by means of word formation, such as derivation, compounding and conversion). All formations in this research were analyzed formally and were viewed as newly created, although some of them might be terminologized words of the general vocabulary. In the case of the legal language, which contains a lot of words which are also used in the general vocabulary, it is often very difficult to establish whether a word was created for legal purposes or was terminologized or transterminologized. As this issue is very problematic in the legal language, terminologization was not addressed in detail in the present research, as the main aim was to establish means of word formation. Umbrasas took a similar approach in his research on legal terminology where he did not distinguish terminologized formations and newly created formations but rather analyzed them together according to the category of word formation and the formant (Umbrasas 2010, 73–74). The word formation analysis applied in this research is synchronic—i.e., the fact of whether a word is derived or not is established on the basis of its current ties with other words in the language and not on the basis of its etymology (Keinys 1999; Urbutis 1978). Terms formed by means of conversion are also attributed to the category of words which come from internal sources. It should be noted that the preference for a particular means of term formation in different languages depends on the language structure and traditions of term formation. Next, one-word terms formed on the basis of external sources are borrowings from foreign languages. The origins of terms which come from external sources in this research were established according to the primary source of the borrowing. Most of the terms which come from external sources are international words. In lexicology, it is common to classify international terms according to the primary source rather than the immediate source of borrowing (Jakaitienė 2010, 232; Umbrasas 2010, 124). For the purpose of the present research an etymological analysis should be carried out, though synchronic analysis of borrowings is also possible and would reveal structural (derivational) relations between the analyzed words and other words currently existing in the language. Such research is beyond the scope of the current thesis; however, it has been carried out and the results have been published in several scientific papers (Mockienė and Rackevičienė 2014, 2015). Finally, terms formed on the basis of internal and external sources were attributed to the category of hybrids.

Similar criteria were used to analyze multi-word terms. First, terms were classified according to the number of constituent words into two-word, three-word and multi-word terms. Next, they were classified according to the sources of their constituent words (internal and external) into terms of native origin (consisting exclusively of native words), terms of foreign origin (consisting of foreign words and/or hybrids) and terms which are multi-word hybrids (consisting of one or more native words and one or more foreign words or hybrids). Finally, they were classified according to the structural models of formation. It should be noted that the analysis of the origin of constituent words of multi-word terms, such as adjectives and participles, posed some problems as the process of distinction between a hybrid and a borrowing is not clear. In the case of nouns, the situation is quite transparent, as the word is usually borrowed in its original form with minor morphological adaptation. In Lithuanian and Russian, an inflection is usually added to adapt the word to the grammatical structure of the language. In this case, an inflection does not perform any derivational function. Its function is purely grammatical. However, in the case of adjectives and participles, the situation is more complicated. In Lithuanian and Russian, not only an inflection but also a native suffix is usually added to form (or adapt) an adjective or participle based on foreign sources. In English, a similar problem occurs with participles when native suffixes "–ing" and "–ed" are added to a foreign base. In such cases it can be difficult to decide whether a word is a hybrid (with a native suffix added to a foreign base) or a borrowing (which was simply adapted to the grammatical structure of the language by means of adding a suffix). Umbrasas discusses this problem in detail (Umbrasas 2010, 164–66). In the present research, this issue is not that relevant since each term was attributed to a certain category on the basis of the criterion of an external source, regardless of whether it is a hybrid or a borrowing.

In the present research, the concept of internal sources embraces vocabulary of various types of a native language: a general native language used by all speakers (including standard variety and regional dialects), as well as special native languages used by specialists of certain areas. From the historical perspective, they also include old borrowings which come from the same protolanguage and have been totally assimilated into the language, and thus synchronically are viewed as native vocabulary because their "foreignness" is not perceptible any longer.

The concept of external sources embraces vocabulary of foreign languages. The main means of term formation on the basis of vocabularies of foreign languages is borrowing of terminology. As it has been noted by numerous authors, in many languages, the majority of borrowed terms are the so-called neoclassical borrowings—international words of Latin and Greek origin (Cabré Castellví and Sager 1999, 88–89). Synchronic analysis of borrowings according to the type of word formation was not carried out in the present research. It is possible in English; however, it is rather problematic in Lithuanian and Russian. For instance, in English it is common to see linguists giving borrowings as examples of words formed by derivation; for example, O'Grady, Dobrovolsky and Aronoff give such examples of derivatives formed by affixation as *assert-ion* and *protect-ion* (1997, 145). However, historically these words came into English as borrowings together with the suffix "-ion". It was not added to the root in English. *Assertion* came either from the Middle French *assertion* or directly from the Late Latin *assertionem* (nominative *assertio*), whereas *protection* came from the Old French *proteccion* and directly from the Late Latin *protectionem* (nominative *protectio*). Nevertheless, synchronically they are viewed as derivatives in English. The same applies to the words *protestant* and *defendant* discussed by Šeškauskienė as examples of suffixation. From the historical point of view, both of them are borrowings: *protestant* came from the German or French *protestant* and from the Latin *protestantem*, whereas *defendant* came from the French *defendant*. Thus, it can be seen that in English synchronic analysis of borrowings according to the type of derivation is widely accepted. The reason for that might be the fact that after English had borrowed numerous words from French and Latin, they were used as models for analogous formations from bases of native origin, and such French suffixes as *-ment* became very productive (Durkin 2011, 98–99).

A combination of internal and external sources (morphemes and words of native and foreign origin) produces hybrids, which take an intermediary position in the given classification. Usually hybrids are formed by means of combining a foreign base and a native affix. However, they can also be formed by means of combining a native base and a foreign affix. A combination of several roots of different origin may produce hybrids which are compounds consisting of a native root and a foreign root (Celiešienė and Džežulskienė 2009, 67, 78; Jakaitienė 2010, 211, 255–56; Umbrasas 2010, 132).

2.3.3. Terminology of constitutional law: a case study

After students are introduced to the main corpus analysis tools and the principles of terminology extraction and linguistic analysis, and before they can carry out their own research on the chosen specific area or domain, they are given an example of a case study on one particular area, namely, constitutional law, in two languages—English and Lithuanian.

The research data for this case study was collected from the primary sources of constitutional law. In the Republic of Lithuania, the primary source of constitutional law is the constitution, which is codified and in the form of a single written document. The Lithuanian terms were collected from the Constitution of the Republic of Lithuania (1992). The sources of constitutional law for the UK are different because of the peculiar nature of the UK constitution. It differs not only from Lithuania or Russia but also from the majority of countries in the world. The UK constitution is not codified and consists of numerous legal acts of a constitutional nature and other sources. The main written sources that are considered to be the basis of the UK constitution are the acts of Parliament, judicial decisions, parliamentary constitutional conventions, the Royal Prerogative and other constitutional sources (Blick 2012). For the purposes of the present research, the following major legal acts of a constitutional nature were chosen: translations into modern English of the Magna Carta (1297), the Habeas Corpus Act (1679), the Bill of Rights (1689) and the Act of Settlement (1700), including the amendments as in force today, and the original texts of the Parliament Act (1949), the Human Rights Act (1998), the House of Lords Act (1999) and the Fixed-term Parliaments Act (2011).

Though only one document was chosen for the analysis in Lithuanian and several in English, their volume is comparable. Moreover, as Anthony (2009) claims, “Even a single text can be considered a corpus if we observe it using the same procedures and tools that we would use to observe more traditional large-scale bodies of texts.” The documents in each language were analyzed by means of the AntConc corpus software tools to extract lexical units and form the list of terms for further analysis. After extracting lexical units from the corpora in the two languages, the lists of terms were reviewed for “noise” and inclusion into the shortlist for further analysis on grounds of relevance for the area of constitutional law. In the English material, 660 terms were found; in the Lithuanian material, 626 terms.

The linguistic analysis of the selected terms first focused on counting the ratio of one-word terms and multi-word terms in Lithuanian and English and comparing them. Different languages give preference to different criteria of term formation. One of them is the length of a term. A concept might be expressed by a single word or a combination of several words with or without a preposition. Terms have a function of not only denoting a concept but also reflecting its meaning; this is why many terms actually consist of several words. In the case of one-word terms, which are more complex than root words, this function is carried out by word-building morphemes.

In English, the ratio of the selected one-word and multi-word terms is as follows: out of 660 terms found in legal acts of a constitutional nature in English, 378 are one-word terms (57%), whereas 282 are multi-word terms (43%). The ratio between one-word and multi-word terms in English is almost equal—i.e., concepts used in legal acts of a constitutional nature in English tend to be expressed through both one-word and multi-word terms, with a dominance of one-word terms. The majority of multi-word terms in English consist of two words (78.5% of multi-word terms). Three-word terms constitute about 13.7% of multi-word terms, whereas terms composed of four or more words constitute 7.8% of multi-word terms (see Figure 1).

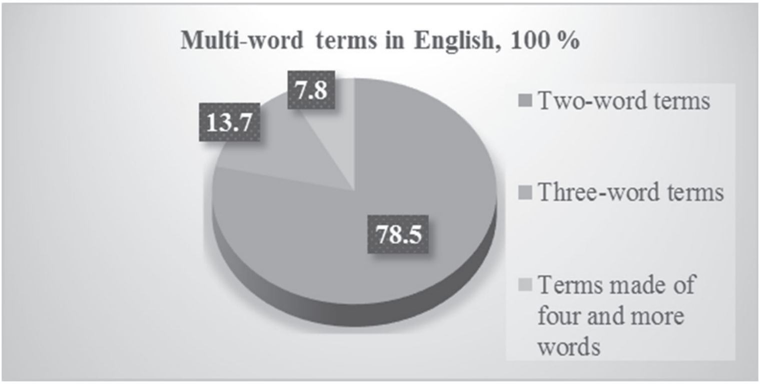


Figure 2-1. Multi-word terms in legal acts of a constitutional nature in English, out of 100%

In Lithuanian, the ratio of the analyzed one-word and multi-word terms is as follows: out of 626 terms found in legal acts of a constitutional nature in Lithuanian, 241 are one-word terms (38.5%), whereas 385 are multi-word terms (61.5%). The ratio between one-word and multi-word terms shows that concepts used in the legal acts of a constitutional nature in Lithuanian tend to be expressed through both one-word and multi-word terms, with a dominance of multi-word terms. The majority of multi-word terms consist of two words (74.5% of multi-word terms). Three-word terms constitute about 17.2% of multi-word terms, whereas terms composed of four or more words constitute 8.3% of multi-word terms (see Figure 2).

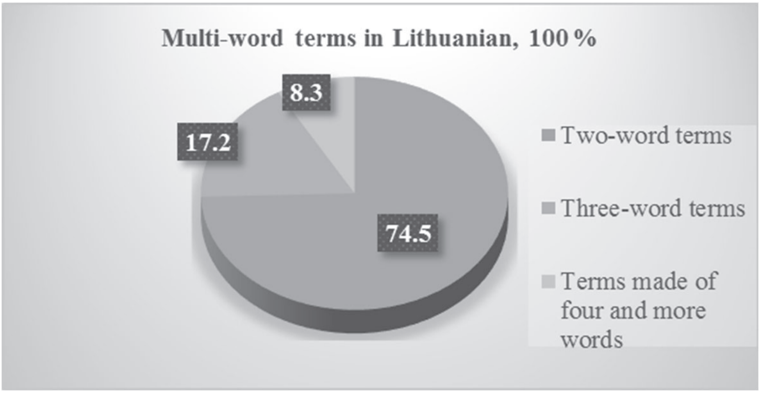


Figure 2-2. Multi-word terms in legal acts of a constitutional nature in Lithuanian, out of 100%

In conclusion, the comparison between the ratio of one-word and multi-word terms in English and Lithuanian (Figure 3) reveals that concepts used in legal acts of a constitutional nature in these languages tend to be expressed by means of both one-word and multi-word terms. In English, one-word terms dominate over multi-word terms. However, Lithuanian gives more preference for multi-word terms. The majority of multi-word terms in both languages are two-word terms.

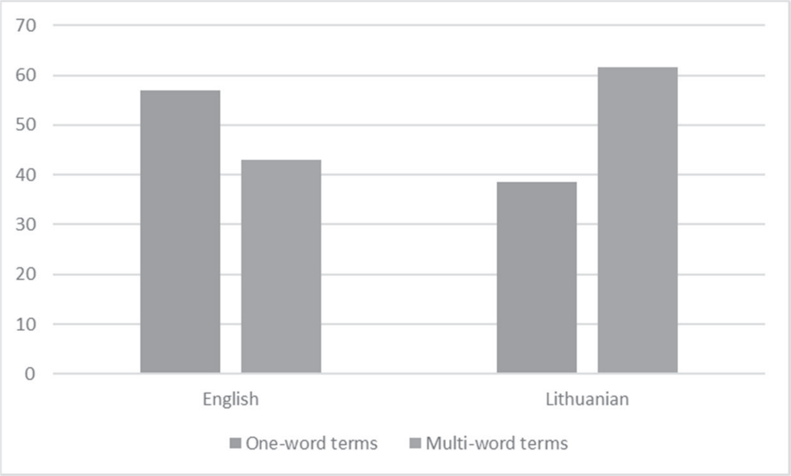


Figure 2-3. One-word and multi-word terms in legal acts of a constitutional nature in English and Lithuanian, out of 100%

Another very important aspect in terminology science to be analyzed is the origin or source of the lexical units used to form a term.

French and Latin have had a great impact on the development of the English legal terminology. In the Middle Ages, the predominant language used to speak in the courts of Medieval England was French, while French and Latin were used to write legal documents. One of the reasons why French and Latin were used in law instead of English could be "the urge to have a secret language and to preserve a professional monopoly" and thus to set the legal profession apart from the rest of the society (Maley 1994). It was only from the end of the 15th century that statutes were printed in English. Nevertheless, English became the official language of law only in the 17th century and gradually took over legal French and legal Latin (Jackson and Zé Amvela 2012, 44–45; Maley 1994). The contacts English had with French and Latin have had a great impact on the formation of the English legal terminology as well as its characteristic features.

Terms found in legal acts of a constitutional nature in English come from internal and external sources, with the great prevalence of the latter. The group of English terms formed on the basis of internal sources consists of words of native origin which are simple in structure (consist of a root) and were formed by means of terminologization (i.e., transformation into terms designating legal concepts), and formations, where all components (the root or roots and affixes) are of native origin. In English, this group of terms includes, first of all, words of Anglo-Saxon origin (Algeo 2010; Jackson and Zé Amvela 2012, 37; Van Gelderen 2006, 300). The Anglo-Saxons came to Britain from a variety of Germanic tribes and the language they spoke is often described as a dialect of Germanic (Van Gelderen 2006, 351). Before the languages of the Germanic branch became differentiated, they had been known as Germanic or Proto-Germanic (Baugh and Cable 2002, 28–29). The languages which descended from Proto-Germanic fall into three groups: East Germanic, North Germanic and West Germanic. North Germanic developed into the modern Scandinavians languages. In its earlier form the common Scandinavian language is referred to as Old Norse. West Germanic is the ancestor of modern English, German, Dutch and Frisian (Baugh and Cable 2002, 28–29; Jackson and Zé Amvela 2012, 25). As was mentioned above, old borrowings which belong to the same protolanguage and have been totally assimilated into the language are also attributed to words which come from internal sources. Additionally, Jackson and Zé Amvela state that due to the close interaction between Old Norse and Old English, "many Scandinavian words resemble their English cognates so closely that it would be impossible to tell whether a given word was Scandinavian or English" (Jackson and Zé Amvela 2012, 42). Algeo also notes that, "Many Old English words of Germanic origin were identical, or at least highly similar, in both form and meaning to the corresponding Modern English words" (Algeo 2010, 90). Moreover, McIntyre also notes that, "The base form of a word in English was often remarkably similar to the base form of the same word in Scandinavian" (McIntyre 2010, 52). Baugh and Cable also observed this similarity between Old English and the language of the Scandinavian invaders and the subsequent difficulty of deciding whether a word is native or borrowed (Baugh and Cable 2002, 87). Thus, words which come from Proto-Germanic are considered to be words which belong to internal sources.

Terms found in legal acts of a constitutional nature formed on the basis of internal sources in English comprise a relatively small group of 52 terms

(14% within the group of one-word terms)—e.g., *bye-law, body, bond, borough, draft, earl, highness, king, kingdom, land, law, lawfulness, leave, oath, poll, queen, right, seat, sheriff, thing, town, witness, work, writ*.

Although borrowing was uncommon in Old English (almost all of it was Germanic, except for about 3% of borrowings, mainly from Latin (Jackson and Zé Amvela 2012, 29; McIntyre 2010, 132)), the number of borrowings in modern English is quite high, being over 70%. As McIntyre states, "The rich vocabulary of English is a result of the extent to which it has borrowed from other languages during the course of its history" (McIntyre 2010, 91). By all means the ratio of native and foreign vocabulary depends on the type of text, its style and scope. Jackson and Zé Amvela suggest that, "Formal style and specialized language use a greater proportion of foreign loans than does everyday conversation" (Jackson and Zé Amvela 2012, 53–54).

The majority of one-word terms found in legal acts of a constitutional nature in English come from external sources and form a group of 314 terms (83%). Most of them are Romance loans predominantly from Latin and French—e.g., *act* (< Latin *actus*); *alien* (< French *aliene, alien* or < Latin *aliēnus*); *applicant* (< Latin *applicantem)*; *compensation* (< Latin *compensātiōn-em*); *concurrence* (< Latin *concurrentia) confidence* (< Latin *confīdentia*); *conviction* (< Latin *convictiōn-em*); *custody* (< Latin *custōdia*); *delegation* (< Latin *dēlēgātiōn-em*); *deportation* (< Latin *dēportātiōn-em*); *discrimination* (< Latin *discrimination-*); *emergency* (< Latin *ēmergentia*); *formality* (< Latin *formālitās, < formālis*); *habeas corpus* (< Latin *habeas corpus*); *injunction* (< Latin *injunctiōn-em*); *instigation* (< Latin *instigātiōn-em*); *juvenile* (< Latin *juvenīlis*); *legislation* (< Latin *lēgislātiōn-em*); *recess* (< Latin *recessus*); *recommendation* (< Latin *recommendation-*); *rejection* (< Latin *rēiectiōn-, rēiectiō*); *remedy* (< French *remède* < Latin *remedium*); *respondent* (< Latin *respondent-, respondens*); *security* (< Latin securitas); *service* (< French *service* < Latin *servitium*, < *servus*); *session* (< French *session* < Latin *sessiōnem* (*sessio*)); *status* (< Latin *statūtum*); *tribunal* (< Latin *tribūnāl, tribūnāle*); *vacancy* (< Latin *vacantia*); *validity* (< Latin *validitas, < validus*); *victim* (< Latin *victima*)—and Greek loans—e.g., *abbot* (< Latin *abbat-*, *abbas* < Byzantine Greek *ἀββάς*); *archbishop* (< Latin *archiepiscopum* < Greek *arkhi-ἐπίσκοπος*); *bishop* (< Latin *episcopus* < Greek *ἐπίσκοπος*); *scheme* (< Latin *schēma* < Greek *σχῆμα*). A further 3% of terms are formed on the basis of combining morphemes of native and foreign origin—e.g., *by-election* (a native prefix *by-* + a base of Latin origin which came into English through the French *election*).

The development of the legal Lithuanian language and professional language of lawyers has a direct connection with the restoration of independence of Lithuania in 1918 (Maksimaitis 2014). Before that, lawyers and administrators used Latin or Slavic to draw up official legal documents. Lithuanian lawyers studied in Latin and Polish (ibid). The Lithuanian legal terminology started to be formed and used in the areas of state governance, politics, economics and other public-life dimensions only after the restoration of independence in 1918 (Umbrasas 2010, 16). The period of 1918–1940 is the most crucial in the history of legal Lithuanian because only after 1918 when Lithuanian became the state language were legal acts published in Lithuanian and the state started to devote considerable attention to legal terminology (Umbrasas 2010, 265). It is also noteworthy that despite the fact that Slavic languages had a significant influence on the Lithuanian legal language at the beginning of its development, borrowings from these languages were step-by-step replaced by Lithuanian equivalents and terms of Latin and Greek origin, borrowed either directly or through intermediary languages. M. Maksimaitis draws a conclusion that during the two decades of the Independence from 1918–1940 a solid foundation for the development of the contemporary Lithuanian legal terminology was established (Maksimaitis 2014). During the later period of 1945–1990 when Lithuania was part of the Soviet Union, general Soviet legal standards were applied to form Lithuanian legal terminology. After the restoration of independence in 1990, Lithuanian again gained its position as the state language and has been used in all spheres of life.

Terms found in the Constitution of the Republic of Lithuania come from internal and external sources, with the great prevalence of the former. Terms formed on the basis of internal sources in Lithuanian, as well as in English, include terminologized words of native origin which are either simple in structure or are formations, which consists of components (the root or roots and affixes) of native origin. This group of terms includes, first of all, native words, words inherited from the Indo-European protolanguage and old borrowings which have been totally assimilated into the language and are not perceived as foreign by native speakers. In Lithuanian, unlike in English, terms formed on the basis of internal sources comprise a very large group of 184 terms (76%)—e.g., *karas* ‘war’; *kraštas* ‘region’; *lytis* ‘sex’; *narys* ‘member’; *straipsnis* ‘article’; *šeima* ‘family’; *tauta* ‘nation’; *valia* ‘will’; *žemė* ‘land’; *žmogus* ‘man’; *įgaliojimai* ‘credentials, powers’; *įgyvendinimas* ‘implementation’; *įsigaliojimas* ‘coming into force’; *įsitikinimai* ‘convictions’; *įstatymas* ‘law’; *kurstymas* ‘incitement’; *nepasitikėjimas* ‘no confidence’; *nesijungimas* ‘non-alignment’; *nusikaltimas* ‘offence, crime’;

nutarimas ‘resolution’; *gynyba* ‘defense’; *taryba* ‘council’; *valdyba* ‘the board’; *rinkėjas* ‘elector’; *teisėjas* ‘judge’; *tautybė* ‘nationality’; *valstybė* ‘the state’; *vyriausybė* ‘government’; *viršenybė* ‘supremacy’.

Only about one fourth of the terms found in the Constitution of the Republic of Lithuania come from external sources, and they form a group of 54 terms (22.4%). All of them are international words. Most of them are of Latin origin—e.g., *aktas* ‘act’ (< Latin *actus*); *asociacija* ‘association’ (< Latin *associatio*); *cenzūra* ‘censure’ (< Latin *censeo*); *dekretas* ‘decree’ (< Latin *decretum*); *deputatas* ‘deputy’ (< Latin *deputatus*); *diskriminacija* ‘discrimination’ (< Latin *discriminatio*); *funkcija* ‘function’ (< Latin *functio*); *institucija* ‘institution’ (< Latin *institutio*); *integracija* ‘integration’ (< Latin *integratio*); *interpeliacija* ‘interpellation’ (< Latin *interpellatio*); *kandidatas* ‘candidate’ (< Latin *candidatus*); *kompetencija* ‘competence’ (< Latin *competentia*); *konstitucija* ‘constitution’ (< Latin *constitutio*); *kultūra* ‘culture’ (< Latin *cultura*); *mandatas* ‘mandate’ (< Latin *mandatam*); *ministerija* ‘ministry’ (< Latin *ministerium*); *ministras* ‘minister’ (< Latin *minister*); *plebiscitas* ‘plebiscite’ (< Latin *plebiscitum*); *pozicija* ‘position’ (< Latin *positio*); *prezidentas* ‘president’ (< Latin *praesidens*); *privilegija* ‘privilege’ (< Latin *privilegium*)—and Greek origin—e.g., *amnestija* ‘amnesty’ (←Greek *amnēstia*); *autonomija* ‘autonomy’ (←Greek *autonomia*); *demokratija* ‘democracy’ (←Greek *dēmokratia*); *kanonas* ‘canon’ (←Greek *kanōn*); *kritika* ‘critics’ (←Greek *kritikē*); *policija* ‘police’ (through German *Polizei* ←Greek *politeia*); *programa* ‘programme’ (←Greek *programma*). These terms entered Lithuanian through or under the influence of intermediary languages, predominantly Western European languages, such as French, German, English and Italian. Unlike English hybrids, which are formed by a combination of either a native base and a foreign affix or a foreign base and a native affix, in Lithuanian the analyzed hybrids are of only one type: they are formed by attaching a native affix to a foreign base. In fact, a combination of a foreign affix and a native base is not productive in Lithuanian. Hybrids constitute only 1.7% of all Lithuanian terms, most of which were formed using bases of Latin or French origin. All analyzed hybrids were formed by combining a base of foreign origin (a verb or an adjective) + a native suffix: *disponavimas* ‘disposal’ ← the verb *disponuoti* ‘to dispose’ (< a base of Latin origin *disponere*) + Lith. suffix *–imas.*

To sum up the main findings of the case study, by means of corpus analysis tools a similar number of terms were extracted in English and Lithuanian from the documents of a constitutional nature. In English, one-word terms dominate over multi-word terms. However, Lithuanian gives more preference for multi-word terms. In the group of multi-word terms, two-word terms dominate in both languages.

The ratio of internal and external sources of one-word terms in English and Lithuanian reveals that the majority of one-word terms in English are borrowings, whereas the Lithuanian language tends to use the resources of the native language as much as possible. The number of hybrids—i.e., terms consisting of morphemes of native and foreign origin—is considerably small in both languages.

Although in Lithuanian multi-word terms prevail over one-word terms, in both analyzed languages developers of multi-word terms adhere to the principle of language economy and tend to create terms which consist of no more than two words.

Nearly a half of multi-word terms in Lithuanian are composed of words which come from internal sources, others are multi-word hybrids, and a very small number of multi-word terms are composed of words which come from external sources only. In English, in contrast to Lithuanian, nearly half of multi-word terms are composed of words which come from external sources, others are multi-word hybrids, and a very small number of multi-word terms are composed of words which come from internal sources only.

In comparison with Lithuanian, English is more open to borrowing from other languages; Lithuanian tends to preserve the national language and make maximum use of the internal resources to create terms, either by terminologizing words of the standard language and dialects or by applying word-building means characteristic of those languages. This does not mean that Lithuanian avoids borrowings; however, borrowings are used quite sparingly.

In comparison with Lithuanian, English terminology is more user-friendly and meets the criteria of language economy and derivability because the majority of constitutional law terms are one-word terms, most of which are simple in structure. However, the fact that the majority of Lithuanian terms are either derivatives or multi-word terms—i.e., are of a more complex structure—means that in both languages the criterion of precision is more important. Multi-word terms not only name a concept; they also to some extent reveal its content and tend to resemble the main features of the concept as fully as possible. In one-word terms, this function can be performed by the means of derivation. The majority of multi-word terms in both analyzed languages are two-word terms. Thus, it is possible to claim that developers of terms in both analyzed languages adhere to the principle of language economy and try to create terms composed of not more than two words. Managing to combine these criteria together when creating a term would produce a term of the optimal length: short enough to be user-friendly and long enough to express the concept as fully as possible.

Contrastive analysis of terminology reveals certain differences in traditions of term formation in different languages and provides a deeper insight into the languages. When translating terms into other languages, translators should also pay attention to the tendencies of the term formation means of the target language and use means and models of term formation characteristic of the target language.

A similar analysis can be carried with the constituent words of the extracted multi-word terms. It is also possible to carry out a more detailed analysis of the selected terms with regard to the structure of both one-word and multi-word terms. One-word terms can be analyzed into simple (root words) and complex (derivative words using different means of word formation). Multi-word terms can be analyzed with regard to their syntactic structure, focusing on the position of the head of the term and the dependents, parts of speech, typical prepositional constructions, most frequent models of term formation and the like.

2.4. Conclusions

By applying corpora building and analysis tools, students of more advanced levels can compile their own ad hoc corpora of a particular domain for term extraction and further linguistic (lexical) analysis as part of their training in languages. Such assignments can be given in the form of a project or case study to students who study philology (linguistics) and have to write course papers and BA and MA theses or even those who study ESP as a part of their course assessment.

The value of such assignments lies not only in the contrastive analysis of linguistic features of several languages but also in the process of building specialized corpora per se. In order to perform the task successfully, students need to conduct lots of individual research and use their analytical thinking, make decisions, learn to select the relevant and discard the irrelevant material, manage the data, time and resources at hand, and be collaborative and creative.

A more focused and detailed linguistic analysis and contrastive study of the linguistic features of several languages give students deeper insights into how languages function and what trends and tendencies exist with regard to term formation.

A great number of other aspects can be analyzed further by means of corpora building and analysis tools, such as collocations of the chosen terms, their syntactic patterns, their derivatives, their frequency and distribution, and even their conceptual meanings. All this research data could contribute to lexicographic, terminographic and translation activities.

References

Akelaitis, G. 2008. “Sudėtinių administracinių terminų struktūriniai modeliai.” *Specialybės Kalba: Terminija ir studijos: mokslinių straipsnių rinkinys*, 5–12.

Akelaitis, G. 2009. “Administracinės kalbos terminija ir kita specialioji leksika.” *Specialybės kalba: Administracinės kalbos vadovėlis*, 36–86. Vilnius: Mykolo Romerio universiteto Leidybos centras.

Aker, A., Paramita, M. L., & Gaizauskas, R. 2013. Extracting bilingual terminologies from comparable corpora. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 402-411.

Algeo, J. 2010. *The Origins and Development of the English Language*. 6th ed. Wadsworth. Cengage Learning: Boston, USA.

Alonso, A., Blancafort, H., de Groc, C., Million, C., & Williams, G. 2012. METRICC: Harnessing Comparable Corpora for Multilingual Lexicon Development. In 15th EURALEX International Congress, 389-403.

Anthony, L. 2009. “Issues in the Design and Development of Software Tools for Corpus Studies: The Case for Collaboration.” In *Contemporary Corpus Linguistics*, edited by P. Baker. 87-104, London, UK: Continuum Press.

Armalytė, O., and L. Pažūsis. 1998. *Anglų-lietuvių kalbų teisės žodynas = English-Lithuanian Law Dictionary*. Vilnius: Alma littera.

Atkins, S., J. Clear, and N. Ostler. 1991. “Corpus Design Criteria.” *Literary & Linguistic Computing* 7 (1): 1–16.

Auksoriūtė, A. 2009. “ES teisės aktų vertimų įtaka tolesnei lietuvių teisinės kalbos raidai.” In *Lietuvių kalba Europos Sąjungos erdvėje: Tarptautinės konferencijos medžiaga, 2008 m. gruodžio 8 d*, 41–49. Vilnius: Valstybės žinios.

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

Baroni, M., & Bernardini, S. 2003. A preliminary analysis of collocational differences in monolingual comparable corpora. In *D. Archer, P. Rayson, A. Wilson and A. McEnery (eds.) Proceedings of Corpus Linguistics 2003*, 366-383, Lancaster: UCREL, Lancaster University

Baugh, A. C., and T. Cable. 2002. *A History of the English Language*. 5th ed. London: Routledge.

Beinoravičius, D., L. Pogožilskaja, and M. Vainiutė. 2013. "Peculiarities of Formal Structure of Terms Denominating Concepts of Human Rights and Freedoms in Lithuanian, German and English." *Адміністративне Право і Процес (АПП): Науково-Практичний Журнал. 2 (4), 249-258*

Bernardini, S. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. In *SYNAPS - A Journal of Professional Communication 26(2013*), 2-13.

Bernardini, S. 2007. Collocations in translated language: Combining parallel, comparable and reference corpora. In *Fourth Corps Linguistics conference held at the University of Birmingham*, 27-30.

Bernardini, S., & Ferraresi, A. 2013. Old needs, new solutions: comparable corpora for language professionals. In *Building and using comparable corpora*, 303-319. Springer, Berlin, Heidelberg.

Biber, D. 1993. "Representativeness in corpus design." *Literary and Linguistic Computing* 8 (4): 243–57.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., J. Grieve, and G. Iberri-Shea. 2010. "Noun Phrase Modification." In *One language, two grammars? Differences between British and American English*, edited by Günter Rohdenburg and Julia Schlüter, 182–93. Cambridge: Cambridge University Press.

Biel, Ł. 2016. Mixed corpus design for researching the Eurolect: a genre-based comparable-parallel corpus in the PL EUROLECT project. In *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora.* Warszawa: Instytut Lingwistyki Stosowanej, 198-208.

Bilić, M., & Gaspar, A. 2018. Extraction of Phrasal Verbs from the Comparable English Corpus of Legal Texts. *International Journal of English Language and Translation Studies*, 184-194.

Bitinaitė, V. 2008. *Mokomasis anglų-lietuvių kalbų teisės terminų žodynas*. Vilnius: Eugrimas.

Blick, A. 2012. *Codifying—or not Codifying—the United Kingdom Constitution: The Existing Constitution.* London: Centre for Political & Constitutional Studies, King's College London.

Cabré Castellví, M. T., A. Condamines, and F. Ibekwe-SanJuan. 2007. *Application-Driven Terminology Engineering*. Amsterdam: John Benjamins.

Cabré Castellví, M. T., and J. C. Sager. 1999. *Terminology: Theory, Methods, and Applications*. Amsterdam: John Benjamins.

Celiešienė, V., and J. Džežulskienė. 2009. *Profesinės kalbos pagrindai: Vadovėlis*. Kaunas: Technologija.

Collin, P. H. 2004. *Dictionary of Law*. 4th ed. London: Bloomsbury.

Corpas, G., and M. Seghiri. 2009. "Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish)." In *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, edited by A. Beeby, P. Rodríguez, and P. Sánchez-Gijón, 75–107. Amsterdam: John Benjamins.

Cottrell, J., and S. Dhungel. 2007. *A Glossary of Constitutional Terms: English-Nepali*. Stockholm: International Institute for Democracy and Electoral Assistance.

Daille, B. 2017. *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*. Amsterdam: John Benjamins.

De Groot, G., and Conrad J. P. van Laer. 2006. "The Dubious Quality of Legal Dictionaries." *International Journal of Legal Information* 34 (1): 64–86.

De Groot, G., and Conrad J. P. van Laer. 2011. "Bilingual and Multilingual Legal Dictionaries in the European Union: An updated bibliography." *Legal Reference Services Quarterly* 30 (3): 149–209.

Deléger, L., & Zweigenbaum, P. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, 2-10.

Delpech, E., Daille, B., Morin, E., & Lemaire, C. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking. In COLING 2012, 745-762.

Durkin, P. 2011. *The Oxford Guide to Etymology*. Oxford: Oxford University Press.

EC (European Commission). 2006. *Terminologijos vadovėlis Europos komisijos vertimo raštu generalinio direktorato lietuvių kalbos departamento vertėjams*. Liuksemburg: EC DGT LT.

Fantinuoli, C. 2018. The use of comparable corpora in interpreting practice and training. *Interpreters' Newsletter No. 23 2018*, 133.

Fišer, D., Ljubešić, N., Vintar, Š., & Pollak, S. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 19-26.

Gaivenis, K. 2002. *Lietuvių terminologija: Teorijos ir tvarkybos metmenys*. Vilnius: Lietuvių kalbos institutas.

Giampieri, P. 2018. Online parallel and comparable corpora for legal translations. Altre Modernità, (20), 237-252.

Goeuriot, L., Morin, E., & Daille, B. 2009. Compilation of specialized comparable corpora in French and Japanese. In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC), 55-63.

Hunston, S. 2008. "Collection strategies and design decisions." In *Corpus Linguistics, An International Handbook*, edited by A. Lüdeling and M. Kytö, 154–68. Berlin: Mouton de Gruyter.

Ignatova, E. 2018. Compiling Comparable Multimodal Corpora of Tourism Discourse. In *Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching*, 96-114.

Infoterm. 2005. *Guidelines for Terminology Policies. Formulating and Implementing Terminology Policy in Language Communities*. Paris: UNESCO (United Nations Educational, Scientific and Cultural Organization).

Ismail, A., & Manandhar, S. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 481-489. Association for Computational Linguistics.

ISO (International Organization for Standardization). 2000. *Terminology work—Principles and methods*. ISO 704:2000 (E). 2nd ed. Geneva: ISO.

Jackson, H., and E. Zé Amvela. 2012. *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology*. 2nd ed. London and New York: Continuum.

Jakaitienė, E. 2010. *Leksikologija: Studijų knyga*. 2nd ed. Vilnius: Vilniaus universitetas.

Janulevičienė, V., and S. Rackevičienė. 2009. "Nusikalstamų veikų pavadinimai lietuvių ir anglų kalbomis." *Socialinių Mokslų Studijos*, Vol. 4 (4), 357–81, Vilnius: Mykolas Romeris University.

Janulevičienė, V., and S. Rackevičienė. 2010. "Lietuvių, anglų ir norvegų kalbų baudžiamosios teisės terminai." *Kalbų Studijos*, Vol. 17, 19–28, Kauno Technologijos Universitetas.

Janulevičienė, V., and S. Rackevičienė. 2012. “Legal Language in Intercultural Communication.” *Santalka: Filologija, Edukologija* 20 (2): 164–73.

Janulevičienė, V., and S. Rackevičienė. 2014. “Formation of Criminal Law Terms in English, Lithuanian and Norwegian.” *LSP Journal—Language for Special Purposes, Professional Communication, Knowledge Management and Cognition* 15 (1): 4–20.

Kageura, K. 2002. *The Dynamics of Terminology: A descriptive theory of term formation and terminological growth*. Amsterdam: John Benjamins.

Kageura, K. 2012. *The Quantitative Analysis of the Dynamics and Structure of Terminologies*. Amsterdam: John Benjamins.

Keinys, S. 1980. *Terminologijos abėcėlė*. Vilnius: Mokslas.

Keinys, S. 1999. *Bendrinės lietuvių kalbos žodžių daryba*. Šiauliai: Šiaulių universitetas.

Keinys, S. 2005a. *Dabartinė lietuvių terminologija*. Vilnius: Lietuvių kalbos instituto leidykla.

Keinys, S. 2005b. “Galūninė lietuviškų terminų daryba.” In *Dabartinė lietuvių terminologija*, 77–112. Vilnius: Lietuvių kalbos instituto leidykla.

Keinys, S. 2005c. “Lietuviškų sudurtinių terminų daryba.” In *Dabartinė lietuvių terminologija*, 129–82. Vilnius: Lietuvių kalbos instituto leidykla.

Keinys, S. 2005d. “Priesaginė lietuviškų terminų daryba.” In *Dabartinė lietuvių terminologija*, 21–76. Vilnius: Lietuvių kalbos instituto leidykla.

Keinys, S. 2005e. “Priešdėlinė lietuviškų terminų daryba.” In *Dabartinė lietuvių terminologija*, 113–28. Vilnius: Lietuvių kalbos instituto leidykla.

Keinys, S. 2005f. “Tarptautiniai elementai lietuvių terminologijoje.” In *Dabartinė lietuvių terminologija*, 193–201. Vilnius: Lietuvių kalbos instituto leidykla.

Keinys, S. 2005g. “Terminologijos kūrimo šaltiniai.” *Dabartinė lietuvių terminologija*, 231–33. Vilnius: Lietuvių kalbos instituto leidykla.

Keinys, S. 2012. *Lietuvių terminologijos raida*. Vilnius: Lietuvių kalbos instituto leidykla.

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

Kilgarriff, A. 2010. Comparable corpora within and across languages, word frequency lists and the KELLY project. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora,* 1-5.

King, B. 2009. “Building and Analysing Corpora of Computer-Mediated Communication.” In *Contemporary Corpus Linguistics*, edited by P. Baker. London and New York: Continuum. Continuum.

Koester, A. 2010. “Building small specialised corpora.” In *The Routledge handbook of corpus linguistics*, 66-79. Routledge, edited by O’Keeffee, A. and McCarthy, M. New York.

Kontutytė, E. 2008. “Įmonių teisinės formos: Vokiškų ir lietuviškų terminų ekvivalentiškumo problemos.” *Kalbotyra*, Vol. 58, 69–79. Vilnius: Vilniaus Universiteto Leidykla

Krishnamurthy, R., & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for academic purposes*, *6*(4), 356-373.

León-Araúz, P., Reimerink, A., & Cabezas-García, M. (2020, May). Representing Multiword Term Variation in a Terminological Knowledge Base: a Corpus-Based Study. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 2358-2367).

Lewandowska-Tomaszczyk, B., & Pęzik, P. 2018. Parallel and comparable language corpora, cluster equivalence and translator education. In *Общество и языки в третьем тысячелетии. Коммуникация. Образование. Перевод= Society and Languages in the Third Millennium. Communication. Education. Translation*, 131-143.

Li, B., & Gaussier, E. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International conference on computational linguistics*, 644-652. Association for Computational Linguistics.

Loock, R. 2015. From comparative grammar to translation studies: the use of DIY comparable corpora in the translation classroom. In *Communication présentée au colloque engcorpora,* 2015, Paris, France, <www.youtube.com/watch.

Losey-Leon, Maria-Araceli. 2015. “Corpus Design and Compilation Process for the Preparation of a Bilingual Glossary (English-Spanish) in the Logistics and Maritime Transport Field: LogisTRANS.” *Procedia—Social and Behavioral Sciences*. Vol. 73, 293 – 299. Elsevier Ltd.

Maksimaitis, M. 2014. “At the Beginning of Lithuanian Legal Language.” *Jurisprudence* 95 (5), 7-13.

Maley, Y. 1994. “The Language of the Law.” In *Language and the law*, edited by J. Gibbons, 11–50. London: Longman.

Marina, V. 2006. “The Analysis of English Metaphorical Terms and their Lithuanian and Russian Equivalents from the Perspective of Linguistic Relativity.” *Tiltai,* Vol. 2, 121–29.

Mattila, H. E. S. 2006. *Comparative Legal Linguistics.* Aldershot, England; Burlington, VT: Ashgate.

Mattila, H. E. S. 2012. *Comparative Legal Linguistics: Language of law, Latin and Modern Lingua Francas*. Farnham, England: Ashgate.

McEnery, T. 2003. "Corpus Linguistics." In *The Oxford Handbook of Computational Linguistics*, edited by R. Mitkov, 448–463. Oxford: Oxford University Press.

McEnery, T., and C. Gabrielatos. 2006. "English corpus linguistics." The handbook of English linguistics, edited by Aarts, B. & McMahon, A. 33–71. Oxford: Blackwell.

McEnery, T., and Hardie, A. 2012. Corpus linguistics: Method, theory and practice. Cambridge University Press.

McEnery, T., and A. Wilson. 2001. *Corpus Linguistics*. 2nd ed. Edinburgh: Edinburgh University Press.

McEnery, T. and Xiao, R. 2007. Parallel and comparable corpora: What is happening? In Anderman G. and Rogers M., editors, *Incorporating Corpora: The Linguist and the Translator.* Clevedon: Multilingual Matters.

McIntyre, D. 2010. *History of English: A resource book for students*. 2nd ed. London and New York: Routledge: Taylor & Francis Group.

Meyer, C. F. 2002. *English Corpus Linguistics*. Cambridge, UK: Cambridge University Press.

Mockienė, L., and S. Rackevičienė. 2014. "Contrastive Analysis of Constitutional One-Word Terms in Lithuanian, Russian and English." *Kalba ir kontekstai = Language in different contexts*. Lietuvos edukologijos universitetas. Filologijos fakultetas. 130-144. Vilnius: Edukologija.

Mockienė L., and S. Rackevičienė. 2015. "Sources of One-Word Terms Used in UK and Lithuanian Constitutional Law Acts." In *Taikomoji kalbotyra*. Vilnius: Vilniaus universitetas.

Mockienė, L., and S. Rackevičienė. 2016. "Analysis of term formation in Lithuanian, Russian and English terminology works." *Terminologija*, Vol. 23, 52–72.

Mockevičius, R. 2002. *Lietuvos Respublikos įstatymuose vartojamų sąvokų žodynas*. 2nd ed. Vilnius: Teisinės informacijos centras.

Morin E., Hazem A., Saldarriaga S. P. Bilingual Lexicon Extraction from Comparable Corpora as Metasearch. *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Jun 2011, Portland, United States, 35-43.

Morin E., Prochasson E. 2011. Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora. *4th Workshop on Building and*

Using Comparable Corpora: Comparable Corpora and the Web, Jun 2011, Portland, United States, 27-34.

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., & Paggio, P. 2011. Creating comparable multimodal corpora for nordic languages. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), 153-160.

O'Grady, W., M. Dobrovolsky, and M. Aronoff. 1997. *Contemporary Linguistics: An Introduction*. 3rd ed. New York: St. Martin's Press.

Ostler, N. 2008. "Corpora of less studied languages." In *Corpus Linguistics, An International Handbook*, edited by A. Lüdeling and M. Kytö, 457–84. Berlin: Mouton de Gruyter.

Paulauskienė, A. S. 2004. *Teisininkų kalba ir bendrosios normos. Monografija*. Vilnius: Justitia.

Pečkuvienė, L. 2009. "Terminų ir kitų administracinės kalbos žodžių daryba." In *Specialybės kalba: Administracinės kalbos vadovėlis*, 139–210. Vilnius: Mykolo Romerio universiteto Leidybos centras.

Pečkuvienė, L. 2013. "Administracinės kalbos žodynas normos ir vartosena." *Socialinių Mokslų Studijos: Mokslo Darbai = Societal Studies: Research Papers* 5(2). 525–539

Plag, I. 2003. *Word-formation in English*. Cambridge and New York: Cambridge University Press.

Postolea, S., & Ghivirigă, T. (2016). Using Small Parallel Corpora to Develop Collocation-Centred Activities in Specialized Translation Classes. Linguaculture, 2016(2), 53-72.

Potemkin, S. (2019). Multiword Terms and Machine Translation. Computational and Corpus-based Phraseology, 133-139.

Rackevičienė, S. 2006. "Baudžiamajai teisei priešingą veiką įvardijantys terminai Lietuvos ir Anglijos-Velso teisės sistemose." *Specialybės Kalba: Tyrimas ir dėstymas*. Vilnius: Mykolo Romerio universitetas, 114-122.

Rackevičienė, S. 2008. "Nusikalstamą veiką ir jos rūšis pagal pavojingumą įvardijantys terminai lietuvių ir anglų kalbomis." *Jurisprudencija*, 98–104.

Reppen, R. 2010. "Building a corpus: what are the key considerations?" In *The Routledge handbook of corpus linguistics*, 59–65. Routledge, edited by O'Keeffee, A. and McCarthy, M. New York.

Rey, A., and J. C. Sager. 1995. *Essays on Terminology*. Amsterdam: John Benjamins.

Rigouts Terryn, A., Hoste, V., & Lefever, E. 2020. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. LANGUAGE RESOURCES AND EVALUATION, 54(2), 385-418.

Rivera, O. M., Mitkov, R., & Pastor, G. C. (2018). A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In *Multiword Units in Machine Translation and Translation Technology*, 166-180. John Benjamins.

Roldán-Riejos, A., & Grabowski, Ł. (2019). Towards a cross-linguistic study of phraseology across specialized genres. Computational and Corpus-based Phraseology, 140.

Rudaitienė, V. 2008. *Leksikos skoliniai administracinėje kalboje*. Vilnius: Mykolo Romerio universiteto Leidybos centras.

Rudaitienė, V. 2012. "Tarptautinių veiksmažodžių vartosena ir lietuvių kalbos normos." *Gimtoji Kalba*, Vol. 8, 3-8.

Rudaitienė, V. 2013. "Kas tinka vienai kalbai, nebūtinai turi tikti kitai." *Gimtoji Kalba*, no. 6, 3–7.

Rundell, M. 2008. "The corpus revolution revisited." *English Today* 24 (1): 23–27.

Sager, J. C. 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

Sager, J. C. 1997. "Term formation." In *Handbook of terminology management*, edited by G. Budin, and S. E. Wright, 25–41. Amsterdam: John Benjamins.

Sager, J. C. 2004. "Terminology in Special Languages." In *Morphologie: Ein Internationales Handbuch zur Flexion und Wortbildung* [Morphology: An International Handbook on Inflection and Word Formation], edited by W. Kesselheim, S. Skopeteas, J. Mugdan, G. E. Booij and C. Lehmann, 1924–28. Berlin: De Gruyter Mouton.

Sánchez-Gijón, Pilar, Patricia Rodríguez Inés, and Allison Beeby Lonsdale. 2009. *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Benjamins Translation Library. Amsterdam: John Benjamins.

Sandrini, P. 1996. "Comparative Analysis of Legal Terms: Equivalence revisited." In *Terminology and knowledge engineering*, edited by C. Galinski and K. D. Schmitz, 342–51, Frankfurt a.M.: Indeks Verlag.

Sandrini, P. 1999. "Legal Terminology. Some Aspects for a New Methodology." *Hermes Journal of Linguistics*, 22, 101–12.

Schäfer, R., A. Barbaresi, and F. Bildhauer. 2013. "The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction." In Proceedings of the 8th web as corpus workshop, edited by S. Evert, E. Stemle, and P. Rayson, 7–15. Accessed at https://www.lt3.ugent.be/media/uploads/publications/2013/wac8-proceedings.pdf

Seghiri, M. 2011. “Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad.” Revista de Lingüística Teórica y Aplicada (RLA) 49, no. 2, II Sem., 13–30.

Šeškauskienė, I. 2013. *Ways with Words: Insights into the English Lexicon and Some Cross-Linguistic Aspects of Study*. Vilnius: Vilnius University Publishing House.

Sharoff, S., Babych, B., & Hartley, A. 2006a. Using collocations from comparable corpora to find translation equivalents. In *LREC*, 465-470.

Sharoff, S., Babych, B., & Hartley, A. 2006b. Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the COLING/ACL 2006*, 739-746.

Sharoff, S., Babych, B., & Hartley, A. 2009. ‘Irrefragable answers’ using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation*, 43(1), 15-25.

Sinclair, J. 1996. Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P. Available at: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html. Accessed March 3, 2020.

Sinclair, J. 2008. “Borrowed ideas.” In Language, People, Numbers: Corpus Linguistics and Society, edited by Andrea Gerbig and Oliver Mason, 21–42. Amsterdam / New York.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlič, Vasiļjevs A., Babych B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., Pinnis, M. 2012. Collecting and Using Comparable Corpora for Statistical Machine Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 438-445.

Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa, M., & Mastropavlos, N. 2010a. A collection of comparable corpora for under-resourced languages. In *Proceedings of the Fourth International Conference Baltic HLT 2010*, 161-168.

Skadiņa, I.; Vasiļjevs, A.; Skadiņš, R.; Gaizauskas, R.; Tufiş; D, Gornostay, T. 2010b. Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, European Language Resources Association (ELRA)*, La Valletta, Malta, May 2010. 6-14.

Smoczyński, W. 2007. *Słownik etymologiczny jezyka litewskiego = lietuvių kalbos etimologinis žodynas*. Vilnius: Vilniaus universiteto leidykla.

Snover, M., Dorr, B., & Schwartz, R. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 857-866.

Ştefănescu, D. 2012. Mining for term translations in comparable corpora. In *The 5th Workshop on Building and Using Comparable Corpora*, 98-103.

Stewart, W. J. 2001. *Collins Dictionary of Law*. 2nd ed. Glasgow: HarperCollins.

Temmerman, R. 2000. *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam: John Benjamins.

Umbrasas, A. 2010. *Lietuvių teisės terminija 1918-1940 metais: Pagrindinių kodeksų terminai*. Vilnius: Lietuvių kalbos institutas.

Urbutis, V. 1978. *Žodžių darybos teorija*. Vilnius: Mokslas.

Valeontis, K., and E. Mantzari. 2006. "The Linguistic Dimensions of Terminology: Principles and Methods of Term Formation." 1st Athens International Conference on Translation and Interpretation Translation: Between Art and Social Science, 13–14 October 2006, Athens, Greece, 1–20.

Van Gelderen, E. 2006. *A History of the English Language*. Amsterdam / Philadelphia John Benjamins.

Vansevičius, S. 2000. *Valstybės ir teisės teorija: Mokomoji priemonė*. Vilnius: Justitia.

Vulic, I., & Moens, M. F. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, 449-459. ACL; East Stroudsburg, PA.

Zanettin, F. 1998. Bilingual comparable corpora and the training of translators. Meta: journal des traducteurs/Meta: Translators' Journal, 43(4), 616-630.

Zanettin, F. 2002a. "CEXI: Designing an English Italian Translational Corpus." *Language and Computers*. 329–43.

Zanettin, F. 2002b. "DIY Corpora: The WWW and the Translator." In *Training the Language Services Provider for the New Millennium*, edited by Belinda Maia, Jonathan Haller and Margherita Urlrych, 239–48. Porto: Facultade de Letras, Universidade do Porto.

Zanettin, F. 2011. "Translation and Corpus Design." *Synap*s 26, 14–23.

Zanettin, F. 2013. Corpus methods for descriptive translation studies. Procedia-Social and Behavioral Sciences, 95, 20-32.

CHAPTER THREE

APPLICATION OF CORPUS ANNOTATION IN LANGUAGE STUDIES AT UNIVERSITY LEVEL

GIEDRĖ VALŪNAITĖ OLEŠKEVIČIENĖ

This chapter focuses on the research on the process of teaching and learning a foreign language at more advanced levels while applying corpus analysis and building tools for corpus annotation. The research participant experience is analyzed in a structural way by aiming to formulate certain conclusions and recommendations for the application of corpus analysis and building tools while teaching and learning a foreign language at more advanced levels. This chapter provides a brief theoretical background focusing on corpora application in language studies, followed by certain issues of discourse management and organization, and closing with insights on principles of teaching and learning with technology and the role of the initial knowledge. The methodological approach of the research is also delineated, and the author provides the grounds for the methodological choices of the qualitative research and describes the research procedures. Finally, the results of the research are presented, and the author provides recommendations for teaching and learning a foreign language at more advanced levels while applying corpus analysis and building tools.

3.1. Theoretical background for the research on applying corpus annotation while teaching a foreign language at more advanced levels

The section comprises certain theoretical insights on the current research. Corpora use in language teaching and learning appears to be an innovative and effective supplement, especially for teaching and learning textual cohesion and coherence at more advanced levels. Pragmatic discourse management has always been important in mastering a foreign language and being able to relate the ideas in the text coherently. Learning with the corpus analysis and building tools and corpora annotation software also involves the necessity to reflect on the principles of learning with technology by theorizing on the teaching and learning environment which provides the learners with the tasks and the supportive information they need to create cognitive schemas.

3.1.1. Corpora development and use in language teaching

The development of corpora—in other words, large language databases—disclosed the potential of language research by using various corpus research techniques focused on examining patterns of lexis, grammar, semantics, pragmatics and textual features. Depending on the research purposes, many corpora are coded according to the parts of speech or examined for grammatical structure, or the investigation is focused on the pragmatic features.

Corpora development has fueled the progress in the advancement of the knowledge concerning lexis, grammar, semantics, pragmatics and textual features (Sinclair 1991; Stubbs 2004). Corpus linguistics is based on the theoretical insights that language varies according to the context related to space and time, which lead to establishing new facts about language on the basis of almost infinite potential. The application of the same theoretical insights to language teaching and learning practices leads to the significant use of corpora in teaching and learning languages. It has already been noted that dictionaries and grammars do not have the capacity to fully describe the language. Thus, corpora application for teaching and learning languages allows both teachers and learners to identify certain regularities and irregularities of the language by researching and relaying the corpora data. According to Aston (2001), another benefit of corpora application is that a corpus-based approach provides real data of live language used in real contexts. The author also identifies that it is important to take into account the frequency information, which might be helpful both for teachers and learners while making real language use choices. Scott and Tribble (2006) identify that while dealing with a language at more advanced levels it also becomes important to acquire definitive knowledge of genres and registers. Granger (2015) supports the idea by carrying out learner corpus research which reveals that the patterns used by relatively advanced language learners have a tendency to exemplify more stylistic discrepancies rather than grammatical problems. The problematic areas for advanced language learners seem to be coherence, cohesion and textual rhetorical features. Thus, cohesive devices and discourse markers become the focus of the researchers' attention as the tools for ensuring textual and discourse management. Research on proper discourse use is looking for answers as to what and how to teach at more advanced levels concerning the matters of textual features. The suggestions offered by the recent research lead to the idea of direct corpus use by language learners and teachers. The studies by Cobb and Boulton (2015) have shown that the application of the innovative idea of using corpora in teaching and learning appears to be effective and efficient. The authors reveal that learners acquire better skills of linking adverbs by using corpus concordances rather than using bilingual dictionaries or grammars. The development of discourse-annotated corpora could be a new step and could provide an innovative supplement in the surplus of teaching/learning materials, especially for more advanced learners in dealing with textual cohesion and coherence.

Corpus linguistics collects and analyzes vast quantities of texts to extend our understanding of language and to provide up-to-date linguistic data. It also produces a wide variety of reference materials which are relevant to teaching and learning foreign or second languages. They appear especially applicable while applying new approaches to language learning; one example of this is data-driven learning (DDL), coined by Johns (1990), which he characterizes as, "The attempt to cut out the middleman as far as possible and to give the learner direct access to the data, the underlying assumption being that effective language learning is a form of linguistic research." According to Boulton and Tyne (2014), DDL embraces a number of key concepts in the existing approaches of language learning, such as authenticity, autonomy, cognitive depth, consciousness-raising, constructivism, context, critical thinking, discovery learning, heuristics, ICT, individualization, induction, learner-centeredness, learning to learn, lifelong learning, (meta-) cognition, motivation, noticing, sensitization and transferability. Thus, the authors advocate for DDT as it may provide the necessary exposure to authentic language. According to usage-based theories (Tomasello 2005), learners need substantial exposure to language, but it is problematic to ensure in a foreign-language classroom. DDL can help to organize and focus the exposure (Gaskell and Cobb 2004). This model embraces the theoretical assumptions that language is not rule-driven but fuzzy and probabilistic in nature (Hanks 2013). It also embraces the premise that the mind works beyond the level of the word in line with dynamic systems theory (Larsen-Freeman and Cameron 2008).

Fawcett (1987) observes that corpus-based teaching and learning could be a promising translator preparation method because the purpose of translator education is to equip trainees with skills applicable to any texts related to any subjects, and corpus-based teaching can provide trainees with such skills. The author stresses that corpus-based translation classes enable students to learn about corpora, corpus analysis tools and their applications for translation. Students can compile corpora on a variety of texts and use them for learning. So, it is useful making students familiar with DIY (do-it-yourself) corpora because they can learn to build their own corpora for any type of texts and use the corpora during their studies and professional careers.

Sinclair (1991) discusses a feature of language patterns known as semantic prosody, which deals with subtle implicatures which are processed subliminally. Corpora concordances may demonstrate how words are imbued with negative or positive meaning due to the contextual collocates. This is a challenging task for translators as they have to take into account the subtle implicatures produced by semantic prosody.

Corpora are a resource for many applications concerning translation, for terminology mining and also for reusing previous translations. The tool which enhances corpora value is annotation—adding linguistic information. The annotations could be used later for research purposes to constrain searches in a corpus for information. Aligned corpora could serve many research purposes; however, the process of automatic alignment is a complex task. Although the translated texts tend to be organized in the same way as the original, there might be different punctuation and the notion of sentences may vary from language to language. So the segmentation of the source and the translated text may differ in terms of the length of the segments and also in the order of the segments.

Learning with corpora during studies gives students experience which they can use later as they will know how to compile corpora on a variety of texts and how to extract information from different types of corpora. Corpus-based training can also be beneficial concerning technical writing and editing. Bowker and Pearson (2002) identify that language for specific purposes (LSP) corpora can be used as writing guides for particular styles or technical texts.

Corpus-based translation classes can also provide students with an opportunity for collaboration and working together; students can create and use their corpus to assess and revise translations by their peers by grounding their criticisms of their peers' translations in corpus-based evidence. Collaboration can help students develop their communicative and interpersonal skills, which are useful for dealing with their colleagues and future clients. In addition, a corpus could be used for revising and editing translations, which is also a necessary skill for a translator.

Kubler and Foucou (2003) identify another benefit of using corpora in teaching translation: developing students' computer skills. Working with corpora demands and stimulates the development of computer literacy and skills. Thus, while using computers and corpora software, future translators develop their computer skills during their studies. Kubler and Foucou (2003) also state that learning to use corpus analysis tools can equip future translators with technical research skills that might be necessary, especially in the translation research field.

3.1.2. Corpora annotation tool used in the research

The current research was based on using the TED-MDB project (Zeyrek, Mendes, and Kurfalı 2018) annotation scheme, which sprung from the Penn Discourse Treebank (PDTB) tool and its annotation scheme, so it is necessary to introduce the tools briefly. The PDTB is an over 2-million term corpus manually annotated for discourse-level information on discourse relations (Prasad, Webber, and Joshi 2014). The PDTB views discourse structure as embracing a logical flow of events, states and propositions. The PDTB annotation scheme includes explicit and implicit discourse connectives, alternative lexicalizations, entity relations and no relations between the annotated binary arguments, namely, Arg1 and Arg2. There is also a system of senses assigned to all discourse relations except for entity relations and no relations. The PDTB's annotation approach is theory-neutral and lexically grounded. The theory-neutral approach means that the annotation is not based on a specific discourse structure or on specific theoretical assumptions. Lexically grounded perception implies that annotator judgments are based on the annotation schema and the instruction set and are effectively elicited, even when there is no explicit discourse connective that links the two arguments (related text spans). The development of the PDTB also stimulated interest in cross-linguistic studies of discourse relations in other languages, giving rise to similar PDTB-based annotation projects—e.g., Turkish (Zeyrek et al. 2013), Arabic (Al-Saif and Markert 2010), Chinese (Zhou and Xue 2012) and Hindi (Oza et al. 2009)—as well as stimulating the multilingual annotation project TED-MDB (Zeyrek, Mendes, and Kurfalı 2018). In line with the annotation project TED-MDB, Lithuanian texts were annotated with discourse relations (e.g., causal, contrastive, elaboration and temporal relations); thus, annotation schema and a set of instructions have been adapted for the Lithuanian language (Oleskeviciene et al. 2018).

In Lithuanian, explicit discourse connectives include expressions from four grammatical classes: subordinating conjunctions—e.g., *kai*, *kol*, *nes*, *kadangi* (*when*, *while*, *because*, *since*); coordinating conjunctions—*ir*, *bei*, *o*, *tačiau* (*and*, *but*, *or*, *however*); sentential relatives—*tam kad*, *tuo metu kai* (*so that*, *at the time when*); and discourse adverbials—*faktiškai*, *galiausiai* (*actually*, *eventually*). In the annotation process it is important to identify whether the words and phrases function as discourse connectives as they can have other functions, especially in the case of adverbials. The argument annotation of explicit discourse connectives and alternative lexicalizations follows the rule that the label Arg2 is attached to the argument which appears in the clause that is syntactically bound to the connective; the other argument is marked Arg1. As in TED-MDB, adverbials called "discourse markers" (Hirschberg and Litman 1987) are not annotated as they signal the organizational structure of the discourse instead of relating two arguments with an abstract object interpretation (Asher 1993). Examples to explain the annotation of Lithuanian texts are taken from TED-MDB Lithuanian part. For example, Lithuanian *Dabar* and its equivalent in English (*Now*) in the examples 1 and 2 below serve to signal discourse organizational structure, so such cases were not annotated.

(1) Dabar kaip matote įtampa, apie kurią girdėjome San Fransiske apie susirūpinimą dėl būsto kainų ir gyventojų išstūmimo ir technologijų kompanijų, kurios atneša daug turto ir įsikuria, yra tikra.

(2) Now you can see, though, that the tensions that we've heard about in San Francisco in terms of people being concerned about gentrification and all the new tech companies that are bringing new wealth and settlement into the city are real.

In annotating implicit connectives, the annotator has to insert a connective that best expresses the inferred relation between two adjacent sentences, where the first sentence is Arg1 (shown in italics) and the second is Arg2 (shown in boldface). For example:

(3) *...sle˙pti galvą sme˙lyje ir negalvoti apie tai*. [Implicit=Bet] **Jei tik galite, priešinkitės tam**. (Implicit) (Comparison: Contrast)

(4) *... Bury our heads in the sand and not think about it*. [Implicit=But] **Resist this, if you can**. (Implicit) (Comparison: Contrast)

Alternative lexalization (AltLex) involves cases when discourse relations between adjacent clauses could be inferred, and a redundancy appears if an explicit connective is inserted. The reason for the redundancy is that the relation is already expressed by some alternatively lexicalized non-connective expression. For example:

(5) *Se˙kme˙ mus motyvuoja, bet beveik pasiekta pergale˙ skatina mus leistis i˛ nuolatinius ieškojimus*. [Vieną iš ryškiausiu˛ to pavyzdžiu˛ pastebime], **kai žvelgiame i˛ skirtumą tarp olimpinio sidabro laime˙toju˛ ir bronzos laime˙toju˛ rungtyne˙ms pasibaigus**. (AltLex) (Expansion: Instantiation)

(6) *Success motivates us, but a near win can propel us in an ongoing quest*. [One of the most vivid examples of this comes] **when we look at the difference between Olympic silver medalists and bronze medalists after a competition**. (AltLex) (Expansion: Instantiation)

Entity relations (EntRel) are annotated between adjacent sentences when it is felt that an entity in one argument is described further in the other argument, as in example 7 below.

(7) *Jie ture˙tu˛ i˛vertinti ir tuos efektyvumo rodiklius, kuriuos vadiname ASV: aplinkosauga, socialiniai klausimai ir valdymas*. **Aplikosauga apima energijos vartojimą, prieigą prie vandens, atlieku˛ tvarkymą ir taršą ir ekonomišką ištekliu˛ naudojimą**. (EntRel)

(8) *Investors should also look at performance metrics in what we call ESG: environment, social and governance*. **Environment includes energy consumption, water availability, waste and pollution, just making efficient use of resources**. (EntRel)

No relation (NoRel) is annotated when there is no relation inferred by the annotator (reader) between the adjacent sentences:

(9) *Tai 4 milijardai viduriniosios klase˙s žmoniu˛, kuriems reikia maisto, energijos ir vandens*. **Dabar jūs tubuūt klausiate savęs: gal tai tik pavieniai atvejai. (NoRel)**

(10) *That's four billion middle-class people demanding food, energy and water*. **Now, you may be asking yourself, are these just isolated cases**? (NoRel)

TED-MDB adds a new top-level category, called hypophora, to the PDTB 3.0 relation hierarchy. The introduction of this category is aimed at capturing rhetorical question-response pairs, where the question is asked and answered by the speaker. TED-MDB annotates hypophora as a case of AltLex where alternative lexicalization is expressed by the question word. Where possible, the additional sense of the Q/R pair may be added.

As in TED-MDB, in Lithuanian, we annotate the question as Arg2 and the answer as Arg1. We consider the question as Arg2 because the AltLex connective expressed by the question word is a part of the question. Thus, AltLex includes the question word (the wh-word or *ar*, a specific question particle used in Yes/No questions, which can also serve as an explicit connective in certain cases in the Lithuanian language). *Ar*, a specific question particle, is presented in example 11 and its equivalent in example 12; this demonstrates their selection as AltLex which marks the discourse relation that exists between the question and the answer. For example:

(11) *Tai verčia mane klausti*, [ar] **šiandienos investavimo taisyklės tinkamos ateities tikslams**. (Explicit) (Expansion: Level-of-detail: Arg2-as-detail)

(12) It makes me wonder [if] **investment rules of today are fit for purpose tomorrow**. (Explicit) (Expansion: Level-of-detail: Arg2-as-detail)

The following examples illustrate how hypophora is annotated in Lithuanian and English. Lithuanian Q/R pairs are annotated for a primary sense first, and then they are tagged as hypophora using the secondary sense.

(13) [Ar] **i˛mone˙s, atsižvelgiančios ˛i tvarumą, išties ftnansiškai se˙kmingos**? *Galintis nustebinti atsakymas yra “taip”* (Explicit) (AltLex: Ar; Expansion: Level-of-detail: Arg1-as-detail; Hypophora)

(14) [Do] **companies that take sustainability into account really do well financially**? The answer that may surprise you is yes. (AltLex: Do) (Hypophora)

(15) [Kode˙l] **kas nors apskritai rinktu˛si toki˛ gyvenimą** - *Atsakymas ˛i ši˛ klausimą gali skirtis, kaip skiriasi ir žmone˙s sutinkami kelyje, bet keliautojai dažnai atsako vienu žodžiu: laisve˙*. (Explicit) (AltLex: Kode˙l; Contingency: cause: Reason; Hypophora).

(16) [Why] **anyone would choose a life like this, under the thumb of discriminatory laws, eating out of trash cans, sleeping under bridges, picking up seasonal jobs here and there**. *The answer to such a question*

is as varied as the people that take to the road, but travellers often respond with a single word: freedom. (AltLex: Why) (Hypophora)

The partial representation of the first two layers of the hierarchy of discourse relation senses could be presented schematically by introducing the main four types of the discourse relation hierarchy, which, on their own, are subdivided further in Table 1 below.

As seen in Table 1, the main four groups in the hierarchy of connective senses are temporal, contingency, comparison and expansion, which on their own include lower hierarchical subdivisions.

Table 3-1. Hierarchy of connective senses after Prasad, Webber, and Joshi (2014)

Discourse relations										
Temporal		Contingency		Comparison		Expansion				
Synchronous	Asynchronous	Cause	Condition	Contrast	Concession	Conjunction	Restatement	Instantiation	Alternative	Exception

3.1.3. Discourse organization and management

As has already been mentioned, teaching and learning a foreign language at more advanced levels brings out the importance of discourse management for the learners. Pragmatic competence has been acknowledged to be one of the key components of linguistic competence in the Common European Framework documents. According to the *Common European Framework of Reference for Languages: learning, teaching, assessment* (Council of Europe 2011), communicative language competence could be defined as comprising certain components: linguistic, sociolinguistic and pragmatic; and each component in its own turn comprises knowledge and skills and know-how. For example, linguistic competence is defined to include lexical, phonological and syntactical knowledge and skills and other dimensions of a language system, and it does not depend on the other components, such as the sociolinguistic and the pragmatic ones. The linguistic component represents not only the range of knowledge (e.g., phonetic distinctions or vocabulary) but also cognitive organization concerning the way the knowledge is stored (e.g., the associative networks lexical of items) and its accessibility (recall and availability). Also, knowledge may be active, which means readily expressible, or it may be not active. Cognitive organization and accessibility of knowledge vary among individuals. They can also depend on the cultural features of the language community in which the individual socializes and operates. Sociolinguistic competence refers to the socio-cultural conditions of language use. This includes rules of politeness, social norms and linguistic codification of certain rituals in the community. The sociolinguistic component is really important and especially impactful in communication between representatives of different cultures. Pragmatic competence is related to the functional use of linguistic resources; for example, production of language functions and speech acts. It is also related to the mastery of discourse, cohesion and coherence, and the identification of text types and forms. This component is even more important in multicultural environments. All the components discussed characterize the areas of communicative language competence which should be internalized by a learner, so the learning process is directed to develop or transform the internalized representations, mechanisms and capacities.

Returning to discourse, according to Baker (2018) discourse embraces certain linguistic features which allow it to be identified as a connected piece of text. First, we observe certain connections of information, how ideas are arranged within and among the sentences to establish the development of the topic. Then, there are surface connections which outline the relations between events and people in the discourse. And finally, there are semantic connections which allow understanding and interpreting of the discourse and of how different parts of it relate to each other. Another important feature of discourse organization is the genre. It embraces the features of discourse organization according to which the process of the interpretation is facilitated by providing the expectations for interpretation.

Thematic structure of discourse could be explained by thematic organization in a clause suggesting that a clause consists of two segments, where the first segment is considered to introduce the theme and the second part the rheme (Halliday 1994). For example:

The play (theme) was interesting. (rheme)

According to Halliday (1994), conjunctions, which are special linking devices (however, because, etc.), and disjunctions, which express the attitude of the speaker (actually, in my opinion, etc.), are considered inherently thematic. However, as they are not a part of the proposition they are not included in the thematic clause analysis.

It should be added that the Hallidayan approach identifies the theme-rheme sequence by ordering the clause elements while the Prague school applies a more complex approach, rejecting sentence position as the essential criterion for identifying theme and rheme structure.

Both cohesion and coherence are related to textual organization and relations, but cohesion represents the network of surface relations such as lexical and grammatical relations in discourse, whereas coherence represents conceptual relations beyond the surface of discourse such as connections of conceptual meaning. According to Hoey (1991), cohesion is a property of the text and coherence is a facet of the reader's evaluation of the text and may vary from reader to reader. The example of implicature is well described by Charolles (1983) in his analysis of the example below:

I went to the cinema. The beer was good.

The author explains that the reception and interpretation might be the following: the speaker says that he went to the cinema and drank beer there and that the beer was good. The author points out that the interpretation naturally provides the necessary links to make the discourse coherent. This way, the implicature deals with the notion that we understand more than it is actually said.

Translation issues of discourse connectives

Discourse connectives signal the manner in which the speaker or writer would like the listener or reader to connect the ideas which are going to be expressed to the ideas which have already been expressed before. According to Baker (2018), a variety of discourse connectives are used to signal different discourse relations, and the discourse relations are expressed by a variety of means. The author provides the example that in English causality could be expressed through both the verbs, such as *cause* or *lead to*, and discourse connectives expressing the causality relation. Languages vary in the frequency of use of certain types of connectives and also in the dominating preferences of the connective use. As the connectives are used to express the relations between the chunks of information, they are related to the chunking of information and produce certain insights into the whole logic of discourse (Smith and Frawley 1983). Some languages have a tendency of expressing relations through complex structures while other languages prefer the use of simpler structures which require marking the relations between the structures explicitly. For example, a noticeable difference between English and Arabic is well described in research literature. While in the English language there is a certain preference to present information in smaller chunks and signal the relations between the chunks, in the Arabic language there is an expressed preference to group information into large grammatical chunks (Holes 1984). This raises the question of how the translators deal when they face the necessity of transferring the multitude of explicit connectives into just a limited number of connectives, or vice versa. The process of adjusting the language-inherent patterns of connectives with target-language specifics and text-type preferences is a complicated task for translators as connectives deal with the logic of the text and they are related to text interpretation. Translators could apply twofold strategies: insertion of additional connectives, although there are no connectives in the original, seeking to make the translation smoother, or preserving all the explicit connectives in the original, although the translation might sound foreign in the target language. In real translation practice, specialists choose something in between or usually apply a bit of both techniques (Baker 2018).

In their recent study, Hoek et al. (2017) examine the types of discourse connectives which have an expressed tendency to be omitted more frequently in translation. The authors propose a hypothesis that cognitively simple relations are expected to be omitted more often than more complex relations. The researchers conduct a parallel corpus study on English parliamentary debates translated into Dutch, German, French and Spanish which demonstrates that some relations types are more prone to omission, identifying that speech-act relations and positive causal relations demonstrate a certain tendency to be omitted. In addition, Dupont and Zufferey (2017) employ translation corpora to research not only omissions but also the effect of register, translation direction shifts of meaning and translator's expertise by focusing on English and French markers of concession. The authors use the TED Talks corpus for their research, so they also discuss the specific characteristics and methodological issues related to one of the TED Talks corpus uses for investigation (Cettolo, Girardi, and Federico 2012). The authors analyze such specific types of translation as subtitling, which raises certain issues discussed by Lefer and Grabar (2015), which include the mix of both spoken and written features, the variety of TED Talks speakers involving non-native speakers or speakers of various regional varieties of English, and the limited expertise of the amateur volunteer translators. The authors compare TED Talks and newspaper articles and arrive at the conclusion that concessive markers usually have one or two most frequent translation equivalents in TED Talks, whereas they demonstrate a wider variety in news articles. Dupont and Zufferey suggest that translations might be more faithful to the original in TED Talks, but they remain careful in their observations because of the "noise arising from the specific translation features of the TED corpus" (2017, 284). They stress that omissions are less frequent in the TED corpus while being more frequent in the other genres investigated, which the authors identify as a surprising result given the space restrictions of subtitles; they provide the explanation of the higher necessity to "maintain highly explicit links in argumentative language" (2017, 286).

3.1.4. Translation competence models

Languages have their own patterns and devices for conveying certain events and relationships and for relating the ideas cohesively. Thus, the topic of cohesion has always been the most useful part of discourse analysis or text linguistics related to translation (Newmark 1987). Cohesion could be defined as the network of lexical, grammatical, and other relations which bind various parts and ideas of any text. These textual relations help to organize the text and, to some extent, to create the text, at the same time requiring the reader or the listener to interpret words, expressions and ideas in relation to other words, expressions and ideas in the surrounding parts of the text—for example, sentences and paragraphs. Cohesion is a textual relation which connects the actual words, expressions and ideas that the reader or listener can see or hear. Halliday and Hasan (2014) identify five main cohesive devices in English: **reference**, **substitution**, **ellipsis**, **conjunction**, and **lexical cohesion**.

Vermeer (1994), in his general theory of translation, supports the idea that the target text needs to convey the specific purpose, or so-called *skopos* (where "*skopos*" is a Greek word which means "aim"), in the target culture. Vermeer's theory advocates for the idea that translation should meet the purpose for which the target text is intended; thus, any strategy can be applied in translation as long as the purpose is observed (Smith 2002). The best example are the ideas by Au (1999), who states that especially while translating adverts translators are free to choose any strategy from the whole range of strategies in order to fulfil the purpose (*skopos*) of the original. Similarly, representative of the functionalist approach Nord (1997) also admits that all the strategies of cultural adaptation, paraphrase, expansion, reduction, modulation, transportation, substitution, loanword, calque, literal translation or even omission are all acceptable translation strategies as long as the purpose is conveyed by the target text. Speaking about the cultural component, many scholars view culture as an indispensable element which influences the translation process and product (Newmark 1988; Sager 1983; Martin and Hewson 1991; Williams 1989).

While teaching cohesion and dealing with cohesion in translation, certain aspects of text linguistics and functional approaches should be kept in mind. For example, it is important to draw together the ideas proposed by scholars such as Snell-Hornby (1988), who advocates an integrated approach to translation, and Baker (2018), who calls for an interdisciplinary approach. This way, development of translation competence (TC) should be a natural consequence of the implementation of integrated approaches. What is more, a trainee translator must develop a level of sufficient target language (TL) socio-cultural experience and language competence to be able to make decisions while translating and using TL without significant deliberation about comparative-contrastive, linguistic and stylistic use.

Hatim and Mason (1997), based on Bachman's (1990) work, present a traditional three-part competence inherited from linguistics—Source Text (ST) processing, transfer, Target Text (TT) processing. Pym (2003) introduces the core of translation competence made up of two key abilities: "the ability to generate a series of more than one viable TT for a pertinent ST" and "the ability to select only one viable TT from this series, quickly and with justified confidence" (Pym 2003). However, Pym (2003) also includes linguistic competence (grammar, rhetoric), extra-linguistic competence (world knowledge), instrumental competence (terminology, computer skills, Internet savvy) and professional translation competence (teamwork cooperation, strategies for getting paid correctly). He introduces the view that translation could be understood as “a problem-solving process”.

The currently widely used PACTE (Action Plan for Business Growth and Transformation) model is made up of a set of sub-competencies which are interrelated in a hierarchical way, with the strategic sub-competence acknowledged as the central one, in a dominant position in the proposed model. The list of sub-competences includes:

- bilingual sub-competence
- extra-linguistic sub-competence
- knowledge about translation sub-competence
- instrumental sub-competence
- strategic sub-competence
- psycho-physiological sub-competence

In the model, the bilingual sub-competence consists of pragmatic, sociolinguistic, textual and lexical-grammatical knowledge in each language. The extra-linguistic sub-competence includes encyclopedic, thematic and bicultural knowledge. The knowledge in the translation sub-competence comprises the knowledge of the principles that guide translation (processes, methods and procedures, etc.) and the profession (types of translation briefs, users, etc.). The instrumental sub-competence consists of knowledge related to the use of documentation sources and information technologies applied to translation. The strategic sub-competence is viewed as the most important one because it includes such crucial factors as solving problems and the efficiency of the process. It also includes planning the process of the translation project, evaluating the process and partial results obtained, activating the different sub-competencies and compensating for deficiencies, identifying translation problems and applying procedures to solve them. The psycho-physiological sub-competence includes such components as cognitive and behavioral ones (memory, attention span, perseverance, critical mind, etc.) and psychomotor mechanisms (PACTE 2003). Any bilingual person has knowledge of two languages and may have extra-linguistic knowledge; we consider that the sub-competencies specific to TC are the strategic, the instrumental and knowledge about translation (PACTE 2003, 2011; Albir 2017).

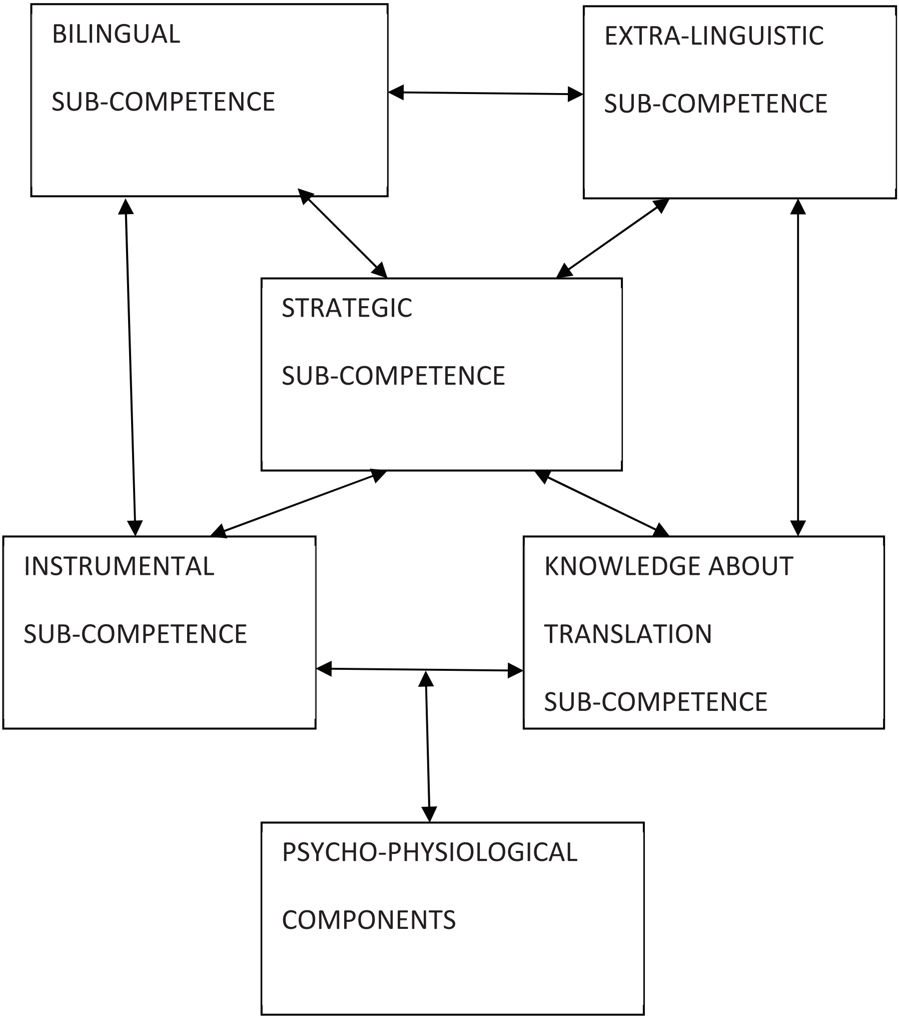


Figure 3-1. PACTE translation competence model, based on PACTE (2003)

The discussion regarding the models of translation competence leads to the discussion on evaluation of translation quality and assessing student translation performance. Adab (2000) suggests three main parameters: students' comprehension of the Source Text (ST); students' production ability in the target language (TL); and students' editing ability in the TL. Additionally, the holistic assessment model was suggested by Biggs and

Tang (2007). According to Biggs and Tang (2007), the student's total performance must be assessed holistically, so at the same time the conceptual framework of assessment must relate the whole to its parts. Thus, even establishing a comprehensive set of assessment components, their proposed model provides an assessment of overall performance and is referred to as a holistic assessment. However, the authors include four main criteria:

1. Translate texts by ensuring that the function and informative intent, and the reasoning and argumentation of texts, are fully and effectively communicated. Target competencies: Translational, linguistic, textual, cultural/encyclopedic, reasoning, strategic.

2. Edit and revise one's own translation to produce readable and typographically grammatically correct texts where target language textual features are used appropriately. Target competencies: Linguistic, textual, strategic.

3. Apply specific methods and techniques effectively while translating specialized texts. Target competencies: Reasoning, strategic.

4. Understand specialized concepts and be able to retrieve correct terminology by being able to appropriately use authoritative resources. Target competencies: Linguistic, cultural/encyclopedic, reasoning, strategic.

The above-discussed criteria for evaluating translation competence reveal that the holistic picture embraces a many-sided view in which linguistic and textual competences are important.

Technologies in translation

The most frequently used technologies in translation are machine translation and computer-aided translation. Machine translation could be defined as a process directed to the core task, which is to produce the translation of a source text. Computer-aided translation embraces the technological tools designed to aid the human translation and the core task is left to the human translator. Technology is becoming central to managing translation tasks, especially the large ones. Corpora are also useful resources for working with the texts. According to McEnery (2003), a corpus is a collection of naturally occurring language data which could be exploited in machine translation if it is made machine-readable.

3.1.5. Principles in learning with technology

Guided discovery learning

The existing approaches to instructional learning could be viewed as a continuum where at one end we observe direct instruction approaches which advocate for explicitly presented learning content to the students through textbooks, teacher demonstrations, etc. (Kirschner, Sweller, and Clark 2006). On the other end we could observe approaches based on learner-centered activities which require students to extract the learning content themselves. Discovery approaches advocate that the material "generation effect" enhances the results of learning (Bertsch et al. 2007). In addition to that, Rocard et al. (2007) recommend scientific discovery learning, which is a process in which students investigate scientifically oriented questions, conduct experiments, formulate explanations and evaluate their explanations in light of the alternatives. De Jong (2006) describes this process as an inquiry cycle, which consists of five phases:

during the orientation phase
students conduct a broad analysis of the domain to identify the main concepts and variables
during the hypothesis stage
students generate a specific statement or a model to be tested
during the experimentation stage
students test the hypothesis by manipulating the variables and interpreting the outcomes
during the conclusion stage
students determine the validity of their hypothesis or models
during the evaluation stage
students reflect on their learning process and the knowledge acquired.

Also, De Jung and Njoo (1992) introduced the concept of regulation, which is the planning and monitoring of the learning process and which could be applied to the whole scientific inquiry cycle. What is more, Mayer (2004) observed that unguided discovery is generally ineffective and students need adequate guidance while applying scientific discovery learning or any other technique related to discovery learning.

Another important point in guided discovery learning is the direct presentation of information. It may seem that guided discovery learning does not allow explicitly presented information, but there are situations when students do not have sufficient prior knowledge or they have difficulty in discovering the required information on their own; then, the necessary information could be presented explicitly to the students. The authors Hmelo-Silver, Duncan and Chinn (2007) demonstrate that after receiving the necessary information students can turn back to the discovery mode. A well-cited example of direct presentation is provided by Klahr and Nigam (2004), who gave an interactive lecture to children about unconfined experimental design enabling them to use CVS (control-of-variables strategy) better than the children who had to use unguided discovery. Similarly, Lazonder, Hagemans and de Jong (2010) proved that students guided by background information demonstrate better results in the experimentation stage; the authors proved that providing background information before and during the inquiry stage is more effective than just presenting information before the inquiry stage. De Jong and Lazonder (2014) observe that the amount of guidance should be well-balanced because too little guidance may impede the learning process and too much guidance may challenge the discovery nature of learning. The amount of guidance depends on students' prior knowledge, and there should also be constant monitoring of the learning process to adapt the guidance required.

Self-regulated learning

Self-regulated learning is only effective if the learners have self-directed learning skills, which means they are able to monitor and control their own learning process. The learners can identify how well they perform a task and whether they are able to choose the appropriate future learning tasks for new learning. However, these are demanding challenges, so in most cases the monitoring and control are performed by the teacher or various design tools.

Psychological literature on cognitive architecture accepts that human knowledge is stored in cognitive schemas. Cognitive load theory (Sweller, Ayres, and Kalyuga 2011) identifies a working memory, which has a limited capacity while processing new information, and unlimited long-term memory, which contains cognitive schemas varying in the degree of richness concerning the number of elements and the connections between the elements. Human expertise, according to the authors, is the result of the availability of rich and automated cognitive schemas, so the processes of schema construction and schema automation are of the utmost importance in learning. The schema construction process encompasses the creation of more complex schemas which incorporate lower-level schemas into higher-level schemas. In such a way, learners who already have relevant cognitive schemas available to incorporate new information, which means learners who have prior knowledge of the subject, can learn more effectively than the ones who do not have the prior knowledge.

Another important factor is schema automation which occurs during the repetitive performance of the tasks. So, according to the theoreticians, learning is the result of schema construction and schema automation. This theory has successfully been applied in the four-component instructional design model (4C/ID) by van Merrienboer and Kirschner (2017) in their research. Their learning environment provides the learners with the tasks and the supportive information to create cognitive schemes, and then the learners perform drill tasks to automate the constructed schemas.

Initial knowledge

The four-component instructional design (4C/ID) model introduced by van Merrienboer, Kirschner and Kester (2003) comprises four components essential for learning, which include: learning tasks, supportive information, procedural information and part-task practice. Each component is accordingly divided into the subcomponents which present the systematic approach to learning. The authors of the 4C/ID model base their design on the assumption of cognitive load theory (Sweller, Ayres, and Kalyuga 2011) that human knowledge is stored in cognitive schemas, and learning processes are closely connected to the construction or reconstruction of human knowledge schemas.

The second component of the 4C/ID model, which is named supportive information, is related to the necessity of sustaining information in order to enhance learning processes. It has a set of subcomponents where one of the elements is prior knowledge. In fact, prior knowledge activation is important in the learning process as it facilitates further development of cognitive schemas based on prior knowledge and stimulates the process of the integration of new knowledge into the existing knowledge foundation. The importance of prior knowledge is discussed by many authors, and various approaches are suggested for how prior knowledge could be activated in the learning process. For example, de Grave, Schmidt and Boshuizen (2001) introduce problem analysis as an appropriate method for the activation of prior knowledge, Machiel-Bongaerts, Schmidt and Boshuizen (1995) suggest mobilization, and Gurlitt et al. (2006) offer concept mapping for activation of prior knowledge.

Van Merrienboer and Kirschner (2013), while conducting their analysis on what builds human expertise, arrived at the conclusion that human expertise is based on the prompt availability of rich and automated knowledge schemas. The schemas ensure the organization and storing of knowledge, so learning may embrace a large number of new elements for a non-experienced person whereas for the experienced person it might be just one element because an experienced person may already have a cognitive schema available which incorporates other elements. As a result, new information is more easily processed by an experienced person with the prior knowledge than by a person without experience.

3.2. Research methodology

The choice of research methodology is presented in this section, which briefly introduces the common move in the research from quantitative research methodologies to qualitative research approaches that bring out the voices of the research participants. Finally, this section provides a detailed description of the research procedures by presenting the research methodology, research stages, sample, data collection and analysis.

3.2.1. Methodological approach

Historically, research methodology, especially in learning environments, has undergone a transition proceeding from a quantitative approach towards a qualitative approach and finally equipping researchers with the ability to successfully apply both approaches as complementing each other. The criticism of the application of purely quantitative methods in educational research features such arguments as the fact that human social life cannot be simply characterized by cause-and-result understanding; what is more, human interactions involve complex processes of negotiation and interpretation, and determined outcomes cannot always be set (Creswell 2007). The author observes that qualitative research is applicable in researching educational processes which are essentially characteristic of ongoing interpretation and interaction, and the application of scientific methods used in quantitative educational research might just lead to standardization.

Qualitative research does not make any assumptions before the research starts, and it focuses on the application of methods which allow us to capture stories of participants' own experience and make it possible to work out the meaning of the research participant experience. For the current research, qualitative inductive content analysis by Elo and Kyngas (2007) has been chosen as a core method in order to structurally analyze teaching and learning experiences while applying corpus analysis and building tools for analyzing textual cohesion through discourse connectives. Qualitative inductive content analysis has been chosen depending on the research question, as the current research is intended to investigate how the participants make sense of teaching and learning while applying corpus analysis and building tools for analyzing textual cohesion using discourse connectives through their own lived experience. The structural analysis of the meaning which research participants give to shared lived experience helps us to examine how things really are and make certain conclusions and recommendations. In education, it could theoretically be known how things should be but it is a sensitive area where regulations and instructions may clash with human realities and research may reveal certain areas for improvement.

3.2.2. Research procedures

The research focuses on the question of what the role of applying corpus analysis and building tools is while teaching and learning a foreign language at more advanced levels. So we performed semi-structured interviews (Ghiglione and Matalon 2001) by asking the following questions:

- How do you perceive the process of applying corpus analysis and building tools for corpus annotation while teaching and learning a foreign language at more advanced levels? What is this process for you?
- How do you perceive the role of applying corpus analysis and building tools for corpus annotation while teaching and learning a foreign language at a more advanced level? How can it help you? How do you use it?
- What do you need in applying corpus analysis and building tools for corpus annotation while teaching and learning a foreign language at more advanced levels? What do you use?
- How do you solve the challenges of using corpus analysis and building tools in teaching and learning a foreign language at more advanced levels? What do you do?
- How could the teaching-learning process be improved by using corpus analysis and building tools for corpus annotation while teaching a foreign language at more advanced levels?

The interview guides were used as flexible tools containing open-ended questions, so they were devised, used, revised and newly devised, which helped us to start each interview and achieve a natural and comfortable conversational flow throughout each interview, as Charmaz (2014, 66) advises to “re-evaluate, revise and add questions throughout the research process”.

The current research is exploratory and descriptive, focusing on the process of teaching and learning a foreign language at more advanced levels while applying corpus analysis and building tools experienced and reflected on by the research participants, and aiming to formulate certain conclusions and recommendations for the application of corpus analysis and building tools for corpus annotation while teaching and learning a foreign language at more advanced levels. A qualitative approach was used in the study in order to conceptualize within the framework of qualitative inductive content analysis the main factors related to the application of corpus analysis and building tools for corpus annotation while teaching and learning a foreign language at more advanced levels. The research is based on a qualitative approach in that we gathered interviews from research participants on their learning experience while applying corpus analysis and building tools for analyzing the cohesion through discourse connectives and comparing them with the original. Two groups of the research participants were exposed to the experience of working with the corpus analysis and building tools annotating and analyzing the cohesion through discourse connectives and comparing them with the original. One group of first-year students majoring in translation and their teacher and a group of the third-year students and their teacher participated in the research, amounting to a research participant number of 26 participants.

Stages of the research

Initial stage: The colleagues agreed to use the software corpus analysis and building tools for annotating discourse connectives as a part of the practical discourse analysis course taught to the first- and third-year students. The students annotated the texts translated by their peers and then analyzed them, comparing them to the original.

Ongoing work: After finishing the course, the research participants were asked to reflect on their experience of using corpus analysis and building tools for annotating discourse connectives and analyzing and comparing the translations with the originals.

Ending: The reflections of the research participants were coded and analyzed using NVivo.

Sample

Case participants were defined by their experience related to the research question. Two groups of students majoring in translation studies were taught by applying corpus analysis and building tools for corpus annotation to raise their awareness of text coherence: one group of the first-year students who just started their studies and had not yet covered a course on syntax and discourse relations within the text, and a group of the third-year students who had covered a course on syntax and discourse relations within the text before. Their teachers were questioned about their observations during the process of teaching/learning by applying corpus annotation. The essential details of the research participants of the study are presented in Table 2.

Table 3-2. Research participants

Student research participants	Study year	Study level	Employed
S1	3	Bachelor's degree	yes
S2	3	Bachelor's degree	no
S3	3	Bachelor's degree	no
S4	3	Bachelor's degree	no
S5	3	Bachelor's degree	yes
S6	3	Bachelor's degree	no
S7	3	Bachelor's degree	yes
S8	3	Bachelor's degree	no
S9	3	Bachelor's degree	no
S10	3	Bachelor's degree	no
S11	3	Bachelor's degree	yes

S12	1	Bachelor’s degree	no
S13	1	Bachelor’s degree	no
S14	1	Bachelor’s degree	no
S15	1	Bachelor’s degree	no
S16	1	Bachelor’s degree	no
S17	1	Bachelor’s degree	no
S18	1	Bachelor’s degree	no
S19	1	Bachelor’s degree	yes
S20	1	Bachelor’s degree	no
S21	1	Bachelor’s degree	no
S22	1	Bachelor’s degree	no
S23	1	Bachelor’s degree	no
S24	1	Bachelor’s degree	no
Teachers			
Research participant	Position	Work experience at university	Level of study programs taught
T1	Lecturer	20	Bachelor’s degree
T2	Lecturer	25	Bachelor’s degree

Data collection

A method of semi-structured interviews for collecting the empirical data was used in the research. The interview method is established as one of the most effective methods, and it is mostly used for collecting data in the qualitative research paradigm (Silverman 2005) as it provides a direct way of obtaining the information about the researched phenomenon. The interview method also enables us to disclose the experience of the research participants expressed in their own words (Kvale 1996; Silverman 2005).

At the beginning of the interview, the interviewees were presented with the aim of the research. Interviewee active participation, objectivity and sincerity were also encouraged during the interviews. Additionally, at the beginning of the interview the topic of the research was introduced and we created an atmosphere of comfort and trust by admitting that the research participants are the experts of their own experience concerning the application of corpus analysis and building tools for corpus annotation while teaching and learning a foreign language at more advanced levels. It was explained that the research question needed to be investigated as fully as possible and that the voices of the research participants are extremely valuable in the research.

According to the established recommendations (Kvale 1996; Silverman 2005), at the end of the interviews the research participants were asked if they wished to elaborate on anything which had not been mentioned or discussed during the interview, and in some cases this helped us collect some additional details. The appreciation and gratefulness for taking part in the research and also for the time spent on the participation in the interview was also expressed. The interviews on average lasted for half an hour (from 30 minutes to 45 minutes), totaling approximately 15 hours of recorded material.

Data analysis

The inductive qualitative content analysis was carried out using NVivo, which is a well-established and efficient software product widely used for organizing and managing data. The researchers instantaneously analyzed the interviews just after they were completed by constantly comparing the structuralized data material. The data have undergone several coding stages, starting with initial open coding, followed by axial coding and selective coding. The data analysis started with a very close inspection at the initial stage, which the researchers later extended into a theoretical stage by applying increasingly more abstract and conceptual processing, ultimately leading to the construction of the core category.

The process of open coding enabled the creation of sub-categories to describe all the aspects of the content. Then the whole pool of sub-categories was grouped under higher order headings; at the same time, memos were written about the conception of the higher order headings leading to higher order categories. The process of formulation of the higher order categories is defined as the procedure of abstraction which continues as long as the core category is identified and methodically related to other categories.

3.3. Research findings

The section of the research findings presents the set of the three broad core categories formulated during the process of structural data analysis—“personal-level factors”, “technology-related factors” and “organizational-level factors”—which are organized in a hierarchical order. The set of the core categories and associated sub-categories describes the necessary conditions, related experiences, and consequences related to the role of applying corpus analysis and building tools for corpus annotation while teaching and learning a foreign language at more advanced levels.

3.3.1. Model of using annotation of discourse relations for teaching and learning a foreign language at more advanced levels

The inductive qualitative content analysis supported by NVivo helped us to build a model of annotating discourse relations for improving teaching and learning a foreign language at more advanced levels, including translation studies. At the beginning, some of the initial codes that emerged from the open coding process allowed us to produce the initial representation of the nodes and divide the application of discourse relation annotation in teaching translation into two big nodes representing the two intertwined processes involved: teaching and learning and annotation procedures which in their own turn are divided into smaller nodes. The teaching and learning node contains: support during the teaching and learning process, motivation, teaching and learning outcomes, teaching and learning challenges, the importance of initial knowledge, and improvement of the teaching and learning process. The annotation procedures node contains such nodes as: using technology, challenges of the annotation process and advantages of the annotation process (see Figure 2).

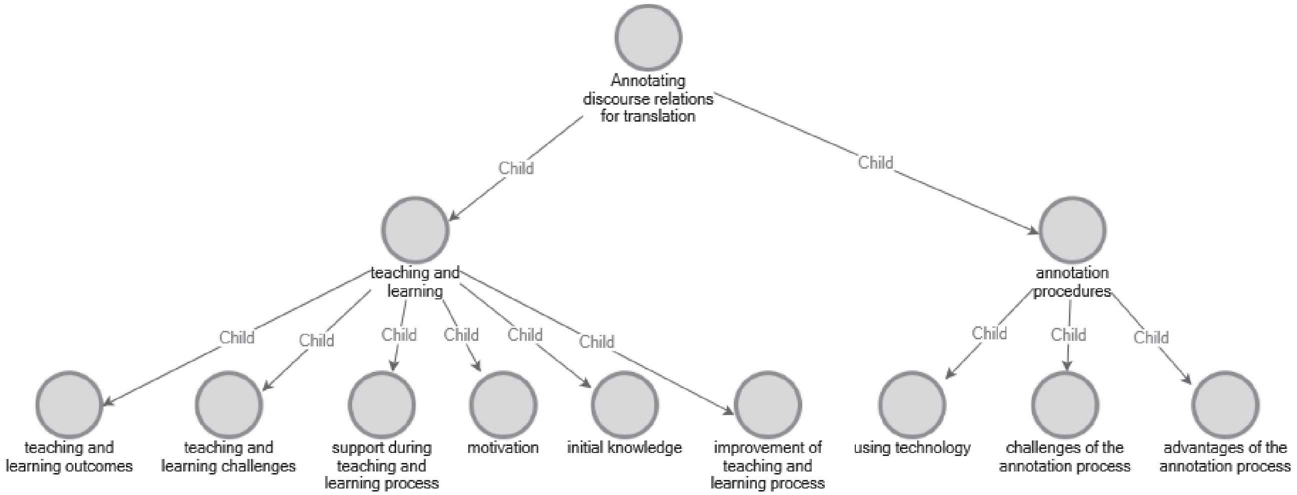
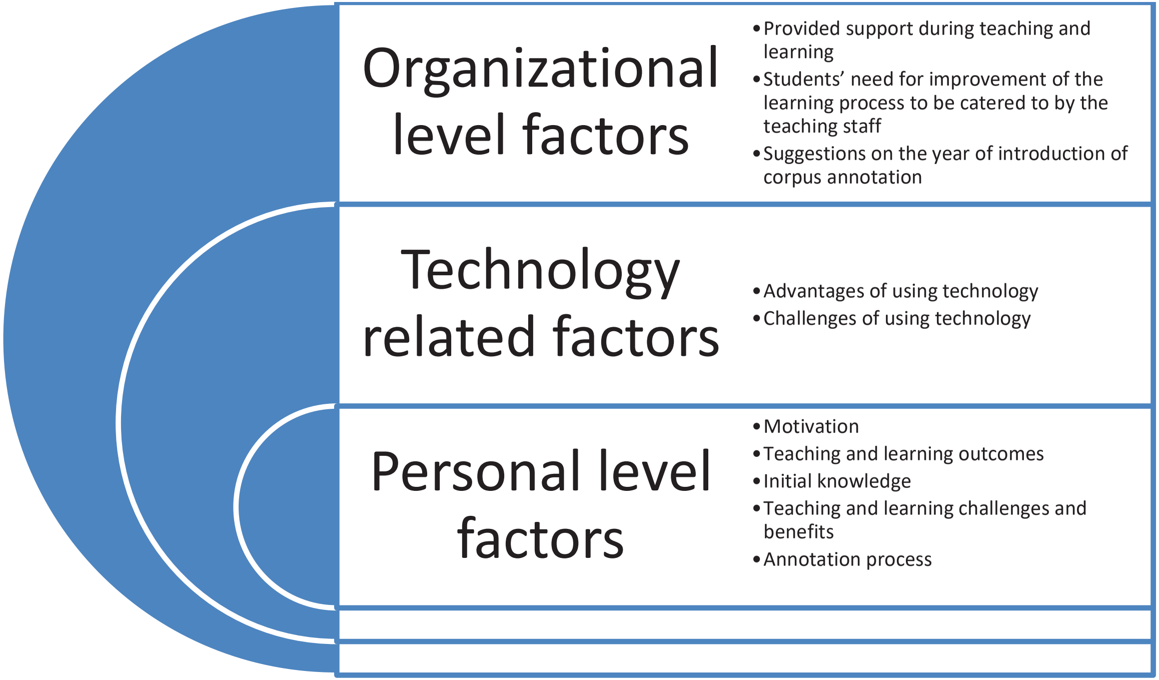

Annotating discourse relations for translation
Child
Child
teaching and learning
annotation procedures
Child
Child
Child
Child
Child
Child
Child
Child
Child
teaching and learning outcomes
teaching and learning challenges
support during teaching and learning process
motivation
initial knowledge
improvement of teaching and learning process
using technology
challenges of the annotation process
advantages of the annotation process

Figure 3-2. Initial representation of the nodes of applying corpus annotation while teaching a foreign language at more advanced levels

A schematic model of annotating discourse relations for improving learning translation is presented in Figure 2 above, and child nodes as main themes are presented and analyzed later in the text.

When all the data were examined, the codes were organized by recurring theme; for instance, there were categories which became prime candidates for stable and common categories, which linked the associated codes. This was the stage of axial coding (Strauss and Corbin [1990] 1998), and it relies on a synthetic technique of making connections between sub-categories to core categories to construct a more comprehensive scheme. This iterative categorization produced a set of three broad core categories of "personal-level factors", "technology-related factors" and "organizational-level factors", which are organized in a hierarchical order. The set of three broad core categories and associated concepts describe the conditions, experiences, and consequences associated with the role of applying corpus annotation while teaching a foreign language at more advanced levels (see Figure 3).



Organizational level factors

- Provided support during teaching and learning
- Students’ need for improvement of the learning process to be catered to by the teaching staff
- Suggestions on the year of introduction of corpus annotation

Technology related factors

- Advantages of using technology
- Challenges of using technology

Personal level factors

- Motivation
- Teaching and learning outcomes
- Initial knowledge
- Teaching and learning challenges and benefits
- Annotation process

Figure 3-3. Applying corpus annotation while teaching a foreign language at more advanced levels

The research reveals that organizational factors (suggestions on the year of introduction of corpus annotation, provided support during teaching and learning, and students' need for improvement of the learning process to be catered to by the teaching staff) are important in applying corpus annotation while teaching a foreign language at more advanced levels. Student research participants were enthusiastic about applying corpus annotation while teaching and learning a foreign language at more advanced levels. They viewed the use of corpus annotation as a useful tool that provided a deeper understanding of textual coherence and cohesion. However, they admitted the importance of prior knowledge in the process of applying corpus annotation while teaching a foreign language at more advanced levels and revealed the difficulties faced while learning without prior knowledge. Later in this study, we analyze all the categories in a more detailed way.

3.3.2. Organizational-level factors

Organizational level factors include three main categories: provided support during teaching and learning, students' need for improvement of the teaching and learning process are catered to by the teaching staff, and suggestions on the year of introduction of corpus annotation. The mentioned categories are analyzed below in a more detailed way.

3.3.2.1. Provided support during teaching and learning

The category of support during teaching and learning contains the subcategories which represent the structured experience of the research participants, revealing the importance of using examples, teacher explanation, learning material, group work and learning by doing.

Table 3-3. Provided support during teaching and learning

Category	Sub-category	Meaning unit
Support during teaching and learning	Using examples	*Before annotating Serbian text I thoroughly worked on English text annotation to see what types of sentences and relations there are. That actually really helped me with Serbian text and after few sentences I actually liked working with the annotator. I was feeling like I know what I was doing. (S1)* *It was good that we had an example of the annotation done. This helped us to see how the junctions of arguments are marked and why. It was easy to track differences between different languages. (S11)*
	Teacher explanation	*It was easy to understand the discourse relation between Explicit and Implicit segments. Because it was well explained by teacher. (S14)* *The most useful things were the slides that our teacher uploaded to Moodle and the material on paper. It was very informative. Also teacher's comments were very helpful as well. (S24)*
	Learning material	*What really helped me to see into the learning process were the slides and the paper material given by the lecturer which was very informative. (S5)*
	Learning by doing	*Also, finding the differences between the relations of the texts got easier with practice. (S5)* *I see that learning by doing is effective in learning annotation. (T1)*
	Group work	*It was useful for students to be divided into groups and then each present a specific connector, or try to think of easy examples. (S23)*

Student research participants observe that examples are really helpful while analyzing the annotation scheme and trying to apply it in one's own language. One of the research participants stresses that examples really helped him to master the process of annotation in his native language: "*Before annotating Serbian text I thoroughly worked on English text annotation to see what types of sentences and relations there are. That actually really helped me with Serbian text and after few sentences I actually liked working with the annotator. I was feeling like I know what I was doing*" *(S1)*.

Also, the research participants discuss the importance of teacher explanation and the learning material prepared by the teacher. They identify that learning material and teacher explanation provide more clarity on the annotation process, as one of the research participants observes: "*The most useful things were the slides that our teacher uploaded to Moodle and the material on paper. It was very informative. Also teacher's comments were very helpful as well*" *(S24)*.

Another important factor is learning by doing, as the research participants observe. Student research participants identify that practice makes it easier to grasp the peculiarities of the process of annotation; for example, "*Also, finding the differences between the relations of the texts got easier with practice.*" *(S5)* It is also observed by teacher research participants that "*leaning by doing is an effective way of learning annotation*" *(T1)*.

Finally, group work is also appreciated by the research participants. Student research participants identify group work as a useful way of leaning annotation: "*It was useful for students to be divided into groups and then each present a specific connector, or try to think of easy examples*" *(S23)*.

3.3.2.2. Improvement of the teaching and learning process

Student research participants also observed their need for the improvement of the learning process, which contains one more category presented in Table 3-4.

Table 3-4. Students' need for improvement of the learning process to be catered to by the teaching staff

Category	Sub-category	Meaning unit
Need for improvement of the learning process	Need for preparation	*I think we need to spend more time on the preparation before starting to annotate. (S19)*
		I think annotation should be used in a teaching/learning process, but only after some preparation and learning about all the different relations. (S13)
	Need for introduction of meta-linguistic knowledge	*I think that at first we should have been introduced to the parts of speech, their names in English, types of conjunctions and how they work in general. (S16)*
		I think it would be useful to have an introduction lecture with detailed explanation of relations, how they differ from each other, how to find and compare them in the text. (S18)
		I would like to get more theoretical knowledge before starting to annotate or before starting any new practice. (S21)
		I think knowledge of basic syntax and semantics is required in combination with sentence relations. *In order to conduct a proper corpus analysis I think I need more basic knowledge, because language is not an easy thing after all. (S17)*

	Need for explanation	*I would like to be given a more detailed explanation on what is annotation itself. (S15)*
		It definitely could be a good way to learn more about conjunctions, their types and about parts of speech; however, everything should be explained in detail. (S19)
	Simpler tasks	*Perhaps instead of doing a whole text from the start students should first be presented with smaller paragraphs and then advance further to bigger discourses. (S12)*
		For instance, annotating only few paragraphs while trying to really mark every relation would make the work less tiring and more efficient. (S22)

The research participants discuss their need for preparation which could make their learning easier; for example, "*I think annotation should be used in a teaching/learning process, but only after some preparation and learning about all the different relations*" *(S13)*. They also stress their need for introduction of meta-linguistic knowledge and need for explanation—both of which, if catered to, could enhance student learning. The research participants identify their preference for simpler tasks as they think that moving from simpler tasks to more complex ones could make their learning more effective: "*For instance, annotating only few paragraphs while trying to really mark every relation would make the work less tiring and more efficient*" *(S22)*. The ideas pointed out by the research participants relate to the theoretical observations mentioned above regarding the necessity of supportive information, procedural information and part-task practice.

3.3.2.3. Suggestions on the study year for the introduction of corpus annotation

The category of research participant suggestions on the study year for the introduction of annotation also demonstrates that students should have some prior knowledge and preparation (Table 5).

Table 3-5. Suggestions on the year of introduction of corpus annotation

Category	Sub-category	Meaning unit
Suggestions on the year of studies to introduce annotation	Suggestions to introduce annotation for second-year students	*I would offer this course in the sophomore years rather than in the freshmen's curriculum. We lack too much the techniques enabling us to fully grasp the material. I would offer this course in the middle of the studies, let us say, end of the second year. (S20)*
		I do not think that is good for the first-course students. What I mean is that we are only the beginners and do not know a lot. From my perspective, it is good for those students, who are not at the start of learning. (S15)
		I mean by that it should be in the third semester because it feels like many people didn't understand it fully and adding it to a second semester may be a bad idea. (S11)
	Suggestions to introduce annotation for third-year students	*I think it should be destined to third-year students who already have a better knowledge of the language rather than first-year students who are still learning a lot of things. (S5)*
		It would be great to have this kind of tasks later, like not on the first course but the third year instead. (S14)
	Suggestions to introduce annotation for master's students	*It depends on the level of the students. Too early for bachelor studies. (T2)*
		Corpus annotation can be a very useful resource for students of linguistics in the last years of bachelor studies and—even better—master studies, maybe a specialized master program. (T1)

The category discloses the information that the first-year student research participants suggest introducing annotation and corpora application later, either in the second or the third year of studies. Their reflections are supported by the reflections of the teacher research participants, suggesting the possibility of introducing annotation and corpora application even to master's students. The suggestions support the above theoretical observations claiming that human expertise is related to the availability of rich and automated knowledge schemas which help an experienced person with the prior knowledge to process new information.

3.3.3. Technology-related factors

It is natural that the category of using technologies appears in coding as the research participants have to use a special software program designed for annotation of discourse relations. It involves the sub-categories of advantages of using technology and challenges of using technology.

Table 3-6. Using Technology

Category	Sub-category 2	Sub-category 1	Meaning unit
Using technology	Advantages of using technology	Usefulness of the program	*I found corpus annotation a very interesting and useful program. A tool that can help us arrange the text and annotate it without much effort because it has all the necessary options. (S1)*
		Easiness of use	*The program is also very easy to use, without any unnecessary things. (S5)*
		Clarity	*With the annotator and different colours a student can clearly see from an example how a certain sentence is divided and just work on that model. (S9)*
		Interest	*It was a bit interesting to see a program for annotation because I have not seen one yet. (S6)*

		Novelty	*It was a great experience to familiarize myself with the Java program which shows different annotation processing tools and to get acquainted with the vast majority of basics while annotating. (S9)*
	Challenges of using technology	Confusing software features	*Also, at the beginning of my work with the annotation it was a bit confusing with all the colours and other options we were supposed to put in about one relation. (S21)*
		Problematic managing of the annotation software	*I had a problem with a few sentences since the programme didn't accept my annotation so didn't really know what to do. (S23)*
		Initial difficulties	*It was my first time when I worked with such a program and it was quite difficult to me. The reason why it was difficult was because it was my first time working with such a program. (S22)*
		Switching between languages	*I have to admit that it was a bit confusing for me to switch from one text to another and find the sentences in two different languages because sometimes they were not in the same positions in the annotator. (S1)*
		Need to memorize software abbreviations	*To me, the memorization of the abbreviations was rather confusing (Explicit, implicit, and all the acronyms on word co-relation.) (S11)*

		Low-quality public computers	*I was feeling a bit anxious when I went to library to finish my annotation because computers there are very slow and it was very inconvenient to use the corpora. It is easier to work with my own laptop, instead of university computers, because all information is always with me. (S19)*
		Need for technical support	*The main challenge was to start the software working. I remember the first day starting to work with the students, the program didn´t function properly and it took an hour and a half and I still couldn´t have the annotation program working. I consulted the IT specialists but it also was very little help, so next day I spent about three hours and finally found the solution, so next lecture the annotation program worked ok. But there was another problem to explain to the students how to use the program and to help them to use the annotation scheme. (T2)*
		Need for knowledge of the annotation software	*I need more knowledge of how to deal with new software so that I could feel confident and help my students. The institution leaves the lecturer on his/her own to grasp new technologies which in many cases are so frustrating. (T2)*

The category of advantages of using technology contains the sub-categories clarity, novelty, interest, easiness of use and usefulness of the software. The research participants identified traditional advantages provided by technologies and find using technologies really beneficial: "*I found corpus annotation a very interesting and useful program. A tool that can help us arrange the text and annotate it without much effort because it has all the necessary options*" *(S1)*.

However, the research participants point out the challenges of using the software for annotation. Some research participants still find some software features confusing: "*Also, at the beginning of my work with the annotation it was a bit confusing with all the colours and other options we were supposed to put in about one relation*" *(S21)*. And they find it problematic to manage the software appropriately: "*I had a problem with a few sentences since the programme didn't accept my annotation so didn't really know what to do*" *(S23)*.

Other research participants stress the clarity, though, leading to the observation that different people have different software managing skills and some need more practice with new software, which is revealed by the sub-theme of initial difficulties: "*It was my first time when I worked with such a program and it was quite difficult to me. The reason why it was difficult was because it was my first time working with such a program*" *(S22)*. In addition, the software features and the annotation process itself pose such challenges as difficulties switching between languages as the sentences move to different positions in the tool—"*I have to admit that it was a bit confusing for me to switch from one text to another and find the sentences in two different languages because sometimes they were not in the same positions in the annotator*" *(S1)*—and the need to memorize software abbreviations and know the meanings of the acronyms—"*To me, the memorization of the abbreviations was rather confusing (Explicit, implicit, and all the acronyms on word co-relation)*" *(S11)*.

Student research participants also speak about the low quality of the public computers, which makes the students spend extra time on annotation as the computers are slow or do not work properly: "*I was feeling a bit anxious when I went to library to finish my annotation because computers there are very slow and it was very inconvenient to use the corpora. It is easier to work with my own laptop, instead university computers, because all information is always with me*" *(S19)*. This sub-theme shows that universities should always consider a constant renewal of the technical resources as a need for it is caused by the constant development of technology and the surplus of new knowledge which requires modern technological solutions.

Teacher research participants mention the need for technical support as they need help while implementing the newest technologies for teaching and learning: “*The main challenge was to start the software working. I remember the first day starting to work with the students, the program didn´t function properly and it took an hour and a half and I still couldn´t have the annotation program working. I consulted the IT specialists but it also was very little help, so next day I spent about three hours and finally found the solution, so next lecture the annotation program worked ok. But there was another problem to explain to the students how to use the program and to help them to use the annotation scheme*” *(T2)*. They also feel that they need knowledge about annotation software because if teachers want to be able to use the newest technologies for teaching and learning they also need certain training and knowledge, but usually they are left to learn everything on their own: “*I need more knowledge of how to deal with new software so that I could feel confident and help my students. The institution leaves the lecturer on his/her own to grasp new technologies which in many cases are so frustrating*” *(T2)*.

3.3.4. Personal-level factors

The core category of personal-level factors appears to be the most populated category; it includes five main categories (motivation, teaching and learning outcomes, initial knowledge, teaching and learning challenges and benefits, and annotation procedures) which consist of smaller sub-categories in turn.

3.3.4.1. Motivation

The node/category of motivation includes two main categories of encouraging motivation and discouraging motivation.

Table 3-7. Motivation

Category	Sub-category	Meaning unit
Encouraging motivation	Success	*The positive memory I have of annotating is when I was annotating a sentence by myself using all the elements I could and that when I compared my annotation with the English text, it was matching and coherent. I felt like I was able to understand how to do it. (S3)*
	Relating theory to practice	*When we started to translate the texts and when we started to learn about annotation itself, we learned to use our theoretical knowledge in practice. I felt content, because I was learning about corpus linguistics, something that is linked with my course. (S21)*
	Interest	*The functions the sentences performed such as reason, result, cause or others were also interesting to analyze. Additionally, in the implicit cases we were supposed to write the implicit conjunction which was supposed to be used in the blank spot and that was also interesting. (S16)*
	Familiarity with the topic	*I was already familiar with a similar topic so the feeling present was nostalgia and happiness to be able to do it again because in the entire education I would single out this to be the matter I most enjoyed dealing with. (S15)*

	Novelty	*When I first heard we will be doing a different project I was interested in learning something new. It was interesting to work with a new programme. (S23)*
Discouraging motivation	Lack of examples	*If the annotation is firstly presented with already done examples of corpus annotation it would be easier for the students to learn it at the beginning because the examples would serve as guidelines. (S15)*
	Mistakes	*And the text I was annotating was not translated well, as there were a lot of grammar and vocabulary mistakes, so sometimes it was very difficult to annotate because sentences did not make any sense in Russian. (S19)*
	Fear of negative assessment	*I really wanted to stop doing this. But at the same time I wanted to do this project because I didn't want to get a negative mark. (S24)*
	Difficulty in understanding	*The negative memory of this annotation was to annotate myself and find out by comparing that it did not correspond to the English annotation; there was something wrong and it gave me the feeling that I could not understand the annotation process. (S3)* *Unfortunately, I did not have many positive experiences while annotating the text, since most of the time I had no idea what I was doing and the task itself wasn't explained well enough to fully understand it. I did not feel well while annotating the text because I was very lost the whole time. (S17)*

Student research participants reveal that motivation is a very important factor in the teaching and learning process and the sub-categories of the main categories of encouraging and discouraging motivation disclose the influential elements shaping student motivation. The category of encouraging motivation embraces the sub-categories of success, the possibility of relating theory with practice, familiarity with the topic, interest and novelty.

Research participants identified success as an important driving force of motivation, as one of the research participants says: "*The positive memory I have of annotating is when I was annotating a sentence by myself using all the elements I could and that when I compared my annotation with the English text, it was matching and coherent. I felt like I was able to understand how to do it*" *(S3)*.

Another motivating element is relating theory to practice, which gives a rewarding feeling of being content with learning: "*When we started to translate the texts and when we started to learn about annotation itself, we learned to use our theoretical knowledge in practice. I felt content, because I was learning about corpus linguistics, something that is linked with my course*" *(S21)*.

Interest is also mentioned by the research participants as a motivating element for learning since the new acquired knowledge builds students' interest. As one of the research participants observes, "*The functions the sentences performed such as reason, result, cause or others were also interesting to analyze. Additionally, in the implicit cases we were supposed to write the implicit conjunction which was supposed to be used in the blank spot and that was also interesting*" *(S16)*. What is more, novelty is related to inducing interest and motivation to learn. The research participants disclose that they were interested in working with new software and learning new knowledge: "*When I first heard we will be doing a different project I was interested in learning something new. It was interesting to work with a new programme*" *(S23)*.

Familiarity with the topic may also increase motivation by bringing enjoyment in studying the topic more. The research participants felt happy to be able to work on a familiar topic again: "*I was already familiar with a similar topic so the feeling present was nostalgia and happiness to be able to do it again because in the entire education I would in single out this to be the matter I most enjoyed dealing with*" *(S15)*.

The category of discouraging motivation includes the sub-categories of lack of examples, mistakes, fear of negative assessment and difficulty in understanding.

The sub-category of lack of examples reveals an interesting observation that sometimes providing a few examples is not enough and some research participants feel that they need examples which could enhance their understanding: "*If the annotation is firstly presented with already done examples of corpus annotation it would be easier for the students to learn it at the beginning because the examples would serve as guidelines*" *(S15)*. It seems that it could be a good idea to provide more examples while introducing the annotation, taking into account that some research participants identified their need for more examples.

Another discouraging element is mistakes in the text which was being annotated. Some mistakes may even lead to the misunderstanding of the text, as one of the research participants observes: "*And the text I was annotating was not translated well, as there were a lot of grammar and vocabulary mistakes, so sometimes it was very difficult to annotate because sentences did not make any sense in Russian*" *(S19)*. Naturally, the mistakes appeared due to the fact that the texts were translated by the students' peers learning the annotation. However, it is worthwhile considering choosing some original texts for teaching and learning annotation, and thereby to avoid the issue of mistakes.

The sub-category of fear of negative assessment raises an ongoing question of motivating students through punishment or reward. The case reflected by research participant S24 can hardly be considered positive as the main motivation to continue the annotation process was reluctance to get a negative grade: "*I really wanted to stop doing this. But at the same time I wanted to do this project because I didn*'*t want to get a negative mark*" *(S24)*.

Finally, speaking about discouraging motivation we arrive at difficulties in understanding, which—depending on the reasons—serve as a discouraging leverage for continuing the task and achieving the results. As can be seen in the reflections of the research participants, difficulties in understanding cause negative emotions lingering for a long time even after the task is finished: "*The negative memory of this annotation was to annotate myself and find out by comparing that it did not correspond to the English annotation; there was something wrong and it gave me the feeling that I could not understand the annotation process*" *(S3)*.

This feeling of inability to understand does not enhance the quality of teaching and learning. The research participants recognized that difficulties in understanding leave them with the feelings of negative experience and feeling lost not knowing what they were doing.

Overall, reflection on encouraging and discouraging motivation leads to the conclusion that providing enough examples and making sure that students experience success in grasping the annotation task and performing it appropriately should be kept in mind while teaching and learning annotation.

3.3.4.2. Teaching and learning outcomes

The node of teaching and learning outcomes comprises such major categories as acquired professional knowledge and skills, which, in their own turn, branch into a number of categories and sub-categories.

3.3.4.2.1. Acquired professional knowledge

The major category of acquired professional knowledge contains the categories of insights on translation and linguistic awareness.

Table 3-8. Teaching and learning results

Category	Sub-category	Meaning unit
Insights on translation	Using dictionaries	*When translating the text there were some words in English for which I couldn't find the equivalent in the Serbian language so I had to use OALD. It was very useful for me because it gave the exact sense of the word in Serbian that I needed. Also the programme itself is more than useful with all the options where you just need to divide the relations. (S1)*
	Experiencing the process of translation	*It was a nice experience to understand what it feels like to be a translator. Also I got experience in writing and translation. (S2)*
	Understanding comparative-contrastive use	*Also, with annotated texts I was able to compare the texts easier and find some differences and similarities in the languages. Now I know that we should pay more attention to the way sentences are constructed, perhaps now while writing or translating a text I will pay attention to the differences between languages. (S1)*
	Editing awareness	*I learned that while translating people cannot avoid errors; therefore, annotations are useful for proof of somebody's state-of-the-art translation. (S11)* *It could be useful in teaching translation because it reveals the drawbacks of translation and gives hints of how the texts could be translated in a better way. (T2)*
	Cultural awareness	*This experience helped me understand that text analysis is not only a technical matter but also deals a lot with the cultural background of a translator. (S11)*

	Accuracy	*I became more attentive to the translation accuracy. This is a motivational experience. I started looking at the text with more care. (S11)*
Linguistic awareness	Understanding semantics	*It develops our skills as professional translators and it helps a lot even to understand the semantics of the language and that is why we can easily translate without linguistic misunderstandings. (S12)*
	Understanding syntax	*I actually understood the text even better (the syntax and everything) after annotating and was able to recognize how sentences are connected in both Serbian and English. (S1)*
	Linguistic knowledge	*Also for my students it develops their linguistic knowledge in the subject. (T2)*
	Mastery of discourse cohesion	*Also, I learned about sentence relations and different kinds of conjunctions that connect sentences that seem not connectable. It made me realize that each sentence is connected on a deeper level and that there are so many types of relations between sentences, which I did not know about. (S10)*
	Deeper linguistic textual understanding	*I started to read different texts more attentively. Sometimes, I even analyze texts unconsciously and it makes me feel language better. It helped me to understand and analyze the texts more deeply. (S5)*

Research participants stated their views in the category of insights on translation, which includes such sub-categories as experience of the process of translation, understanding comparative-contrastive use of languages, accuracy, editing awareness, using dictionaries and cultural awareness. The research participants observed that they experienced the process of translation and had a feeling of what it is to be a translator: "*It was a nice experience to understand what it feels like to be a translator. Also I got experience in writing and translation*" *(S2)*. They realized the importance of comparative-contrastive understanding of source and target languages and understood that editing is an essential part of the translation process: "*I learned that while translating people cannot avoid errors; therefore, annotations are useful for proof of somebody's state-of-the-art translation*" *(S11)*. Teacher research participants also pointed out the importance of editing and looking for how the text could be translated in a better way: "*It could be useful in teaching translation because it reveals the drawbacks of translation and gives hints of how the texts could be translated in a better way*" *(T2)*. The research participants stress that through the process of annotation they found out the importance of cultural awareness: "*This experience helped me understand that text analysis is not only a technical matter but also deals a lot with the cultural background of a translator*" *(S11)*. The sub-categories have certain connotations to the theoretical models of translator competence which involve the contrastive-comparative competence of source and target languages, cultural competence of both languages and editing competence.

The category of linguistic awareness also has some connotations to linguistic competences in source and target languages. It involves the sub-categories of understanding semantics and syntax, linguistic knowledge, deeper linguistic textual understanding, mastery of discourse cohesion and analytical abilities. The research participants revealed that understanding semantic, syntactic and discourse layers of the text enabled them to become better translators: "*I started to read different texts more attentively. Sometimes, I even analyze texts unconsciously and it makes me feel language better. It helped me to understand and analyze the texts more deeply*" *(S5)*. Teacher research participants acknowledged that the students develop linguistic knowledge necessary for translation during the annotation process: "*Also for my students it develops their linguistic knowledge in the subject*" *(T2)*. In addition, student research participants revealed that the development of analytical abilities makes them better translators: "*The useful part of corpus annotation in general is that you can get the best insight into sentence structure and explore syntax, semantics and cohesion in a most thorough way. This analysis can help you to understand the structure and system of sentences*" *(S15)*.

3.3.4.2.2. Skills

The major category of skills includes the categories of acquired skills reflected on by the research participants and necessary skills generally needed during the studies identified by the research participants.

Acquired skills

The category of acquired skills embraces such sub-categories as independent studies, independent research, analytical abilities or analytical thinking, dealing with novelty, attentiveness, time management, linguistic skills, patience, managing technology and communication skills. It could be observed that most of the acquired skills mentioned by the research participants are general skills relevant to any study program or specialty, and only linguistic skills are closely related to the research participants' study program of translation and editing.

Table 3-9. Acquired skills

	Sub-category	Meaning Unit
Acquired skills	Independent studies	*It made me read more extra materials about the translation topic and I found an interesting reading on this topic:* Developing Linguistic Corpora: a Guide to Good Practice *by Geoffrey Leech, Lancaster University, 2004. (S11)*
	Independent research	*I spent a lot of time doing my own research, which encouraged me to spend more time learning by myself. (S17)*
	Analytical abilities	*The useful part of corpus annotation in general is that you can get the best insight into sentence structure and explore syntax, semantics and cohesion in a most thorough way. This analysis can help you to understand the structure and system of sentences. (S15)* *The analytical abilities to understand the construction of sentences (S23)* *Also, you spend a lot of time annotating and analyzing texts, but then you have check it again, and of course, you find mistakes, because*

		sometimes sentences can be annotated similarly, but meanings of them differ, and you have to think, why. (S4) *The students are learning to stop and think. (T1)*
	Dealing with novelty	*It was completely new for me. So basically everything we did during the annotation project brought new knowledge to me. (S3)*
	Linguistic skills	*Translation skills, grammar skills, writing skills (S2)* *It develops our skills for translation. (S22)* *Logical thinking, concentration, improves translation and syntactic skills and knowledge (S5)*
	Attentiveness	*The use of corpus annotation develops attentiveness (S10)* *It has increased my attention to details because every sentence can have different ways of annotating it. (S20)*
	Patience	*It can help the process of developing more patience, more insight and can improve the way we look at work as a whole. (S20)*
	Managing technology	*I learned many things, and first and foremost is how to use the annotator programme. (S1)* *Corpus annotation develops our computer skills.* *It develops the new skills of using the software. (T2)*
	Communication skills	*The whole learning experience taught me communicating with people. (S8)*

Firstly, the sub-categories of independent studies, independent research, analytical abilities and dealing with novelty could be related to the process of learning to learn. The research participants admitted that they had to deal with novelty, which directed them to acquiring a lot of new knowledge; some of them even stressed that everything they learned during the annotation was new to them: "*It was completely new for me. So basically everything we did during the annotation project brought new knowledge to* *me*” *(S3)*. In the process of dealing with novelty, the research participants observed that they also studied independently looking for extra material; they reflect that in this way the novelty and challenges made them search for reading and study material on their own: “*It made me read more extra materials about the translation topic and I found an interesting reading on this topic:* Developing Linguistic Corpora: a Guide to Good Practice *by Geoffrey Leech, Lancaster University, 2004*” *(S11)*. In addition, the research participants carried out their own research, which they acknowledged lead them to developing their independent studies: “*I spent a lot of time doing my own research, which encouraged me to spend more time learning by myself*” *(S17)*. The research participants also had to analyze the study material and develop their analytical thinking and analytical abilities.

The research participants also mentioned their acquired linguistic skills as they are closely related to their study program and are of the utmost importance. The research participants observed that they acquired or improved their “*Translation skills, grammar skills, writing skills*” *(S2)*.

Then the research participants expressed their impressions that attentiveness, patience and time management are the skills which they practiced most during learning annotation. They also mentioned managing technology and communication skills, which they learned during group work.

Analytical skills and managing technology skills were discussed by student and teacher research participants, which might show that teachers pay attention to students’ analytical abilities: “*The students are learning to stop and think*” *(T1)*. Acknowledging analytical thinking is one of the most important skills in the study process. Speaking about managing technology, some student research participants identified that it was one of the easiest parts of learning annotation; however, there are students who identified occasional struggling with managing technologies and see it as a new skill: “*I learned many things, and first and foremost is how to use the annotator programme*” *(S1)*. Meanwhile, teacher research participants considered technology management as an important skill: “*It develops the new skills of using the software*” *(T2)*.

Necessary skills

While talking about the acquired skills, many research participants digressed into reflecting on the necessary skills they believe they need to acquire during their study process.

Table 3-10. Necessary skills

Category	Sub-category	Meaning unit
Necessary skills	Linguistic skills	*We need a high level of spoken and written English, more structure and academic vocabulary (S10)*
	Time management	*We need time management (S14)*
	Communication skills	*What we need is good communication skills. (S9)*
	Information search and processing	*Knowing how to quickly and purposefully find needed information is a very useful skill to have (S17)*
	Conducting research	*I need to learn working independently, conducting linguistic research and thesis writing skills (S18)*
	Managing technology	*I also need skills in the usage of Microsoft Office programs and other programs. (S19)*
	Patience	*I should get more patience. (S20)*
	Attentiveness	*I should pay more attention to details. (S21)*

Surprisingly enough, most of the necessary skills mentioned coincide with the acquired skills, which allows us to conclude that the teaching and learning process satisfies students' needs.

Firstly, research participants stress that they need linguistic skills as they expect to acquire them during the study program: "*We need a high level of spoken and written English, more structure and academic vocabulary*" *(S10)*. The research participants also mention that they need to acquire time management skills, which are closely related to patience and attentiveness. Some research participants feel that they need managing technology skills and communication skills.

Research participants also spoke about the necessary skills which have not been discussed as the acquired ones. Here the research participants stressed the importance of information search and processing skills and conducting research, which are naturally needed during their studies: "*Knowing how to quickly and purposefully find needed information is a very useful skill to have (S17)*"; "*I need to learn working independently,* *conducting linguistic research and thesis writing skills (S18)*". It could be seen that the abilities to process information and carry out independent research are valued by the research participants, which is natural in the academic environment.

One of the research participants mentions almost all the necessary skills in their reflections: "*In my ongoing learning process I need my knowledge of English and other foreign languages (Korean and Norwegian), also my skills in the usage of Microsoft Office programs. I also have to know how to choose reliable sources on the Internet and how to search for the information properly. Time management is also very important because I need to distribute my time mindfully between all the subjects and extracurricular activities I have*" *(S19)*.

3.3.4.3. Initial knowledge

The majority of the students who already had some prior knowledge before starting the annotation process stressed the advantages of having and applying their prior knowledge. They observed that it made their learning process easier, upgraded their prior knowledge, was helpful in the process of annotating the discourse relations, and enhanced their insights while comparing and contrasting discourse relations in different texts. Table 11 reveals the category of prior knowledge benefit, which includes such sub-categories as easier learning process, helping the annotation process, and helping the process of comparing and contrasting discourse relations.

Table 3-11. Prior knowledge benefit

Category	Sub-category	Meaning unit
Prior knowledge benefit	Easier learning process	*After working on a few sentences it was way easier for me since I have some previous knowledge from syntax and Transformational Generative Grammar that was useful for annotation. (S9)*
		Also I understood that it is easier if the students have some background syntax and pragmatics knowledge, otherwise it is very difficult for them to manage the annotation scheme. (T1)

	Helping the annotation process	*Since I have encountered similar processes of sentence analysis in Transformational Generative Grammar Analysis, I found the process of annotation very interesting. (S1)*
		My syntax knowledge was of great help while working on the annotation project since I knew where to divide sentences and how they were connected. (S5)
		I could distinguish the connectors I already have known; for instance: conjunction, condition, contrast, comparison. (S4)
	Helping the process of comparing and contrasting coherence relations	*Sometimes, my linguistic knowledge and skills help me with my studying when I need to find something or contrast and compare. (S2)*
	Upgrading previous knowledge	*I upgraded my previous knowledge about sentence structures and got an insight into sentential meanings connected to cohesion and semantics. (S10)*
		The most positive thing was remembering the classification of different connectors, as we were learning it before. (S7)

In their reflections, the research participants revealed that having prior knowledge about syntax and textual coherence makes their learning process easier; for example, one of the research participants says, "*It was way easier for me since I have some previous knowledge from syntax and Transformational Generative Grammar that was useful for annotation*" *(S9)*. This view is also supported by the opinion of a teacher research participant: "*It is easier if the students have some background syntax and pragmatics knowledge; otherwise it is very difficult for them to manage the annotation scheme*" *(T1)*.

An additional point revealed by the research participants shows that prior knowledge helps the annotation process. The research participants stress that the syntax knowledge helped them a lot while working on the annotation project as they knew how to divide sentences and how the sentences were related.

The research participants also stressed that prior knowledge helped during the process of comparing and contrasting coherence relations; for instance, one of the research participants reveals: "*My linguistic knowledge and skills help me with my studying when I need to find something or contrast and compare*" *(S2)*.

The category of difficulty without prior knowledge shows that the absence of prior knowledge, on the contrary, was the cause of the learning process becoming more difficult. The category contains such sub-categories as difficulty in understanding, more learning and novelty of the information (Table 12).

Table 3-12. Difficulties caused by the lack of prior knowledge

Category	Sub-category	Meaning unit
Difficulties without prior knowledge	Difficulty in understanding	*I had no idea what certain parts of speech were called in English and the manuals did not help since the information there wasn't explained in a simple way. (S12)*
	More learning	*It was hard to understand the main concept of annotation, because we had a lot of information to learn and remember. (S15)*

		It was difficult, because I did not know them before and I needed to learn various types of relations in the sentences, phrases or between them. (S20)
	Novelty of the information	*The whole text annotation thing was new for me and it was far from easy. (S22)*
		It was difficult to understand all the terms and relations between segments because I never used this corpus analysis before. (S11)
		It was difficult, because we never had such practice before. (S14)

Research participants observed that without prior knowledge they ran into certain difficulties. First, they faced difficulty in understanding and had to put more efforts into learning. One of the research participants reveals: "*It was hard to understand the main concept of annotation, because we had a lot of information to learn and remember*" *(S15)*.

The research participants identified that the novelty of the information also caused difficulty; for example, "*The whole text annotation thing was new for me and it was far from easy*" *(S22)*.

3.3.4.4. Teaching and learning challenges and benefits

The category/node of teaching and learning challenges is related to the categories/nodes of initial knowledge and improvement of the teaching and learning process as the latter nodes provide additional insights by the research participants as to how teaching and learning challenges could be overcome.

Table 3-13. Teaching and learning challenges

Category	Sub-category	Meaning unit
Teaching and learning challenges	Time-consuming	*It takes quite a lot of time to analyze each sentence. (S10) It takes quite a lot of time, so I worked day and night. (S2)*
	Stressful	*It took a long time and a lot of nerves. (S22)*
	Mastering new software	*New technologies bring new challenges for me because I need to train myself and I need time and sometimes advice on how to do everything. (T2)* *I think when introducing a new program, students should be given more time to understand fully how to use it as it heavily impacts their annotation work. (S6)*
	Large amount of new information	*At the beginning it was a bit confusing because you get a lot of information at once and I simply didn't know what to do, a lot of terms and things to understand, but it happened only at the beginning. (S3)*
	Linguistic differences	*I think that the hardest part was the comparison of English and Ukrainian annotations because they are completely different languages. (S12)*
	Extra efforts	*Nothing about this was easy; everything required extra work and research at home. (S17)*
	Complexity	*The most confusing part was understanding how to annotate in accordance with a complicated classification of connectors. Of course, the manuals were provided but still I found it difficult to distinguish what I need to write. The most difficult challenge was to understand and apply the types of relations, since there are many different variants of them and it took a lot of time to figure out which ones to use. (S5)*

The research participants named such challenges of the annotation process as the fact that it is time-consuming and stressful, also that it requires mastering new software, putting in extra efforts, dealing with a large amount of new information, and that it involves linguistic differences and complexity. Research participants identified that at the beginning they needed to cope with the large amount of new information, which was quite confusing: "*At the beginning it was a bit confusing because you get a lot of information at once and I simply didn't know what to do, a lot of terms and things to understand, but it happened only at the beginning*" *(S3)*. In addition, they needed to master new software, which required additional time: "*I think when introducing a new program, students should be given more time to understand fully how to use it as it heavily impacts their annotation work*" *(S6)*. Teacher research participants also admitted that before teaching they needed to master the new software as well, which required additional teacher preparation time: "*New technologies bring new challenges for me because I need to train myself and I need time and sometimes advice on how to do everything*" *(T2)*. So the two sub-themes lead to other sub-themes, such as time-consuming and extra efforts, because naturally acquiring new information and mastering software lead to devoting one's time—"*It takes quite a lot of time to analyze each sentence (S10)*"; "*It takes quite a lot of time, so I worked day and night*" *(S2)*—and extra efforts—"*Nothing about this was easy, everything required extra work and research at home*" *(S17)*. As the research participants noticed, it may seem stressful at times: "*It took a long time and a lot of nerves*" *(S22)*. Also, research participants mentioned the complexity of the whole annotation and mastering discourse relations process: "*The most confusing part was understanding how to annotate in accordance with a complicated classification of connectors. Of course, the manuals were provided but still I found it difficult to distinguish what I need to write. The most difficult challenge was to understand and apply the types of relations, since there are many different variants of them and it took a lot of time to figure out which ones to use*" *(S5)*. Linguistic differences also added to the complexity as analyzing and comparing different languages also poses a certain challenge: "*I think that the hardest part was the comparison of English and Ukrainian annotations because they are completely different languages*" *(S12)*.

Benefits of teaching and learning corpus annotation

The category of benefits of teaching and learning corpus annotation in a way represents recommendations provided by the research participants and their insights into why teaching and learning corpus annotation presents value for the research participants. This category contains the sub-categories of teaching and learning linguistic knowledge, raising awareness of linguistic differences, raising translation awareness, developing analytical skills and usefulness for research.

Table 3-14. Benefits of teaching and learning corpus annotation

Category	Sub-category	Meaning unit
Benefits of teaching/learning corpus annotation	Teaching/learning linguistic knowledge	*Corpus annotation should most definitely be used in both teaching and learning processes because it enables the best way of teaching sentence structure in connection to syntax, semantics and cohesion. (S15)*
	Raising awareness of linguistic differences	*It can really help me to find the differences in language clearly. It might be useful because it gives a deeper understanding of how languages work. (S16)*
	Raising translation awareness	*In my opinion, annotation should be used to show how translation and original texts can differ. (S4)*
	Useful for research	*Corpus annotation is also a very useful tool for research. (S2)*
	Developing analytical skills	*It helps to understand and analyze the texts more deeply; (S7)*

The research participants acknowledged that teaching and learning corpus annotation equips them with the linguistic knowledge of syntax, semantics and discourse structure: "*Corpus annotation should most definitely be used in both teaching and learning processes because it enables the best way of teaching sentence structure in connection to syntax, semantics and cohesion*" *(S15)*. They also see the usefulness of the process in getting a more rounded understanding of the differences between languages: "*It can really help me to find the differences in language clearly. It might be useful because it gives a deeper understanding of how languages work*" *(S16)*. As the study program of the research participants is translation and editing, they stress the usefulness of raising translation awareness: "*In my opinion, annotation should be used to show how translation and original texts can differ*" *(S4)*. The research participants also acknowledge its usefulness for research and learning analytical skills: "*Corpus annotation is also a very useful tool for research (S2)*"; "*It helps to understand and analyze the texts more deeply*" *(S7)*.

3.3.4.5. Advantages and challenges of the annotation process

The category/node of the annotation procedures contains three categories/nodes closely related to the procedures of annotation, which include the nodes of using technology during the annotation (which has been discussed above in the section on technology-related factors) and the two categories discussed here, including the challenges and advantages of the annotation process itself.

Advantages of the annotation process

The category of the advantages of the annotation process involves the sub-categories of easily distinguishable discourse relation and value of practice, which in turn are broken into smaller sub-categories.

Table 3-15. Advantages of the annotation process

Category	Sub-category 2	Sub-category 1	Meaning unit
Advantages of the annotation process	Easily distinguishable discourse relations	Processing first-level classification of connectives	*The easy part for me personally was to distinguish between the explicit, implicit, NoRel and EntRel features of sentences. (S15)*
		Processing explicit connectives	*Explicit connectives were easy to distinguish because they are clearly visible in the text. (S18)* *The easy thing to me was to annotate when a connector was obvious and explicit in a sentence. (S3)*
		Processing equivalent connectives	*When I annotated it was easy to do it when relations in the translated text and in the original were the same. (S18)*
	Value of practice	Practice provides ease	*At first it was hard to find differences in annotated texts, but with more practice it became easy. (S7)*
		Practice provides clarity	*At first it was very confusing, but after some time of working with this annotation, it became clearer. (S10)*
		Practice leads to analysis	*I compared and analyzed differences in the annotated translated version and the original. (S13)*

In the sub-category of easily distinguishable relations, we find that according to the research participants it is easy to process first-level classification of discourse connectives as this level includes the major groups of connectives such as explicit, implicit, no relations and entity relations: "*The easy part for me personally was to distinguish between the explicit, implicit, NoRel and EntRel features of sentences*" *(S15)*. Also, it is easier to process explicit discourse relations as the connectives are explicitly expressed in the text: "*Explicit connectives were easy to distinguish because they are clearly visible in the text*" *(S18)*; "*The easy thing to me was to annotate when a connector was obvious and explicit in a sentence*" *(S3)*. In addition, the process of annotation is easier in the cases when the connectives are equivalent in the original text and the translated one; then there is a possibility to compare and reflect on discourse relations in the texts: "*When I annotated it was easy to do it when relations in the translated text and in the original were the same*" *(S18)*.

The sub-category of value of practice leads to the observation of the research participants that annotation practice leads to better analysis of the text and equips the students with more clarity and ease of understanding the discourse relations. The research participants claim that, "*At first it was hard to find differences in annotated texts, but with more practice it became easy*" *(S7)*; "*At first it was very confusing, but after some time of working with this annotation, it became clearer*" *(S10)*; "*I compared and analyzed differences in the annotated translated version and the original*" *(S13)*.

Challenges of the annotation process

The category of the challenges of the annotation process involves the sub-categories of confusing discourse relations, the necessity of experience with the system of discourse relations and the annotation process being time-consuming. The sub-categories split into smaller sub-categories which indicate the subtleties of the research participant insights on annotating discourse relations.

Table 3-16. Challenges of the annotation process

Category	Sub-category 2	Sub-category 1	Meaning unit
Challenges of the annotation process	Confusing discourse relations	Not obvious implicit connectives	*It was hard to mark the annotation when the discourse connectives were implicit and they were not visibly obvious. (S18)*
		Confusing AltLex feature	*The most difficult to learn was the AltLex feature of corpus annotation and hypophora. (S15)*
		Confusion between implicit and entity relations	*The difficult thing to me was to distinguish between implicit relations and entity relations. (S3)*
		Difficulty choosing discourse relations	*It was difficult because sometimes it was not clear which relation to decide on. (S3)*
		Difficulty understanding discourse relation differences	*I did not understand the differences between the types of relations and what they look like in a sentence. (S21)*

	Necessity of experience with the system of discourse relations	Lack of experience	*I have never done anything like this before and therefore it was challenging. (S18)*
		Difficulty memorizing the system of discourse relations	*All the relations were difficult to memorize and while annotating I felt like sometimes I did not know what fits better for a certain sentence or a part of a sentence. (S19)*
		Need for additional knowledge of syntax	*It is a tedious task, which requires a lot of additional syntax knowledge about sentences and their relations and without this knowledge it can be confusing and quite difficult. (S10)*
	Time-consuming	Slow process	*I was going sentence by sentence and even though it was going very slow I think I did a good job since it is the first time I encountered a program like an annotator. (S1)*
		Demanding attention	*You have to be very patient and like to do monotonous work; also you have to be careful and attentive. (S19)*
		Lengthy texts	*This project took a long time, because the text was so long. (S21)*
		Demanding reflection	*Annotation implies much reflection and is time-consuming. (T1)*

The sub-category of confusing discourse relations reveals student research participants' reflections that some discourse relations are not easy to decipher and identify. Such difficulties include not obvious implicit discourse connectives, which imply certain difficulties in annotating discourse relations: "*It was hard to mark the annotation when the discourse connectives were implicit and they were not visibly obvious*" *(S18)*. Other confusing features mentioned by the research participants include AltLex (alternative lexicalization) and hypophora (representing rhetorical questions and answers): "*The most difficult to learn was the AltLex feature of corpus annotation and hypophora*" *(S15)*. In addition, the research participants mentioned confusion between implicit and entity relations as sometimes it is difficult to decide for sure: "*The difficult thing to me was to distinguish between implicit relations and entity relations*" *(S3)*. Some research participants also mentioned that it is difficult for them to choose discourse relations during annotation as they were not able to understand the differences between discourse relations: "*It was difficult because sometimes it was not clear which relation to decide on*" *(S3)*; "*I did not understand the differences between the types of relations and what they look like in a sentence*" *(S21)*.

The sub-category of the necessity of experience with the system of discourse relations includes such sub-categories as difficulty memorizing the system of discourse relations, need for additional knowledge of syntax and lack of experience. The research participants feel that their annotation would improve if they had background syntax knowledge and experience in annotation of discourse relations. Their observations are valid as the background knowledge and experience allow a better understanding and more effective annotation process: "*It is a tedious task, which requires a lot of additional syntax knowledge about sentences and their relations and without this knowledge it can be confusing and quite difficult*" *(S10)*; "*I have never done anything like this before and therefore it was challenging*" *(S18)*.

Another difficulty mentioned by the research participants is the problem of memorizing the system of discourse relations as it represents quite a complicated construct: "*All the relations were difficult to memorize and while annotating I felt like sometimes I did not know what fits better for a certain sentence or a part of a sentence*" *(S19)*.

The research participants also observed that the annotation process is time-consuming, and the sub-categories reveal the reasons why the annotation process takes so much time. The research participants feel that it is a slow process as it is done sentence by sentence and involves a lot of reflective efforts: "*I was going sentence by sentence and even though it was going very slow I think I did a good job since it is the first time I encountered a program like an annotator*" *(S1)*. Teacher research participants also admitted that annotation demands a lot of reflection: "*Annotation implies much reflection and is time-consuming*" *(T1)*. Also, student research participants mentioned lengthy texts and stated that it would be easier to annotate shorter texts: "*This project took a long time, because the text was so long*" *(S21)*. This allows us to reflect on how to make annotation not so time-consuming, and one of the straightforward solutions could be offering shorter texts for annotation.

3.4. Discussion

The core category of the organizational-level factors reveals two key important elements while applying corpus annotation while teaching a foreign language at more advanced levels. One element is the support needed and provided by the teacher, and the other element is the need for initial knowledge. The research participants acknowledge the benefits of using examples, teacher explanation, learning material, group work and learning by doing, which in a way have certain connotations to Hattie's (2012) ranking. Hattie (2012) ranked the system of indicators and found that anything with a score above 0.40 had a direct correlation to student achievement. He ranked direct instruction 0.60 and cooperative learning vs. individualistic learning 0.55. As can be seen in the research findings, the sub-categories of explanations and learning material are related to direct instruction, and cooperative vs. individualistic learning is related to group work. Students also stressed their need for preparation before applying corpus annotation and for introduction of meta-linguistic knowledge and explanation.

All the mentioned categories are related to the other important element—initial knowledge. The research participants even reflected on the year of introducing annotation and corpora application and express their suggestions for later introduction of corpus annotation, either in the second or the third year of studies, which is related to the students' concern about acquiring the prior knowledge. So it could be observed that, merging the initial knowledge and learning by doing, both elements could be related to Hattie's (2012) one indicator among student learning strategies called "strategy to integrate with prior knowledge" with the ranking of 0.93, revealing the importance of prior knowledge and its integration. Actually, the organizational factors emphasize the staff's and the institution's responsibility to cater to students' needs and plan the teaching-learning process having in mind the integration of the prior knowledge. Also, especially concerning the issues of prior knowledge the organizational-level category also intertwines with the personal-level category, revealing that initial linguistic knowledge is essential while teaching/learning a foreign language at more advanced levels while applying corpus annotation and comparing and contrasting coherence relations. The research participants observed that prior knowledge of syntax and textual coherence ensures an easier learning process; it also helps the annotation process and the process of comparing and contrasting discourse relations. What is more, the lack of prior knowledge is admitted to lead to certain difficulties in the learning process; for example, the novelty of the information causes difficulty in understanding and leads to additional learning. The observations of the research participants resonate with the ideas expressed by van Merrienboer and Kirschner (2013) that new information is more easily processed by an experienced person with the prior knowledge than by a person without experience. In addition, the research participants identified certain needs for enhancing the teaching/learning process. They identified their need for preparation, need for introduction of meta-linguistic knowledge and need for explanation; also, the research participants expressed their preference for simpler tasks as they believe it could make their learning more efficient. The needs identified by the research participants resonate with the observations by the authors van Merrienboer, Kirschner and Kester (2003) on the necessity of supportive information, procedural information and part-task practice.

The core category of technology-level factors reveals two elements in the attitudes of the research participants to the software used for corpus annotation. On one hand, the research participants perceived usefulness and ease of using the software, which is expressed by the categories such as usefulness of the program, easiness of use, clarity, interest and novelty. The results of the current study relate to the research on a student group by Smarkola (2007), who, by using the Technology Acceptance Model (TAM), confirmed that perceived usefulness and perceived ease of use are the important factors predicting the user acceptance of computer technology.

On the other hand, the research participants observed certain barriers to using the software for corpus annotation which are expressed by the categories of need for knowledge of the annotation software, confusing software features, problematic managing of the annotation software, initial difficulties, problematic switching between languages, need to memorize software abbreviations, low-quality public computers and need for technical support. In fact, the mentioned categories represent three main elements analyzed in scientific literature: lack of knowledge and skills, lack of access or shortage of technical supply, and lack of technical support. Recent research by Siddiquah and Salim (2017) shows that students are experts at simple skills like MS Word and MS Power, but they lack skills in using more sophisticated software, which could be one of the barriers to ICT use. The authors also reveal that students mention such factors as slow speed of computers and poor working condition of computers as main impediments to ICT use. It could be seen that the current research relates to the research carried out by Siddiquah and Salim (2017) as the research participants mentioned that their knowledge of the annotation software and their studies are affected by low-quality public computers. Actually, shortage of technological supply might be seen as one of the impeding barriers to using innovative software. Having the Internet and computer access is now becoming viewed as one of the fundamental essentials like shelter and water (Cisco 2011). Cisco's (2011) research reveals that one in three college students consider the Internet to be just as fundamental as air, water, shelter and food are. It seems that an Internet connection is becoming perceived as vital by the majority. Lack of technical support also has been researched as one of the barriers hindering the use and integration of ICT in teaching and learning (Pelgrum 2001).

The core category of personal-level factors includes motivation, teaching and learning outcomes, initial knowledge, teaching and learning challenges and benefits, and annotation procedures. According to Hattie (2012), motivation ranks 0.42, which means it has a direct correlation to student achievement. In addition, Ambrose et al. (2010) state that student motivation activates, directs and sustains student learning. In the current research, student research participants mentioned success as one of the motivating factors and fear of failure (such as mistakes and negative assessment) as one of the factors hindering motivation. Ambrose et al. observe that "when students successfully achieve a goal and attribute their success to internal causes (for example, their own talents or abilities) or to controllable causes (for example their own efforts or persistence), they are more likely to expect future success" (Ambrose et al. 2010, 77). But, according to the authors, when a student fails to achieve their learning goals and attributes the failure to their lack of ability, the student motivation is likely to be low. What is more, other sub-categories of the current research such as interest and relating theory to practice also correlate with the factors mentioned by Ambrose et al. (2010), including student interest and real-world tasks relevant to students' future professional lives.

In the category of teaching and learning outcomes, the current research reveals such major categories as acquired professional knowledge and skills. Acquired professional knowledge is closely related to the translation competence model introduced and researched by the PACTE group—the so-called PACTE model (PACTE 2003). The sub-categories, including experience of the process of translation, understanding comparative-contrastive use of languages, accuracy, editing awareness, using dictionaries and cultural awareness, have close connotations to the four criteria for evaluating translation competence introduced by Biggs and Tang (2007) based on the PACTE model. It could be seen that the criteria of translating texts by ensuring the function and informative content, editing and revising, applying specific methods effectively and understanding specialized concepts are related to the above-mentioned sub-categories. The research participants acknowledged that they need to know the process of translation and be able to understand and apply the comparative-contrastive use of both the source and target languages; the research participants admitted that they need to improve their editing skills and also use dictionaries and demonstrate a certain cultural awareness of the source and target languages.

In addition, the category of linguistic awareness, which involves the sub-categories of understanding semantics and syntax, linguistic knowledge, deeper linguistic textual understanding, mastery of discourse cohesion and analytical abilities, has definite connotations to the communicative language competence defined by the *Common European Framework of Reference for Languages: learning, teaching, assessment* (Council of Europe 2001) as comprising the linguistic, sociolinguistic and pragmatic components. Here, the research participants mentioned all three components by admitting that they need understanding of semantics and syntax, mastery of discourse cohesion, and analytical abilities concerning languages and their understanding of them.

The categories of acquired skills and necessary skills reveal that the research participants identified that they both acquire and need professional skills related to the study program as well as transversal skills, such as analytical research skills, communication, ICT skills, self-discipline and ability to learn independently, which are certainly related to the ERI-Net (2013) conceptualized framework of transversal skills. This framework includes five domains encompassing critical and innovative thinking, interpersonal skills, intrapersonal skills, global citizenship, and media and information literacy.

It could be seen that the research participants observed that both their professional translation competence and their linguistic competence are developed together with transversal skills essential in their later professional and personal lives.

The importance of initial knowledge at the personal level was discussed by the research participants. They admitted the benefits of prior knowledge such as an easier learning process, upgrading their prior knowledge, help while annotating the discourse relations, and help while comparing and contrasting discourse relations in different texts. The observations expressed by the research participants resonate with the ideas discussed by van Merrienboer and Kirschner (2013) that new information is more easily processed by an experienced person with the prior knowledge than by a person without experience. In addition, the category of difficulties without the prior knowledge also supports the idea expressed above by van Merrienboer and Kirschner (2013).

The category of teaching and learning challenges and benefits provides insights on what the research participants identified as challenging in teaching and learning through corpus annotation and what they find beneficial. These insights are especially useful for formulating the recommendations.

The challenges of teaching and learning through annotation are identified as, first, that it might be viewed as a time-consuming and stressful process; it also requires mastering new software, putting in extra efforts, and dealing with a large amount of new information, linguistic differences and complexity. It could be observed that teaching and learning with new technology poses certain challenges related to processing new information discussed by van Merrienboer and Kirschner (2013).

However, the benefits are related to professional growth as the research participants identified such benefits as teaching and learning linguistic knowledge, raising awareness of linguistic differences, raising translation awareness, developing analytical skills and usefulness for research.

The annotation process itself features its own challenges and benefits perceived by the research participants. The annotation software and its use was perceived by the research participants in a twofold way. On one hand, the research participants named such benefits as easily distinguishable discourse relation and value of practice. On the other hand, they named such challenges as confusing discourse relations, the necessity of experience with the system of discourse relations and the annotation process being time-consuming. It could be seen that some discourse relations were clearly represented and were easily grasped by the research participants, but there were some discourse relations which were not so easily distinguishable, and thus the research participants saw the annotation process as time-consuming and observed that they needed some experience with the system of discourse relations. Van Merrienboer, Kirschner and Kester (2003) suggest the necessity of supportive information, procedural information and part-task practice as a solution while dealing with the challenges posed while dealing with the new information in the teaching and learning process.

3.5. Recommendations

Following the research participant observations on teaching and learning corpus annotation for raising awareness in translation, certain recommendations could be drawn.

The first recommendation, based on all the present study, is that corpora, corpus building and annotation tools could be used in teaching a foreign language at more advanced levels. All the positive insights expressed by the research participants leads to the identification that the application of corpus annotation while teaching a foreign language at more advanced levels provides enhancement of the teaching and learning process and is beneficial for the learners. As the study program of the research participants is translation and editing, they stressed the usefulness of the application of annotation, which adds to raising translation awareness, acquiring linguistic knowledge and raising awareness of linguistic differences. The research participants also acknowledged the usefulness of acquiring certain transversal skills such as analytical skills for research and learning.

The second recommendation, based on student research participants' reflections on encouraging and discouraging motivation, and their needs in the teaching and learning process, is that it is imperative that the teaching and learning process should be carefully planned and it should embrace well-prepared examples, teacher explanations, task gradation and continuous support, which should ensure students experience success in their learning process. It should be noted that the research participants identified their need for preparation which would make their learning easier. They also emphasized their need for introduction of meta-linguistic knowledge and need for explanation which could enhance their learning. The research participants expressed their preference for simpler tasks as they believe it could make their learning more efficient moving from simpler tasks to more complex ones.

The third recommendation, based on the research participants' insights on the importance of prior knowledge and their suggestions, envisions introducing annotation and corpora application later in the study process, possibly either in the second or the third year of studies or even to master's students. The student research participant reflections about second or third study years are followed by the suggestion by their teachers of the possibility of introducing annotation and corpora application even to master's students. The later introduction of corpus annotation for teaching and learning a foreign language at a more advanced level could be also supported by the theoretical idea that human expertise relies on the availability of rich and automated knowledge schemas, which help an experienced person with the prior knowledge to process new information.

The fourth recommendation, based on the research participant insights on using technology (i.e., the annotation software), encourages the universities to constantly renew their public computer lot and provide technical and methodological support for the teaching staff concerning ICT use in the teaching and learning process. The student research participants observed that the low-quality public computers slow down the annotation process, which could be considered as an impediment in their learning process. The teacher research participants expressed their wish for technical and methodological support as sometimes they feel left on their own to grapple with new technologies and learn at their own expenses of personal time.

3.6. Conclusions

First, the category of usefulness of teaching and learning corpus annotation leads to reflecting on the benefits provided by the research participants and their insights into why teaching and learning corpus annotation presents some value for the research participants. The category contains the sub-categories of teaching and learning linguistic knowledge, raising awareness of linguistic differences, raising translation awareness, developing analytical skills and usefulness for research. The research participants identified that teaching and learning corpus annotation enables their professional growth as it provides them with the linguistic knowledge of syntax, semantics and discourse structure. They also stressed the usefulness of the process, which allows them to get a more rounded understanding of the differences between languages.

Next, speaking about the provided support in the teaching and learning process, it appears that examples are really helpful while analyzing the annotation scheme and trying to apply it in one's own language. Also, the research participants discussed the importance of teacher explanation and the learning material prepared by the teacher. Another important factor is learning by doing, as the research participants observed. Student research participants identified that practice makes it easier to grasp the peculiarities of the process of annotation. Finally, group work was also appreciated by the research participants. In addition, the research participants identified certain needs for enhancing the teaching/learning process. They identified their need for preparation, need for introduction of meta-linguistic knowledge and need for explanation; also, the research participants expressed their preference for simpler tasks at the beginning of the teaching and learning process as they consider it could make their learning more efficient. The needs identified by the research participants reveal the necessity of supportive information, procedural information and part-task practice. In fact, the research participants identified the importance of a gradual teaching and learning process based on teacher support and practice.

Further, student research participants revealed that motivation is an important factor in the teaching and learning process, and the sub-categories of the main categories of encouraging and discouraging motivation disclose the influential elements shaping student motivation. The current research reveals that the encouraging motivation includes success, the possibility of relating theory with practice, familiarity with the topic, interest and novelty. The research participants identified success as an important driving force of motivation. Another motivating element was relating theory to practice, which gives a rewarding feeling of being content with the results of learning. Interest was also mentioned as a motivating element for learning by the research participants since the new acquired knowledge maintains students' interest. What is more, novelty is related to inducing interest and motivation to learn. The research participants disclosed that they were interested in working with new software and learning new knowledge. Familiarity with the topic may also increase motivation by bringing enjoyment in studying the topic more. The research participants felt happy to be able to work on a familiar topic again.

The category of discouraging motivation includes the sub-categories of lack of examples, mistakes, fear of negative assessment and difficulty in understanding. The sub-category of lack of examples reveals an interesting observation that sometimes providing a few examples is not enough and some research participants felt that they needed examples which could enhance their understanding. It seems that it could be a good idea provide more examples while introducing the annotation, taking into account that some research participants identified their need for more examples. Another discouraging element is mistakes in the text which was being annotated. Some mistakes may even lead to the misunderstanding of the text, as one of the research participants observes. Naturally, the mistakes appeared due to the fact that the texts were translated by the peers learning the annotation. However, it is worthwhile considering choosing some original texts for teaching and learning annotation and thereby to avoid the issue of mistakes. The sub-category of fear of negative assessment raises an ongoing question of motivating students through punishment or reward. The case reflected by one research participant can hardly be considered positive as the main motivation to continue the annotation process was reluctance to get a negative grade. Finally, speaking about discouraging motivation we arrive at difficulties in understanding, which—depending on the reasons—serve as a discouraging leverage for continuing the task and achieving the results. As can be seen in the reflections of the research participants, difficulties in understanding cause negative emotions lingering for a long time even after the task is finished. This feeling of inability to understand does not enhance the quality of teaching and learning. The research participants recognized that difficulties in understanding leave them with the feelings of negative experience and feeling lost not knowing what they were doing. Overall, reflection on encouraging and discouraging motivation leads to the conclusion that providing enough examples and making sure that students experience success in grasping the annotation task and performing it appropriately should be kept in mind while teaching and learning annotation.

The current research reveals that initial linguistic knowledge is essential while teaching/learning a foreign language at more advanced levels while applying corpus annotation and comparing and contrasting coherence relations. The majority of the students with the prior knowledge stressed the advantages of having and applying their prior knowledge. They identified that it makes their learning process easier, upgrades their prior knowledge, helps them to annotate the discourse relations, and assists them while comparing and contrasting discourse relations in different texts. The research participants observed that having prior knowledge of syntax and textual coherence ensures an easier learning process. Another point observed by the research participants proves that prior knowledge helps the annotation process. The research participants pointed out that the syntax knowledge was of great help while working on the annotation project since they knew where to divide sentences and how the sentences are connected. It also was revealed that prior knowledge helped during the process of comparing and contrasting coherence relations. The research participants observed that prior knowledge of syntax and textual coherence ensures an easier learning process; it also helps the annotation process and the process of comparing and contrasting discourse relations. The observations expressed by the research participants support the ideas discussed that new information is more easily processed by an experienced person with the prior knowledge than by a person without experience.

On the other hand, the category of difficulties without prior knowledge reveals that the absence of prior knowledge, on the contrary, makes the learning process more difficult. The category reveals that lack of prior knowledge may lead to difficulty in understanding, more learning and grappling with the novelty of the information. All in all, the lack of prior knowledge is admitted to lead to certain difficulties in the learning process; for example, the novelty of the information causes difficulty in understanding and leads to additional learning. It could be concluded that the research participants support the idea that new information is more easily processed by an experienced person with the prior knowledge than by a person without experience. Finally, the research participants reflect on the year of introducing annotation and corpora application and express their suggestions for later introduction of corpus annotation, either in the second or the third year of studies, which is also related to the students' concern about acquiring the prior knowledge.

Considering the use of technology (ICT), the category of advantages of using technology contains the sub-categories of clarity, novelty, interest, easiness of use and usefulness of the software. The research participants identified traditional advantages provided by technologies and found using technologies really useful. However, the research participants pointed out certain challenges of using the software for annotation. Some research participants still found some software features confusing, and they found it problematic to manage the software appropriately. Other research participants stressed the clarity, though, leading to the observation that different people have different software managing skills and some need more practice with new software, which is revealed by the sub-category of initial difficulties. In addition, the software features and the annotation process itself pose such challenges as difficulties switching between languages and the sentences of the text moving to different positions in the tool, and the need to memorize software abbreviations and know the meanings of the acronyms used in the tool. Student research participants also mentioned the low quality of the public computers which makes the students spend extra time on annotation as the computers are slow or do not work properly. This sub-category reveals that universities should always consider a constant renewal of the technical resources. The need for this renewal is caused by the constant development of technology and the surplus of new knowledge which requires modern technological solutions. Teacher research participants mentioned the need for technical support as they need help while using the newest technologies for teaching and learning and implementing the newest technological solutions. They also felt that they need knowledge of annotation software because if teachers want to be able to use the newest technologies for teaching and learning they also need certain training and knowledge, but usually they are left to learn everything on their own. Thus, even though use of new computer software is perceived as positive, there are still certain issues to be solved such as the constant renewal of public computers at universities and technical and methodological support for teachers while using new software in teaching and learning.

References

Adab, B. 2000. "Evaluating translation competence." *Benjamins Translation Library*, 38, 215–28.

Albir, A. H., ed. 2017. *Researching translation competence by PACTE group*. Vol. 127. Amsterdam: John Benjamins.

Al-Saif, A. and K. Markert. 2010. "The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic." *LREC*, 2046-2053.

Ambrose, S. A., M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman. 2010. *How learning works: Seven research-based principles for smart teaching*. San Francisco, CA: John Wiley & Sons.

Asher, N. 1993. *Reference to abstract objects in discourse*. Boston: Kluwer Academic Publishers.

Aston, G. 2001. "Learning with corpora: An overview." In *Learning with corpora*, edited by G. Aston, 7–45. Houston: Athelstan.

Au, K. K. L. 1999. "Cultural transfer in advertisement translation." *Babel* 45 (2): 97–106.

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, M. 2018. *In other words: A coursebook on translation*. London: Routledge.

Bertsch, S., B. J. Pesta, R. Wiscott, and M. A. McDaniel. 2007. "The generation effect: A meta-analytic review." *Memory & cognition* 35 (2): 201–10.

Biggs, J., and C. Tang. 2007. *Teaching for quality learning at university*. 3rd ed. Maidenhead, England: McGraw-Hill.

Boulton, A., and H. Tyne. 2014. *Corpus-based study of language and teacher education*. New York: Routledge.

Bowker, L., and J. Pearson. 2002. *Working with specialized language: a practical guide to using corpora*. London: Routledge.

Cettolo, M., Girardi, C., and Federico, M. 2012. "Wit3: Web inventory of transcribed and translated talks." In *Conference of European Association for Machine Translation*, 261–68.

Charolles, M. 1983. "Coherence as a principle in the interpretation of discourse." *Text—Interdisciplinary Journal for the Study of Discourse* 3 (1): 71–98.

Cisco. 2011. Connected World Technology Report. San Jose, California. <http://newsroom.cisco.com/press-release-content?articleId=474852.> Accessed August 14, 2020.

Cobb, Th., and A. Boulton. 2015. Classroom Applications of Corpus Analysis." In *The Cambridge Handbook of English Corpus Linguistics*, edited by D. Biber-Reppen, 478–97. Cambridge: Cambridge University Press.

Council of Europe (Council of Europe Council for Cultural Co-operation, Education Committee, Modern Languages Division. 2011. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Strasbourg: Cambridge University Press.

Creswell, J. W. 2007. *Qualitative inquiry and research method: Choosing among five approaches*. 2nd ed. Thousand Oaks, CA: SAGE.

De Grave, W. S., H. G. Schmidt, and H. P. A. Boshuizen. 2001. "Effects of problem-based discussion on studying a subsequent text: A randomized trial among first year medical students." *Instructional Science*, 29, 33–44.

De Jong, T. 2006. "Technological advances in inquiry learning." *Science* 312 (5773): 532–33.

De Jong, T., and A. W. Lazonder. 2014. "The guided discovery learning principle in multimedia learning." In: *The Cambridge handbook of multimedia learning*, 2nd ed., edited by R. E. Mayer, 371–90. New York: Cambridge University Press.

De Jong, T., and M. Njoo. 1992. "Learning and instruction with computer simulations: Learning processes involved." In *Computer-based learning environments and problem solving*, 411–27. Berlin, and Heidelberg: Springer.

Dupont, M., and S. Zufferey. 2017. "Methodological issues in the use of directional parallel corpora." *International journal of corpus linguistics* 22 (2): 270–97.

Elo, S., and H. Kyngas. 2007. "The qualitative content analysis process." *Journal of Advanced Nursing* 62 (1): 107 15.

ERI-Net Research Programme. 2013. 2013 Asia-Pacific Education Research Institutes Network (ERI-Net) regional study on: transversal competencies in education policy and practice (Phase I): regional synthesis report. <http://unesdoc.unesco.org/images/0023/002319/231907E.pdf>. Accessed August 20, 2020.

Fawcett, P. 1987. "Putting translation theory to good use." In *Translation into Modern Languages Degree*, edited by Hugh Keith and Ian Mason, 31–38. London: CILT.

Gaskell, D., and T. Cobb. 2004. "Can learners use concordance feedback for writing errors?" *System* 32 (3): 301–19.

Ghiglione, R., and B. Matalon. 2001. *O Inquérito: Teoria e Prịtica*. Oeiras: Celta Editora.

Granger, S. 2015. "Contrastive Interlanguage Analysis: A Reappraisal." *International Journal of Learner Corpus Research* 1: 7–24.

Gurlitt, J., A. Renkl, M. Motes, and S. Hauser. 2006. "How can we use concept maps for prior knowledge activation—different mapping-tasks lead to different cognitive processes." In *Proceedings of the 7th International Conference of the Learning Sciences*, edited by S. A. Barab, K. E. Hay and D. T. Hickey, 217–21. Mahawah, NJ: Lawrence Erlbaum.

Halliday, M. A. K. 1994. *An introduction to functional grammar*. London: Edward Arnold.

Halliday, M. A. K., and R. Hasan. 2014. *Cohesion in english*. London: Routledge.

Hanks, P. 2013. *Lexical analysis: Norms and exploitations*. London: Mit Press.

Hatim, Basil, and Ian Mason. 1997. *The Translator as Communicator*. London and New York: Routledge.

Hattie, J. 2012. *Visible learning for teachers: Maximizing impact on learning*. London: Routledge.

Hirschberg, J., and D. Litman. 1987. "Now let's talk about now: Identifying cue phrases intonationally." In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, 163–71. Association for Computational Linguistics.

Hmelo-Silver, C. E., R. G. Duncan, and C. A. Chinn. 2007. "Scaffolding and achievement in problem-based and inquiry learning: a response to Kirschner, Sweller, and Clark." *Educational psychologist* 42 (2): 99–107.

Hoek, J., S. Zufferey, J. Evers-Vermeul, and T. J. Sanders. 2017. “Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study.” *Journal of pragmatics* 121, 113–31.

Hoey, M. 1991. *Patterns of lexis in text*. Oxford: Oxford University Press.

Holes, C. 1984. *Colloquial Arabic of the Gulf and Saudi Arabia*. London: Routledge & Kegan Paul Books.

Johns, T. 1990. “From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning”. *CALL Austria* 10, 14–34.

Kirschner, P. A., J. Sweller, and R. E. Clark. 2006. “Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching.” *Educational psychologist* 41 (2): 75–86.

Klahr, D., and M. Nigam. 2004. “The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning.” *Psychological science* 15 (10): 661–67.

Kubler, N., and P. Y. Foucou. 2003. “Teaching English verbs with bilingual corpora: Examples in the computer science area.” *Contrastive Linguistics and Translation Studies*. 1-15. Amsterdam: Rodopi.

Kvale, S. 1996. *InterViews: An Introduction to Qualitative Research Interviewing*. Thousand Oaks, CA: SAGE.

Larsen-Freeman, D., and L. Cameron. 2008. “Research methodology on language development from a complex systems perspective.” *The Modern Language Journal* 92 (2): 200–213.

Lazonder, A. W., M. G. Hagemans, and T. De Jong. 2010. “Offering and discovering domain information in simulation-based inquiry learning.” *Learning and Instruction* 20 (6): 511–20.

Lefer, M. A., and N. Grabar. 2015. “Super-creative and over-bureaucratic: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates.” *Across Languages and Cultures* 16 (2): 187–208.

Machiel-Bongaerts, M., H. G. Schmidt, and H. P. A. Boshuizen. 1995. “The effect of prior knowledge activation on text recall: An investigation of two conflicting hypotheses.” *British Journal of Educational Psychology*, 65, 409–23.

Martin, J., and L. Hewson. 1991. *Redefining Translation: The Variational Approach*. London: Routledge.

Mayer, R. E. 2004. “Should there be a three-strikes rule against pure discovery learning?” *American psychologist* 59 (1): 14.

McEnery, A.M. 2003. Corpus Linguistics. In: Mitkov, R. (ed.) The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, pp. 448-463.

Newmark, P. 1987. *The use of systemic linguistics in translation analysis and criticism*. Amsterdam/Philadelphia: John Benjamins.

Newmark, P. 1988. *A textbook of translation*. Vol. 66. New York: Prentice Hall.

Nord, C. 1997. "Functional translation units." *AFinLAn vuosikirja*. 41-49.

Oleskeviciene, G. V. Zeyrek, D. Mazeikiene, V. & Kurfalı, M. 2018. "Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian." In *Proceedings of the Workshop on Annotation in Digital Humanities Co-Located with ESSLLI*, *2155*, 53–58.

Oza, U., R. Prasad, S. Kolachina, D. M. Sharma, and A. Joshi. 2009. "The Hindi Discourse Relation Bank." In Proc. of the 3rd Linguistic Annotation Workshop, pages 158–61. Association for Computational Linguistics.

PACTE (Action Plan for Business Growth and Transformation). 2003. "Building a translation competence model". In *Triangulating Translation: Perspectives in process oriented research*, edited by Fabio Alves. 43-68. Amsterdam: John Benjamins.

PACTE (Action Plan for Business Growth and Transformation). 2011. "Results of the validation of the PACTE translation competence model: Translation problems and translation competence." In *Methods and Strategies of Process Research: Integrative Approaches in Translation Studies*, edited by Alvstad, C., Hild, A., & Tiselius, E., 4–5. Amsterdam: John Benjamins.

Pelgrum, W. J. 2001. "Obstacles to the integration of ICT in education: results from a worldwide educational assessment." *Computers & education* 37 (2): 163–78.

Prasad, R., B. Webber, and A. Joshi. 2014. "Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation." *Computational Linguistics*. *40*(4), 921-950.

Pym, A. 2003. "Redefining translation competence in an electronic age. In defence of a minimalist approach." *Meta: journal des traducteurs/Meta: Translators' Journal* 48 (4): 481–97.

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Wakberg-Henriksson, H., & Hemmo, V. 2007. *Science education now: A renewed pedagogy for the future of Europe*. Luxemburg: Office for Official Publications of the European Communities.

Sager, J. C. 1983. "Quality and Standards: the Evaluation of Translation." In *The Translator's Handbook*, edited by C. Picken, 121–28. London: Aslib.

Scott, M., and C. Tribble. 2006. *Textual patterns: Key words and corpus analysis in language teaching*. Amsterdam: John Benjamins.

Siddiquah, A., and Z. Salim. 2017. "The ICT facilities, skills, usage, and the problems faced by the students of higher education." *EURASIA Journal of Mathematics, Science and Technology Education* 13(8): 4987–94.

Silverman, D. 2005. *Doing qualitative research: A practical handbook*. London: SAGE.

Sinclair, J. M. 1991. *Corpus, concordance collocation*. Oxford: Oxford University Press.

Smarkola, C. 2007. "Technology acceptance predictors among student teachers and experienced classroom teachers." *Journal of educational computing research* 37(1): 65–82.

Smith, K. 2002. "Translation as secondary communication. The relevance theory perspective of Ernst-August Gutt." *Acta Theologica* 22(1): 107–17.

Smith, R. N., and W. J. Frawley. 1983. "Conjunctive cohesion in four English genres." *Text-Interdisciplinary Journal for the Study of Discourse* 3(4): 347–74.

Snell-Hornby, M. 1988. *Translation studies: An integrated approach*. Amsterdam: John Benjamins.

Strauss, A. L., and J. Corbin. (1990) 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. London: SAGE.

Stubbs, M. 2004. "Language corpora." In *The handbook of applied Linguistics*, edited by A. Davies and C. Elder, 106-113. London: Blackwell Publishing.

Sweller, J., P. Ayres, and S. Kalyuga. 2011. *Cognitive load theory*. New York: Springer.

Tomasello, M. 2005. "Beyond formalities: The case of language acquisition." *The Linguistic Review* 22(2–4): 183–97.

Van Merrienboer, J. J. G., and P. A. Kirschner. 2013. *Ten steps to complex learning*. 2nd ed. New York: Routledge.

Van Merrienboer, J. J., and P. A. Kirschner. 2017. *Ten steps to complex learning: A systematic approach to four-component instructional design*. New York: Routledge.

Van Merrienboer, J. J. G., P. A. Kirschner, and L. Kester. 2003. "Taking the load off a learner's mind: Instructional design for complex learning." *Educational Psychologist* 38, 5–13.

Vermeer, H. J. 1994. “Translation today: Old and new problems.” *Translation studies: An interdiscipline*, 3–16.

Williams, M. 1989. “The assessment of professional translation quality: Creating credibility out of chaos.” *TTR: traduction, terminologie, rédaction* 2(2): 13–33.

Zeyrek, D., I. Demirsahin, A. Sevdik-Çallı, and R. Çakıcı. 2013. “Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language.” *Dialogue and Discourse* 4(2): 174–84.

Zeyrek, D., A. Mendes, and M. Kurfalı. 2018. “Multilingual extension of PDTB-style annotation: The case of TED Multilingual Discourse Bank.”, *Proceedings of the 11th Language Resources and Evaluation Conference-LREC’2018*, 1913-1919. European Language Resources Association.

Zhou, Y., and N. Xue. 2012. “PDTB-style discourse annotation of Chinese text.” In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers—Volume 1, pages 69–77. Association for Computational Linguistics.