University of Zadar
Universitas Studiorum
Jadertina | 1396 | 2002 |

CORPORA IN LANGUAGE LEARNING, TRANSLATION AND RESEARCH

Proceedings of the International Conference Corpora in Language Learning, Translation and Research 2023

Edited by Larisa Grčić and Marija Brkić Bakarić

Corpora in Language Learning, Translation and Research

Proceedings of the International Conference
Corpora in Language Learning, Translation and Research
held at the University of Zadar (August 23–24, 2023)

Edited by

Larisa Grčić and Marija Brkić Bakarić

Proofreader: Leonarda Lovrović

Publisher

University of Zadar

For the publisher
Josip Faričić, rector

ISBN: 978-953-331-535-5

Reviewers

Benedikt Perak
Špela Vintar

Paper reviewers

Marco Angster (University of Zadar, Croatia), Paula de Santiago (University of Valladolid, Spain), Katja Dobrić Basaneže (University of Pula, Croatia), Ivana Lalli Paćelat (University of Pula, Croatia), Benedikt Perak (University of Rijeka, Croatia), Vanda Mikšić (University of Zadar, Croatia), Jelena Parizoska (University of Zagreb, Croatia), Ivanka Rajh (University of Zagreb, Croatia), Lucija Šimičić (University of Zadar, Croatia), Špela Vintar (University of Ljubljana, Slovenia)

Organizing Committee

Larisa Grčić (President), Vanda Mikšić,
Maja Bahnik (University of Zadar, Croatia)

Programme Committee

Maja Bratanić (Croatia), Paula de Santiago (Spain), Adriana Mezeg (Slovenia), Maja Lončar (Croatia), Ana Ostroški Anić (Croatia), Mojca Pecman (France), Ivanka Rajh (Croatia), Špela Vintar (Slovenia)

Design and layout: KaramanDesign

CORPORA IN LANGUAGE LEARNING, TRANSLATION AND RESEARCH

Edited by
Larisa Grčić and Marija Brkić Bakarić

University of Zadar
Universitas Studiorum
Jadertina | 1396 | 2002 |

Zadar 2024

Table of contents

Preface

In the realm of modern linguistic studies, corpora play a foundational role that extends across diverse disciplines and applications. These extensive language resources are not merely repositories of linguistic data but serve as indispensable tools for language learners, translators, and researchers alike. The articles compiled in this conference proceedings delve into interconnected themes such as corpora, AI technologies, language education, and cultural translation. Together, they celebrate the dynamic synergy between language, technology, and culture in modern linguistic research. Corpus-driven studies shed light on language use and cultural expressions, revealing how languages adapt within diverse cultural contexts. The task of cultural translation highlights the complexities of conveying cultural nuances across languages and cultures. By exploring language corpora, educators can leverage innovative strategies to enhance language proficiency among learners, thus influencing language pedagogy and curriculum design, and individuals can gain deeper insights into language structure, usage variations, and cultural expressions. Through an exploration of the development and utilization of both general and specialized corpora, the selected collection of articles seeks to enrich our understanding of language and its multifaceted dimensions in today's rapidly evolving global context.

Each of the nine articles applies a methodology appropriate to its specific research objectives, both for analysing and understanding linguistic and cultural phenomena.

In "Learning about Corpora, Learning with Corpora", **Larisa Grčić** opens the question of using corpora in learning environment and explains their potential for the improvement of students' language competencies, the development of their analytical skills, and a deeper understanding of linguistic structures. **Ivana Havelka**'s article "ChatGPT & Co: AI Transforming Terminological Preparation in Interpreting" examines the impact of artificial intelligence on terminology preparation in translation. A case study is used to evaluate selected AI tools and discuss how they can improve translation efficiency. **Jana Kegalj** and **Mirjana Borucinsky**, in their paper "Data-Driven Learning for Writing Skills Development", present the ways that corpora, corpus tools, and corpus methods can be used to develop students' writing skills, and at the same time enable them to improve their digital competencies.

The paper entitled "Adapting the Generic English-Croatian NMT Model to a Religious Domain", co-authored by **Košuta Estera Lerga**, **Lucia Načinović Prskalo**, and **Marija Brkić Bakarić**, examines the adaptation of a generic neural machine translation (NMT) model to a specific domain and illustrates how models can be optimized for specific linguistic and cultural contexts. A combination of traditional corpus linguistic methods and Natural Language Processing methods is illustrated in **Frane Malenica's** paper entitled "Picking Up the Scraps—Creating a

Specialized Corpus Using Web Scraping Tools". In her contribution "Nursing Corpus and the Academic Collocation List", **Kaja Mandić** combines results from the two specialized corpora and the academic collocation list with the aim to generate a field-specific academic vocabulary list as a teaching material. In the article "Definitional Patterns in Specialized Resources for Schoolchildren", **Ana Ostroški Anić** uses a corpus-driven methodology to identify definitions and concept characteristics with the aim to improve the way complex concepts are taught to children. In "Translating Elements of Culture Using the Example of the Series Only Fools and Horses", **Nikolina Palašić** and **Klaudia Križanec** explore strategies for translating cultural elements and understanding the challenges of translating culturally specific content. Finally, **Eriola Qafzezi** devotes her contribution to "Inside Out: A Corpus-Driven Study of Expressions with Parts of the Body in the Albanian Language", as the title reads; with richness of detail, she ponders the potential of corpus use in examining cultural and linguistic patterns within the cross-linguistic perspective.

In preparing this volume we have benefited from the generous help of several people. Apart from the contributors themselves, we wish to thank all the reviewers for their valuable comments and suggestions.

Larisa Grčić and Marija Brkić Bakarić

Larisa Grčić

Department of French and Francophone Studies, University of Zadar
lgrcic@unizd.hr

Learning about Corpora, Learning with Corpora

Abstract

Integrating corpora in university education has not been an easy task and still represents a major challenge. Despite the acknowledged advantages of the use of corpora in the realms of language teaching, translation, and research, the implementation of data-driven techniques and tools at university is very slow and partial. The aim of this paper is to offer an insight into the different ways in which corpora can be exploited in a high education environment, especially for language learning and translation.

Keywords: corpora, language learning, language teaching, translation education

1. Introduction

Corpus linguistics, a field that emerged with the advent of corpora, has revolutionized linguistic research. Corpora has gained popularity in a range of language-related disciplines over the past five decades allowing a shift from an intuitive and introspective armchair approach to empirical descriptive linguistics. As a complementary resource to dictionaries and grammars, corpus data provide large and diverse authentic datasets allowing the identification but also an in-depth analysis of linguistic patterns, language variation, diachronic changes, language evolution. Corpora have become valuable resources for sociolinguistics, comparative linguistics, contrastive analysis, lexicography, language typology, forensic linguistics, media studies, gender studies, and many others. They have had a significant impact on the development of language technologies as they rely on vast amounts of linguistic data to function effectively and provide accurate results. In the age of artificial intelligence and NLP, corpora still play a crucial role in training and testing language models, chatbots, and translation algorithms (Absalom 2021).

In this paper, we attempt to present a review of applied corpus-related research that confirms the innovative use of corpora regarding language learning, language pedagogy, and translator education. Introducing corpora into university education offers an opportunity to apply an empirical method of studying language in use supporting inductive learning. Both teachers and students can benefit immensely from integrating corpora in their investigations dedicated to various aspects of

language, including morphology, syntax, phraseology, semantics, discourse analysis, pragmatics, terminology, and many others. As corpus linguistics is a wide area of research, it is impossible to present the whole range of possible uses. Therefore, our literature review has the purpose to provide basic background information for those who are not yet familiar with corpus methodology and its advantages and are interested in starting to use corpora.

The article begins with the overview of the key groundbreaking moments in the evolution of the corpus approach. Following the introduction, Section 2 is divided into three subsections providing background information about the main corpus features such as compilation criteria, different corpus types, and query techniques. In Section 3, we give a brief overview of the corpus activities aimed at developing comprehension and production skills in language learning. The contribution of corpus-based pedagogy to translator education is described in Section 4. Outlining some of these earlier experiences in corpus pedagogy, we aim to illustrate the potential of corpus use in higher education.

2. Designing and Building Corpora

Before being applied to various spheres of research, first corpora were compiled for lexicographical purposes, like the one used for *Oxford English Dictionary* published in 1928. With the arrival of computer units at universities in the 1960s, computational lexicography started developing. The largest computer corpus of the English language was created in a groundbreaking research project at the University of Birmingham led by J. Sinclair, as a support for creating a corpus-based COBUILD dictionary published in 1987. The corpus-based approach was innovative in considering language use as evidence for pre-existing linguistic theoretical statements.

A more creative vision of language was developed within a corpus-driven perspective which aimed to make theoretical hypotheses based on observation and empirical evidence. This approach was successfully incorporated within the framework of advocacy for the so-called Data-Driven Learning (DDL) introduced by Johns (1990, 1991). The basic postulate of the DDL approach was that students should acquire language knowledge inductively, using authentic data and the independent discovery of rules in them. As the learning-centred approach aligned with constructivist approaches to language acquisition, data-driven methodology was integrated not only in foreign language learning but also in LSP and CLIL classes. The direct use of corpora for teaching grammar and facilitating lexis acquisition or developing academic literacy and linguistic reflection skills was introduced at university level. However, the benefits of applying corpora were also confirmed for lower-level learners (Braun 2007; Ackerley and Coccetta 2007). Indeed, as Tognini-Bonelli pointed out, in the field of language teaching, corpus linguistics has changed "both the object to be taught and the way it is taught" (2001: 23).

Although both corpus-based and corpus-driven research have shown countless

possibilities for exploring the creativity, innovation, and potential of language phenomena, linguists emphasize that each corpus has its limitations, as no corpus can fully represent the entire language use. Some of the main issues related to corpus design are presented in the next chapter.

2.1. Corpus Compilation

While discussing methods in corpus linguistics, most frequent issues concern the corpus content, its reliability, size, and form. As these questions address fundamental concerns regarding corpus compilation, we suggest revisiting the definition of a corpus provided by Sinclair (1991): "a corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." According to the mentioned features, the author assumes three essential principles for corpus compilation. The first refers to the selection of language samples or sampling, the second to the principle of balance, and the third to the principle of representativeness.

In corpus design, sampling implies choosing adequate texts according to the research purpose. Since no sample can be representative of the entire language system, they should be selected according to their communicative function and not according to their language features (cf. Clear 1992). In terms of size, there is no consensus on how large a text sample must be in order to be representative of a given text as a whole. However, selecting or excluding certain parts of the text for the sample can influence the research results, given that certain features can be distributed differently in the text. Size is crucial in corpus linguistics, but for certain purposes, such as LSP studies, a smaller corpus of several hundreds of thousands of words can be sufficient. The general language corpus, on the other hand, is never finite enough to be adequate for a comprehensive description. In 1963, the one-million-word *Brown Corpus* (first large computerized corpus of American English for linguistic analysis) was considered representative, while the size of the web-based contemporary *COCA Corpus* is one billion words. Besides the difference in size, the content of the two corpora is inherently different as the first is systematically assembled, while the second is a web-derived corpus. Table 1 illustrates the composition of the Brown corpus (cf. Weisser 2016 :17).

Table 1. Composition of the Brown corpus

Label	Text Category/Genre	No. of Texts
A	Press: Reportage	44
B	Press: Editorial	27
C	Press: Reviews	17
D	Religion	17

Label	Text Category/Genre	No. of Texts
E	Skills & Hobbies	36
F	Popular Lore	48
G	Belles Lettres, Biography, Essays	75
H	Miscellaneous: Government Documents, Foundation Reports, Industry Reports, College Catalogue, Industry House Organ	30
J	Learned	80
K	General Fiction	29
L	Mystery & Detective Fiction	24
M	ScienceFiction	6
N	Adventure & Western Fiction	23
P	Romance & Love Story	29
R	Humour	9

As mentioned above, collecting a corpus implies the application of certain criteria to decide as objectively as possible which texts are representative of a particular genre, time period, language variety, and the like. For this reason, it is necessary to always explicitly state the parameters for compiling the corpus, that is, the selection of representative texts, the number of texts, the size of samples, and all other parameters that determine a corpus.

In addition to sampling, the principle of balance is an important criterion for compiling the corpus as it concerns the extent to which different text types, genres, and language modalities are represented. Depending on its function, the corpus can be composed of formal (manual, poem, newspaper article, novel, textbook, script, etc.) or informal (chats, forum posts, wiki pages, tweets, blogs, comments on social networks) genres. Ideally, the representation of genres in the corpus should be equal to the representation of genres in language use. Although most general corpora are composed of written texts of various genres, the tendency of the world's national corpora is to include up to 10% of spoken language samples collected from informal conversations. In spoken and written records, the balance can also pertain to the age, gender, and origin of the author. Another dimension of balance concerns the selection of texts. The corpus compiler should decide whether to choose a significant text or author who is influential or well-known, or to make a random selection of texts and authors, or to adapt the texts to meet the linguistic criteria. A combined approach that selects from a wider range of text types has proven to be best.

Although there is no universal recommendation regarding the size, texts, genres, and language modality in which the examples for the corpus are given, scholars agree on the importance of the third principle of representativeness which implies the quality of corpus data. This principle ensures the proportionate collection of diverse sources that consider specific verbal environment (co-text), as well as the situational and cultural parameters (context) according to corpus function. The

distinction of two types of corpus representativeness suggested by Egbert et al. (2022) sheds light on this important criterion. The first type, the domain representativeness, is defined as a set of text types selected to be included in the corpus according to their variability and relevance for the domain. The second type, the linguistic distribution representativeness, suggests the appropriateness of the selected data for the specific linguistic research goal.

Since corpus evidence can be reliable only as much as the corpus, it is important to specify the corpus design criteria before compilation starts so that corpus has principled underpinning. All linguists agree on one thing—there is no single corpus that would serve all purposes, and each corpus is only an approximate sample of linguistic variety we want to explore. That is why each corpus is compiled in particular ways with a specific purpose in mind.

2.2. Different Corpora for Different Purposes

In this section, we aim to provide an overview of the main corpus types used within Corpus Studies without attempting to create an exhaustive list. As mentioned above, the corpus type and its features depend on its primary purpose.

Before the corpus era, dictionaries were considered as the main source of reference and a gold standard for language use. With the advancement of corpora, as large samples of language, witnessing empirical evidence and based on principles of balance and representativity, they were soon accepted as reference material. This is the primary function of monolingual corpora, and nowadays all major languages, as well as many minor ones, have their own reference corpora available. According to Leech (2002) a "reference corpus is designed to provide comprehensive information about the language ...] It has to be a general corpus of wide coverage of the language, and hopefully it will be treated by its user community as some kind of "standard" for the language." Reference corpora thus contrast with specialised corpora dealing with a specific field of knowledge or a web corpus that illustrates a non-standard variety as it does not follow the linguistic design criteria.

The compilation of a multilingual corpus implies selecting sources from two or more languages that are either parallel or comparable. Parallel corpora suggest an aligned resource containing original texts and their translation, and they can be unidirectional (if the sources are primarily selected in language A and their translations in language B) or bidirectional (if the original texts are also selected in language B with their translations in language A). An example of bilingual unidirectional parallel corpus is the Pavia Corpus of Film Dialogue (PCFD), consisting of the transcriptions of 12 original film dialogues and their dubbing translations (Freddi and Pavesi 2009). An example of bilingual bidirectional parallel corpus is the English-Norwegian Parallel Corpus (ENPC) designed at the University of Oslo which consists of original texts and their translations, English to Norwegian and Norwegian to English. Large-scale multilingual parallel corpora such as the JRC-Acquis,

DGT-Acquis, EUR-Lex, and Europarl corpora are released by European Union organisations (Steinberger et al. 2014). A large collection of freely available parallel corpora is available at OPUS[1] as described by Tiedemann (2012). The CLARIN[2] infrastructure provides access to 82 parallel corpora (40 bilingual and 41 multilingual corpora) suitable for comparative research as many of them are also sentence-aligned. Some of the comparable corpora offered by CLARIN are social media corpora, corpora of academic texts, parliamentary corpora ParlaMint, newspapers corpora, and historical corpora. They contain original texts in two or more languages and share similar thematic, textual, discursive, and pragmatic parameters. Comparable corpora are suitable for contrastive studies but also for educating translators in the specialized domain to acquire specific disciplinary knowledge and its terminology. For this purpose, it is often necessary to compile a targeted DIY corpus. An interesting example of comparable corpora is the International Corpus of English (ICE) which comprises 27 national and regional varieties of English. Each of the corpora is considered comparable as they all follow the common corpus design, as illustrated in Table 2.

Table 2. The design of ICE corpora[3]

SPOKEN (300)	**Dialogues (180)**	**Private (100)**	**Face-to-Face Conversations (90)**
			Phone Calls (10)
		Public (80)	Classroom Lessons (20)
			Broadcast Discussions (20)
			Broadcast Interviews (10)
			Parliamentary Debates (10)
			Legal Cross-Examinations (10)
			Business Transactions (10)

1 Accessed January 13, 2024, https://opus.nlpl.eu/corpora

2 Accessed January 17, 2024, https://www.clarin.eu

3 The precise explanation of sampling criteria aims to illustrate the comparability between corpora: "Numbers in brackets indicate the number of 2,000-word texts in each category. The texts in the corpus date from 1990 or later. The authors and speakers of the texts are aged 18 or above, were educated through the medium of English, and were either born in the country in whose corpus they are included or moved there at an early age and received their education through the medium of English in the country concerned. The corpus contains samples of speech and writing by both males and females, and it includes a wide range of age groups. The proportions, however, are not representative of the proportions in the population as a whole: women are not equally represented in professions such as politics and law, and so do not produce equal amounts of discourse in these fields. Similarly, various age groups are not equally represented among students or academic authors." Accessed February 12, 2024, https://www.ice-corpora.uzh.ch/en/design.html

SPOKEN (300)	Dialogues (180)	Private (100)	Face-to-Face Conversations (90)
	Monologues (120)	Unscripted (70)	Spontaneous Commentaries (20)
			Unscripted Speeches (30)
			Demonstrations (10)
			Legal Presentations (10)
		Scripted (50)	Broadcast News (20)
			Broadcast Talks (20)
WRITTEN (200)	Non-Printed (50)	Student Writing (20)	Student Essays (10)
			Exam Scripts (10)
		Letters (30)	Social Letters (15)
			Business Letters (15)
	Printed (150)	Academic Writing (40)	Humanities (10)
			Social Sciences (10)
			Natural Sciences (10)
			Technology (10)
		Popular Writing (40)	Humanities (10)
			Social Sciences (10)
			Natural Sciences (10)
			Technology (10)
		Reportage (20)	Press News Reports (20)
		Instructional Writing (20)	Administrative Writing (10)
			Skills / Hobbies (10)

2.3. Corpus inquiry

The way to use corpora depends on the specific research questions, but also on the possibilities that the corpus can provide. Corpus investigation software allows various quantitative and qualitative analytical searches of corpus data, such as extracting word lists according to their frequency and distribution, generating part-of-speech and semantic annotations, concordances, thesaurus, calculating n-grams and clusters, producing different visualizations of corpus data, and many others.

Concordances, as the most accessible level of corpus use, have wide application in language teaching as they allow us to examine the occurrences and behaviour of different word forms. Using the KWIC format (Key Word in Context), learners can identify a listing of node words in a specific context. J. R. Firth's (1957) catchphrase is well known: "You shall know a word by the company it keeps." John Sinclair (1991) was the first to introduce the so-called 'idiom principle', according to which every speaker has a large number of semi-prepared or preconstructed phrases at his/her disposal. Native speakers use them unconsciously, while others have to adopt them

and learn to use them. It is precisely this aspect of usage that is the most difficult when acquiring a foreign language, and by examining the concordance, students can discover the repertoire of language specific patterns. This empirical approach emphasizes the mutual connection between grammar and lexis as advocated by Stubbs (2001: 18): "It is not the words that tell you the meaning of the phrase, but the phrase tells you the meaning of the individual words in it." While corpora provide empirical evidence on word and phrase frequency across registers, concordances help explore word and phrase meanings in context. However, it is important to understand that the corpus inquiry is only the first step towards discovering different aspects of language. As Weisser (2016: 9) points out, "once we actually have extracted some relevant data from a corpus, this is rarely ever the 'final product'. Such data generally either still needs to be interpreted, filtered, or evaluated as to its usefulness, if necessary, by (re-)adjusting the search strategy or initial hypotheses and/or conclusions, or, if it's to be used for more practical purposes, such as in the creation of teaching materials or exercises, to be brought into an appropriate form."

3. Corpora in Language Teaching and Learning

In the context of the increasingly widespread use of technological advances, language corpora have become an integral part of the new subfield of language teaching, Data-Driven Learning (DDL), focused on promoting autonomy and language awareness among learners. Numerous works dedicated to the use of electronic corpora in language teaching testify to the great potential of this method. The possible ways of using language corpora in teaching both the first and second language are very diverse and can be divided into several types. In the first case, a corpus is used as a supplement to dictionaries and grammars. Numerous authors (among many others, see Bernardini 2004; Boulton 2017, 2021; Boulton and Vyatkina 2021; Meunier 2002; Vyatkina 2020; Xu 2022) single out the use of corpora as a reference material in which certain language patterns and rules can be checked. For example, corpora have transformed the way language learners acquire vocabulary by placing it on the syntagmatic level and providing learners with real-world examples of words in context where a more nuanced understanding of word usage, collocations, and idiomatic expressions is offered. Corpora also enable learners to observe part-of-speech (POS) patterns in vast amounts of authentic sentences, which aids in grasping complex grammatical rules and syntactical nuances. The new emphasis on the interrelation between grammar and lexis has led to the discovery of a wide range of recurrent language-specific patterns and networks of paradigmatic and syntagmatic relations. By providing empirical evidence on word frequency, exploring meanings in context, and offering authentic language examples for contextual learning, teachers can enhance corpus-based pedagogical grammars and underlie the phraseological approach to pedagogy (Römer 2006; Vaughan and McCarthy 2016).

Apart from exposing students to native speaker corpora, the other important use of corpora refers to the possibility of identifying common mistakes made by language learners. This kind of learner corpora presents a rich and promising field, as demonstrated in previous research (Granger and Lefer 2020; Pérez-Paredes and Mark 2022; Granger and Lefer 2023). By analysing the language data in corpora, teachers can anticipate learners' errors and provide targeted feedback, ultimately helping students correct their linguistic shortcomings. One of the biggest learner corpora is the International Corpus of Learner English (ICLE)[4] created at the University of Louvain (Granger et al. 2020). It is the result of almost 30 years of international collaboration between universities, and contains essays written by upper-intermediate to advanced learners from 16 different non-native L1 backgrounds.

For reasons of space, many other significant techniques developed for exploiting language corpora have not been mentioned, but interested readers can refer to the corpus-based pedagogy area of research that is still expanding. New areas of learners' use of corpora in Second Language Writing (SLW), such as corpus-aided writing and mobile assisted language learning (MALL), are highlighted by Schmidt (2023). Various corpus-based case studies on teaching English for Academic Purposes (EAP) writing and disciplinary writing are presented by Flowerdew (2022). Recent studies (Szudarski 2023) also bring important advances in learning collocations in second language acquisition based on corpus use.

However, despite the existence of the mentioned possibilities, as well as many theoretical works dealing with them, corpora are still rarely integrated in language learning practices. At this point, we may note that educating prospective DDL practitioners is essential. Previous findings (Breyer 2009; Ebrahimi and Faghih 2017; Leńko-Szymańska 2017; Chen et al. 2019; Lin 2019) showed the importance of implementing DDL methods and techniques in general teacher education programmes. The adoption of technology implies not only introducing user-friendly corpora tools and focusing on hands-on experience but also encouraging direct and indirect use of authentic data with emphasis on the pedagogical knowledge for exploiting corpus results. It is therefore important to incorporate practical aspects of corpora throughout the curriculum and to equip students to integrate DDL in their future teaching.

4. Corpora in Translator Education

The results of corpus linguistic studies have been applied in translation regarding particular methods (among many others, see Bowker 1998; Dash and Ramamoorthy 2019; McEnery and Wilson 1997; Olohan 2002; Sinclair 2003, 2004a) as well as in terms of the general concept of translation, with special reference to translation competence issues (Bernardini 2022; Johansson 1998; Martín 2014; Pietrzak 2015;

4 Accessed February 23, 2024, https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Zanettin 2014). By accessing a vast repository of authentic, representative language data, translators' education can include different browsing activities such as frequency word-listing, producing and interpreting concordances, identifying patterns, semantic preferences, linguistic nuances, and common mistranslations, extracting keywords or key terms, sorting results, leading to more accurate and contextually appropriate translations. Additionally, corpus-based approaches allow the error annotation offering valuable insights for developing reflective activities about translator strategies. Stimulating curiosity about language and searching for answers in corpora is essential as highlighted by Bernardini and Castagnoli (2008: 44): "Indeed, successful corpus work requires first and foremost an inquisitive frame of mind, a critical attitude and an ability to detect patterns, and only secondarily (some) technical skills."

The development of the field of corpus-based translations studies (CBTS) started with Mona Baker's (1993) seminal work and the creation of the first Translation English Corpus (TEC). The empirical corpus studies of translation were first mainly oriented towards the typical translation regularities, such as the study of language simplification, lexical creativity, or stylistic preferences, which led to a series of studies devoted to translation universals (Baker 1995; Laviosa 1997), the translator's style (Baker 1999), and translation norms (Kenny 1998). While looking for translation universals, independent of language pairs, interesting culture-specific aspects were discovered, which contributed to the advancement of intercultural translation studies. Very soon the field was extended by Shlesinger (1998) into corpus-based translation and interpreting studies (CTIS) and continued to grow in three directions: literary translation, translation theory, and intercultural studies. Each of them relies on monolingual or bilingual/multilingual corpora, regardless of whether they are comparable or parallel.

Translation or parallel corpora were recognized as a valuable source for understanding the cultural nuances in translation, as they are likely to reveal patterns and detect unexpected cross-linguistic equivalents. Vilceanu (2019: 1483) reports that "(...) by their very nature, translation or parallel corpora are made up of written texts belonging to a specific genre and type of discourse, which can be considered both an advantage (in the sense of allowing for in-depth analysis and interpretation of findings as guidelines for quality assurance in translation) and a disadvantage (their application is limited to certain cross-linguistic studies)." The biggest limitation of parallel corpora is that they are rarely available for the specific language pair and specific subject.

According to Tognini Bonelli (2001: 133), the translator should ideally consider evidence of both parallel and comparable corpora. The same is stated by Bernardini (2022: 493): "To fully exploit the potential of parallel and comparable corpora, these should be used together: parallel corpora (from the public domain) may provide suggestions about translator strategies and translation equivalents, while (self-made) specialized comparable corpora of non-translated target language texts may

be used to (dis)confirm the general currency of the choices made by translators." While parallel, sentence aligned corpora can be used for developing hypotheses, comparable corpora allow the possibility of testing them (through the insight into the two similar samples of L1 and L2). From the viewpoint of the translator and translation students, both types of corpora present invaluable sources for comparing and revealing not only the degree of mutual correspondence between lexical items but also general cross-linguistic similarities and differences.

5. Conclusion

In this paper, we have given a very brief insight into the advantages of corpus methodology, simply to 'set the scene', rather than to provide an extensive coverage of the multitude of corpus applications. Due to space constraints, we were unable to cover here all the variety of corpus linguistic issues, so the goal was to provide the basis for understanding how corpus investigation can be integrated into language learning and translation. More extensive coverage of the theoretical and practical issues is available in manuals like Beeby et al. (2009), Crawford and Csomay (2016), Facchinetti (2007), Hunston (2002), Ji et al. (2016); Partington (1998), Pérez-Paredes and Mark (2021), Sinclair (2003, 2004b), Stubbs (2001), Szudarski (2023), Tognini-Bonelli (2001), Weisser (2016), Zanettin (2012) among others.

As noted at the beginning of this paper, data- or corpus-driven methodology had the fundamental role in LLM training and is at the core of the main technological innovations we are witnessing. By describing some of the multiple exploitation possibilities of the corpus, we aimed to illustrate how learning to use corpora changes the way students perceive and understand languages. The evidence feedback that can be provided from corpora brings reassurance about language and makes it easy to detect cross-linguistic features as well as innacuracies and incompatibilities. The benefits of thrustworthy corpus data is even more enhanced in the age of deep learning and big data. Along with the countless opportunities of AI tools, it is vital for learners to have access to authentic and attested information for a reliable analysis to be made.

Furthermore, this enhances even more the importance of the implementation of DDL methods and techniques in general teacher training programmes to equip students with the technical and pedagogical skills needed to exploit the existing corpora and create their own corpus-based and corpus-driven material suitable for language teaching and translation tasks. By achieving a more realistic learning experience, learners will hopefully keep the curiosity towards further research into language features and develop inspiration for discovering its potential.

References

Absalom, Matthew. 2021. "Digital corpora: language teaching and learning in the age of big data." In Beaven, T. & Rosell-Aguilar, F. (Eds.) *Innovative language pedagogy report*. 97–101.

Ackerley, Katherine; Coccetta, Francesca. 2007. "Enriching language learning through a multimedia corpus." *Recall* 19(3). 351–370.

Baker, Mona. 1993. "Corpus linguistics and translation studies: Implications and applications." In Francis, G. & Tognini-Bonelli, E. (Eds.) *Text and technology: In honour of John Sinclair*. Amsterdam: Benjamins. 233–250.

Baker, Mona. 1995. "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research." *Target* 7(2). 223–243. doi: 10.1075/target.7.2.03bak

Baker, Mona. 1999. "The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators." *International Journal of Corpus Linguistics* 4(2). 281–298. doi: 10.1075/ijcl.4.2.05bak

Bernardini, Silvia. 2004. "Corpora in the classroom. An overview and some reflections on future developments." In Sinclair, J. M. (Ed.) *How to use corpora in language teaching*? Amsterdam/Philadelphia: John Benjamins. 15–36.

Bernardini, Silvia; Castagnoli, Sara. 2008. "Corpora for translator education and translation practice." In Yuste, E. (Ed.) *Topics in Language Resources for Translation and Localisation*. 39–57. John Benjamins.

Bernardini, Silvia. 2022. "How to use corpora for translation?" In 0'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge Handbook of Corpus Linguistics*. 485–498. doi: 10.4324/9780367076399-34

Beeby, Allison; Rodríguez Inés, Patricia; Sánchez-Gijón, Pilar. (Eds.) 2009. *Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: Benjamins.

Boulton, Alex. 2017. "Corpora in language teaching and learning". *Language Teaching* 50(4). 483–506. doi: 10.1017/S0261444817000167

Boulton, Alex. 2021. "Research in data-driven learning." In Pérez-Paredes, P. & Mark, G. (Eds.) *Beyond the concordance: Corpora in language education*. John Benjamins. 9–34. doi: https://doi.org/10.1075/scl.102.01bou

Boulton, Alex; Vyatkina, Nina. 2021. "Thirty years of data-driven learning: Taking stock and charting new directions over time." *Language Learning & Technology*. 25(3), 66–89.

Breyer, Yvonne. 2009. "Learning and teaching with corpora: reflections by student teachers." *Computer Assisted Language Learning* 22(2). 153–172.

Bowker, Lynne. 1998. "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study." *Meta* 43(4). 631–651. doi: 10.7202/002134ar

Braun, Sabine. 2007. "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora." *Recall* 19/3. 307–328.

Chen, Meilin; Flowerdew, John; Laurence, Anthony. 2019. "Introducing in-service

English language teachers to data-driven learning for academic writing." *System* 87. 102–148. doi: https://doi.org/10.1016/j.system.2019.102148

Clear, Jeremy. 1992. "Corpus sampling." In Leitner, G. (Ed.) *New directions in English language corpora*. Berlin: Mouton de Gruyter. 21–31.

Crawford, William J.; Csomay, Eniko. 2016. *Doing Corpus Linguistics*. Oxford: Routledge.

Dash, Niladri Sekhar; Ramamoorthy L. 2019. "Corpus as a Primary Resource for ELT." In Dash, N. S. & Ramamoorthy, L. (Eds.) *Utility and Application of Language Corpora*. Singapore: Springer. 91–103. doi: 10.1007/978-981-13-1801-6

Ebrahimi, Alice; Faghih, Esmail. 2017. "Integrating corpus linguistics into online language teacher education programs." *ReCALL* 29(1). 120–135.

Egbert, Jesse; Biber, Douglas; Gray, Bethany. 2022. "A Practical Framework for Corpus Representativeness." In Egbert, J., Biber, D. & Gray, B. (Eds.) *Designing and Evaluating Language Corpora*. Cambridge University Press. 52–67. doi: https://doi.org/10.1017/9781316584880.003

Facchinetti, Roberta. (Ed.). 2007. *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi.

Firth, John Rupert. 1957. *Papers in Linguistics 1934-1951*. London: Oxford.

Flowerdew, Lynne. 2022. "Using corpora for writing instruction." In 0'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge Handbook of corpus linguistics*. 444–457. doi: 10.4324/9780367076399-31

Freddi, Maria; Pavesi, Maria. 2009. "The Pavia Corpus of Film Dialogue: Methodology and Research Rationale." In Freddi, M. & Pavesi, M. (Eds.) *Analysing Audiovisual Dialogue. Linguistic and Translation Insights*. Bologna: CLUEB. 95–100.

Granger, Sylviane; Dupont, Maïté; Meunier, Fanny; Naets, Hubert; Paquot, Magali. 2020. *The International Corpus of Learner English. Version 3*. Louvain la-Neuve: Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877

Granger, Sylviane; Lefer, Marie-Aude. 2020. "The Multilingual Student Translation corpus: a resource for translation teaching and research." *Language Resources and Evaluation* 54(4). 1183–1199. doi: 10.1007/S10579-020-09485-6

Granger, Sylviane; Lefer, Marie-Aude. 2023. "Learner translation corpora." *International journal of learner corpus research* 9(1). 1–28. doi: 10.1075/ijlcr.00032.gra

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge University Press.

Ji, Meng; Oakes, Michael; Defeng, Li; Hareide, Lidun. 2016. *Corpus Methodologies Explained. An empirical approach to translation studies*. Oxford: Routledge.

Johansson, Stig. 1998. "On the role of corpora in cross-linguistic research." In Johansson, S. & Oksefjell, S. (Eds.) *Corpora and Cross-linguistic Research*. Amsterdam and Atlanta: Rodopi.1–24.

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

Johns, Tim. 1991. "Should you be persuaded: Two samples of data-driven learning materials." In Johns, T & King, P. (Eds.) *Classroom concordancing. English*

Language Research Journal 4. 1–16.
Kenny, Dorothy. 1998. "Creatures of Habit? What Translators Usually Do with Words." *Meta* 43(4). 515–523. doi:10.7202/003302ar
Laviosa, Sara. 1997. "How Comparable Can 'Comparable Corpora' Be." *Target* 9(2). 289–319. doi:10.1075/ target.9.2.05lav
Leech, Geoffrey. 2002. *The Importance of Reference Corpora*. Invited speech at the conference of ZIO Corpus. University of the Basque Country.
Leńko-Szymańska, Agnieszka. 2017. "Training teachers in data-driven learning: Tackling the challenge." *Language Learning & Technology* 21(3). 217–241.
Lin, Ming Huei. 2019. "Becoming a DDL teacher in English grammar classes: A pilot study." *The Journal of Language Teaching and Learning* 9(1). 70–82.
Martín, Ricardo Muñoz. 2014. "Situating Translation Expertise: A Review with a Sketch of a Construct." In Schwieter, S. & Ferreira, A. (Eds.) *The Development of Translation Competence: Theories and Methodologies from Psycholinguistics and Cognitive Science*. Cambridge Scholars Publishing. 2–56.
McEnery, Tony; Wilson, Andrew. 1997. "Teaching and Language Corpora (TALC)." *ReCALL* 9(1). 5–14. doi: 10.1017/S0958344000004572
Meunier, Fanny. 2002. "The pedagogical value of native and learner corpora in EFL grammar teaching." In Granger S., Hung J. & Tyson, S. (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Benjamins: Amsterdam/Philadelphia. 119–142.
Olohan, Maeve. 2002. "Corpus Linguistics and Translation Studies: Interaction and Reaction." *Linguistica Antverpiensia* 1. 419–429. doi: 10.52034/LANSTTS.V1I.29
Partington, Alen. 1998. *Patterns and Meanings: Using corpora for English language research and teaching*. John Benjamins Publishing.
Pérez-Paredes, Pasqual; Mark, Géraldine. 2022. "What can corpora tell us about language learning?" In 0'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge Handbook of corpus linguistics*. 313–327. doi: 10.4324/9780367076399-22
Pietrzak, Paulina 2015. "Translation competence." In Bogucki, Ł., Góźdź-Roszkowski, S. & Stalmaszczyk, P. (Eds.) *Ways to translation*. Łódź-Kraków: Łódź University Press & Jagiellonian University Press. 317–338.
Römer, Ute. 2006. "Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments." *Zeitschrift für Anglistik und Amerikanistik* 54(2). 121–134. doi: https://doi.org/10.1515/zaa-2006-0204
Shlesinger, Miriam. 1998. "Corpus-based interpreting studies as an offshoot of corpus-based translation studies." *Meta* 43 (4). 486–493.
Schmidt, Nicole. 2023. "Unpacking second language writing teacher knowledge through corpus-based pedagogy training." *ReCALL* 35(1). 40–57. doi:10.1017/S0958344022000106
Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, John. 2003. *Reading Concordances*. London: Longman.

Sinclair, John. 2004a. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, John (Ed.). 2004b. *How to use corpora in language teaching*? Amsterdam/Philadelphia: John Benjamins.

Steinberger, Ralf; Ebrahim, Mohamed; Poulis, Alexandros; Carrasco-Benitez, Manuel; Schlüter, Patrick; Przybyszewski, Marek; Gilbro, Signe. 2014. "An overview of the European Union's highly multilingual parallel corpora." *Lang Resources & Evaluation* 48, 679–707. doi : https://doi.org/10.1007/s10579-014-9277-0

Stubbs, Michael. 2001. *Words and phrases: Corpus studies of lexical semantics.* Oxford: Blackwell.

Szudarski, Pawel. 2023. *Collocations, Corpora and Language Learning.* Cambridge University Press. doi: 10.1017/9781108992602

Tiedemann, Jorg. 2012. "Parallel data, tools and interfaces in OPUS." In *Proceedings of LREC*. Istanbul, Turkey.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. John Benjamin Publishing.

Vaughan, Elaine; McCarthy, Michael. 2016. "Research in Corpora in Language Teaching and Learning." In Hinkel, E. (Ed.). *Handbook of Research of Second Language Teaching and Learning*. Oxford: Routledge. doi: 10.4324/9781315716893.CH13

Vyatkina, Nina. 2020. "Corpora as open educational resources for language teaching." *Foreign Language Annals*. 1–12. doi: 10.1111/FLAN.12464

Weisser, Martin. 2016. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. Wiley Blackwell.

Xu, Jiajin. 2022. "A Historical Overview of Using Corpora in English Language Teaching." In Jablonkai, R. R. & Csomay, E. (Eds.) *The Routledge Handbook of Corpora and English Language Teaching and Learning.* doi: 10.4324/9781003002901-3

Zanettin, Federico. 2014. "Corpora in Translation." In House, J. (Ed.) *Translation: A Multidisciplinary Approach*. Palgrave Macmillan. 178–199. doi: 10.1057/9781137025487_10

Zanettin, Federico. 2012. *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Oxford: Routledge.

Ivana Havelka

University of Vienna
ivana.havelka@univie.ac.at

ChatGPT & Co: AI Transforming Terminological Preparation in Interpreting

Abstract

Interpreting, as an interdisciplinary field, inherently requires specialized terminology due to its transdisciplinary applications (Lušicky 2019). Terminological preparation is thus a prerequisite for each interpreting assignment. With the advent of new AI tools, this process is undergoing a transformative shift. ChatGPT and similar tools are already used in academia (e.g., Halaweh 2023; Passmore and Tee 2023) but might become more prevalent in terminological preparation (Hsu 2023). This case study delves into the implementation and benefits of chatbots in terminological preparation. Beyond merely providing terminology lists, ChatGPT and Co support the process by facilitating an interactive dialogue that aids in research. This approach not only streamlines terminology extraction but also fosters a deeper understanding and tailored adaptation to specific project needs. The findings of this study show that AI-powered solutions hold the potential to significantly enhance terminological preparation, paving the way for future innovations in the field.

Keywords: interpreting, terminology, terminological preparation, chatbot

1. Introduction

As a form of communication, interpreting aims to bridge gaps in understanding. It is inherently interdisciplinary, nearly always occurring in a specific field or facilitating the specialized communication between interlocutors. Interpreting assignments occur in medical, legal, economic, or any other contexts, where communication takes place among professionals or between professionals and lay persons. As a result, interpreters are continuously involved in exploring the communicative settings, specialized communication, and terminology across various domains. This often presents unique challenges for interpreters, as they may not possess the same level of expertise in the specific field as expert interlocutors. To minimize this discrepancy, thorough understanding of the assignment (e.g., involved interlocutors and the purpose of communication) and terminological preparation are essential prerequisites for each interpreting assignment (Chiocchetti et al. 2023; Lušicky 2019). Terminological preparation is not only necessary for facilitating specialized communication; it also serves as a tool for quality assurance and can thus

be considered a measure of risk management (KÜDES 2018: 30; Lušicky 2019: 67). Quality in interpreting is enhanced, among other factors, by the identification and use of terminological equivalents. Terminological equivalents ensure adequacy and accuracy. Depending on its subject and purpose, terminology work can yield different outcomes (Ramos and Guzmán 2023: 376). Some terminology work requires introducing neologisms or technical expressions in order to fill terminological gaps (KÜDES 2018: 70, 82). Terminological preparation not only ensures accuracy in terms of terminology but also contributes to a more seamless and fluid interpreting process (Lušicky 2019: 67). Additionally, the immediate nature of interpreting means that preparatory work in terminology significantly relieves the interpreter's cognitive load.

Terminology work varies in approach and content based on its objectives and available methods. Depending on the purpose, the choice is between ad hoc research for rapid problem resolution, or extensive documentation of a subject area's terminology through systematic conceptual comparison (Engberg 2023). The third form of terminology research is text-related terminology (KÜDES 2018: 66). Documenting the results of terminology research helps save time and reduce stress by eliminating redundant efforts, benefiting future assignments and colleagues. Given that terminological preparation requires considerable time yet is essential for interpreters' work, the use of large language models (LLMs) should be assessed, especially in light of recent advancements in AI (e.g., Russell and Norvig 2016). Large language models incorporated in chatbots have been tested in academic education (Halaweh 2023; Passmore and Tee 2023), but only a few studies have focused on their terminological application (Ahn 2023; Hsu 2023). This article introduces a case study that evaluates the use of LLM-powered chatbots in the legal terminological preparation of interpreters. It focuses on the efficiency of the legal terminological preparation process, the adequacy and accuracy of results, as well as their user-oriented features.

2. Digital Literacy for Interpreters

While the use of terminology databases and management programs is already common practice for translators, interpreters have only begun to gradually adopt technology in the last decade. Interpreters typically use technology for transmitting their interpretation, such as equipment for telephone (e.g., Kelly and Pöchhacker 2015) or video interpreting (e.g., Braun 2019; Seeber and Fox 2022), as well as various tools to aid during the interpreting process (e.g., Corpas Pastor and Dúran-Muñoz 2018; Fantinuoli 2018). This second category is broad and encompasses a variety of supports within the interpreting process. Computer-assisted interpreting (CAI) tools, such as terminology databases, aid in the preparation, during interpreting, and the review phases of terminology management. Other CAI tools offer capabilities for digital note-taking or simultaneous-consecutive (Sim-Cons) mode,

demonstrating the wide range of technological applications designed to enhance interpreting efficiency and reduce the cognitive load.

In a 2021 study, Iacono et al.[1] (2021) showed how interpreters adapt and respond to changing working conditions. The study reveals that 84.27% of interpreters use digital media for assignment preparation and follow-up, greatly enhancing efficiency. Digital tools also play a crucial role in storing data (81.46%), managing tasks (73.88%), and facilitating digital communication with colleagues (74.72%) and clients (70.51%). Additionally, 70.22% of interpreters offer remote audio and video interpreting services, adapting to the demand for modern communication technologies. Although less common, on-site interpreting with technical support is still practised by 37.64%, reflecting the diversity of interpreting settings. Despite the digital trend, a significant majority (90.17%) show a clear preference for traditional note-taking with a notepad and pen over digital note-taking with a tablet and stylus or Sim-Cons mode. For terminology preparation, a substantial number of interpreters rely on digital dictionaries (83.15%) and online databases (77.25%), yet a considerable portion (37.08%) still uses physical dictionaries and Excel lists. The study also notes the value of terminology databases and crowd intelligence as resources, indicating a blend of digital and traditional approaches in the interpreting field. These data suggest that interpreters choose a mix of proven and innovative tools to meet the demands of their profession.

Digital media significantly support the organizational, communicative, and professional aspects of the interpreter's work. As technology becomes increasingly integral to society, the importance of digital competencies has been recognized by the European Commission, which has proposed a framework outlining five key competency areas (Kluzer et al. 2018): Information and Data Literacy, Communication and Collaboration, Digital Content Creation, Safety, and Problem Solving. These areas align with the aforementioned applications of technology among interpreters. Within the competency area of Information and Data Literacy, the following skills are essential for interpreters (Havelka et al. 2021: 223):

- Searching for and filtering data and terminology related to the assignment,
- Analysing data and terminology related to the assignment,
- Selecting an appropriate terminology and data management system and implementing it into the interpreting process (including preparation and follow-up).

1 The online survey was conducted between March 23 and April 23, 2021. Altogether, 356 interpreters, primarily with German in their language combination, completed the survey (for further details on the study and its methodology, refer to Iacono et al. 2021).

3. ChatGPT and Co

The launch of LLMs through multilingual chatbots such as ChatGPT (Chat Generative Pre-Trained Transformer) (OpenAI 2024b), Gemini (Google 2024), Microsoft Copilot (Microsoft 2024), and Perplexity.ai (Perplexity 2024) has marked a milestone in the digital era. Large language models are pre-trained with vast amounts of information about language and the world. This enables them to generate texts, summaries, and perform machine translation by processing natural language. Notably, they can answer questions and powering chatbots (Jurafsky and Martin 2024: 214). They can be used to perform linguistic or stylistic text optimizations, such as grammar checking, style editing, and register adjustment. LLMs can perform tasks such as creating preliminary outlines by generating a list of topics and subtopics. They can also take on time-consuming tasks in the writing process, such as researching information or organizing ideas. LLMs are becoming increasingly capable of generating text-type-specific texts, such as news articles, blog posts, and academic papers. This type of text generation, known as mechanical writing, is understood as a schematic and functional type of writing. In contrast, critical thinking and writing is a creative act that emerges in the course of a writing process, such as analysing complex ideas, developing arguments, and writing persuasively (Bishop 2023).

ChatGPT was the first LLM-powered chatbot to be made available to the general public for the first time in November 2022 (AlZaabi et al. 2023). As a chatbot, ChatGPT utilizes a dialogical approach for both user interaction and knowledge queries. The human-machine interaction takes place within a task-based dialogue system (Jurafsky and Martin 2024: 320–21). Similar to ChatGPT, other LLM-powered chatbots like Google Gemini, Microsoft Copilot, and Perplexity.ai also employ this approach. Currently supporting over 50 languages, ChatGPT automatically adjusts its output language to match the user's input. This allows for seamless conversations across various languages. The input is also referred to as a prompt (see for more Chapter 12 in Jurafsky and Martin 2024). Prompt engineering describes the process of formulating a prompt by specifying a role for the chatbot to adopt, an action the chatbot should perform, and by delineating the expected outcome (OpenAI 2024a).

Trained on massive multilingual datasets, chatbots recognize not only text but also speech and images, further enhancing their ability to interact with users in natural and diverse ways (Haleem et al. 2022). ChatGPT, Gemini, Microsoft Copilot, and Perplexity.ai offer the possibility to resume older conversations and topics. In this way, discussed topics are not lost (AlZaabi et al. 2023) but content can be fine-tuned. The ability to upload and analyse files was introduced with the GPT-4 version, which was released in March 2023. Depending on the chatbot, the generated texts or images can be exported, or the generated content can be shared via a link. Through the process of terminology extraction, these innovative systems are capable of automatically detecting terms within texts. Such technology proves crucial

for the creation of terminology databases or the translation of specialized texts, thus offering essential tools for experts across a multitude of disciplines. For most interpreters, terminology work involves acquiring new terminology. In the course of preparing the terminology, details such as the linguistic properties of words, e.g., grammatical gender, and the context of the term, can be methodically displayed in a table format. Also, crowdsourcing, a phenomenon associated with social media and the internet generation, can be used in the search for terminology through micro-queries. Additionally, LLMs leverage the power of swarm intelligence, often referred to as the wisdom of the crowd (Bishop 2023).

The greatest disadvantages of LLMs lie in the loss of human autonomy due to the opacity of the background processing workflows (van Dis et al. 2023: 224). Further disadvantages arise from so-called hallucinations and imprecise statements, which are especially problematic in text production and knowledge queries. Therefore, an application where the truthfulness or correctness of the information is crucial is not recommended (Jurafsky and Martin 2024: 240). Human verification is necessary in all cases. However, correcting these errors requires specialized knowledge (van Dis et al. 2023: 224). In this regard, the issue of missing sources is a fundamental concern. These sources are provided in the paid version of ChatGPT-4. Recently, Microsoft Copilot, Gemini, and Perplexity.ai have started to offer sources as well. Nonetheless, the lack of sources and transparency concerning the origin of information—whether it comes from single or multiple sources—remains crucial. This situation leads to the original authors of the content being unknown and unverifiable to the end user. Jurafsky and Martin (2024: 240) recommend that language models include datasheets or model cards, offering comprehensive and replicable information about the training corpora. Another issue commonly associated with language models is the reinforcement of content through written text data in widely spoken languages, such as English. As a result, the use of LLMs is associated with priming and biased information (van Dis et al. 2023: 224).

4. Case Study: Chatbots as Terminological Tools

The performance of LLMs can be tested in two ways: extrinsic and intrinsic evaluation (Jurafsky and Martin 2024: 38). Due to the limited scope of the present study, a simplified version of the extrinsic evaluation will be utilized in this case study (Silverman 2024: 70–73). Extrinsic evaluation examines the improvement and efficiency of an LLM within a specific workflow. For this case study, workflow steps during ad hoc terminology research are adapted to the interpreting assignment (KÜDES 2018: 66). By following these steps consistently, it allows for a direct comparison among various chatbots. For the case study, an interpreting assignment in divorce proceedings under Austrian law is used as an example. In this fictional setting, one of the applicants speaks exclusively Croatian, resulting in the need for an interpretation for the Croatian–German language pair.

The analysis is performed based on chat logs for each LLM-powered chatbot. The results of the following workflow steps are examined in comparison, by evaluating their impact on the adequacy and accuracy of the results and user-oriented features. In this study, user-oriented features include free access, the possibility of importing and exporting files, a public link for sharing, and the provision of sources and visuals, as is shown in Table 1.

Table 1. User-oriented features

Chatbot	Free access	Importing files	Exporting files	Public link for sharing	Sources	Visuals
ChatGPT-4	ChatGPT 3.5 is free	✓	✓	✓	✓– very few	–
Gemini	✓	–	✓	✓	✓– very few	–
Microsoft Copilot	✓	✓	✓	✓– yes, but only for one prompt each	✓	–
Perplexity.ai	✓	✓	–	✓	✓	✓

While adequacy refers to the specific context and function of interaction (Steps 1 and 2), accuracy means the conceptual equivalence between the source and target languages (Steps 3 and 4).

The following steps were applied:

Step 1. In dialogue interpreting, knowing the context, communication mode (in-person, video, or phone), meeting purpose (business, medical, legal), interpreting mode (simultaneous or consecutive), and conversation structure or protocols (e.g., court hearing, medical examination) is essential.

Step 2. Identifying participants and their communication purpose entails understanding the interlocutors (professionals with professionals or laypersons) and their objectives, helping to define roles and refine terminology.

Step 3. Anticipating the terminological relevance of documents like presentations, reports, contracts, and technical materials is key to interpreting preparation. Identifying relevant background or parallel texts and understanding the communication's register (formal, informal, technical, specialized) based on the subject and participants is crucial.

Step 4. The next step involves extracting key terms from this information and compiling a table of the most critical terms, their definitions, and translations into the target language.

To generate this information through LLMs, it is necessary to formulate appropriate prompts. A prompt is delineated by assigning a specific role, which in turn

narrows down the scope of the task at hand. The selection of a role not only specifies the nature of the task but also establishes the register of the language to be used. This approach ensures that the language model adopts a tone and style appropriate to the intended function and context of the interaction.

For the current study, the following prompts were formulated in German and applied in the LLM-powered chatbots: ChatGPT (Open AI), Gemini (Google), Microsoft Copilot, and Perplexity.ai. The workflow steps and formulated prompts are given in Table 2.

Table 2. Workflow steps and prompts

Setting and context	
German	Antworte als Experte für Scheidungen. Nenne wesentliche Informationen zu einem Scheidungsverfahren, nenne welche Arten einer Scheidung das österreichische Gesetz vorsieht und welche Voraussetzungen dafür das österreichische Gesetz vorsieht?
English	Act as a divorce expert. Outline key information about divorce proceedings, describe the types of divorce provided for under Austrian law, and detail the prerequisites established by Austrian law for these divorces.
Participants/Interlocutors	
German	Antworte als Scheidungsexperte. Wer ist in einer Gerichtsverhandlung zu einer einvernehmlichen Scheidung anwesend. Welche Aufgaben haben die einzelnen Verfahrensbeteiligten? Antworte nacheinander.
English	Act as a divorce expert. Explain who is present at a court hearing for an uncontested divorce. Describe the specific duties of each participant in the proceedings. Answer the questions in order.
Documents and parallel texts	
German	Antworte als Scheidungsexperte. Welche Dokumente sind nach österreichischem Gesetz wesentlich für eine einvernehmliche Scheidung? Erstelle einen fiktiven Ablauf einer einvernehmlichen Scheidungsverhandlung.
English	Act as a divorce expert. What documents are essential according to Austrian law for an uncontested divorce? Create a fictional scenario of how an uncontested divorce hearing would typically unfold.
Term extraction and table	
German	Antworte als Terminologieexpertin: Erstelle eine Tabelle mit den 15 wichtigsten Termini aus diesem Chatverlauf zum Thema Scheidung. Diese Termini sollen zur terminologischen Vorbereitung dienen. Bereite die Termini in einer Tabelle auf, mit den Spalten Deutsch, Definition des Terminus auf Deutsch sowie die Entsprechung in kroatischer Sprache sowie eine Definition in kroatischer Sprache.
English	Act as a terminology expert. Create a table with the 15 most important terms from this chat log related to divorce. These terms will be used for terminological preparation. Prepare the terms in a table, with columns for German, the definition of the term in German, the equivalent in Croatian, and the definition in Croatian.

The performance of the chatbots is given in Table 3.

Table 3. Performance of chatbots

Chatbot	Public chatbot link	Term extraction and translation by chatbots	Participants	Documents and parallel texts
ChatGPT-4	https://chat.openai.com/share/8bb658bb-915f-4d9b-9a2e-87286dab2166	Einvernehmlich Scheidung – sporazumni ravod Obsorge – skrbništvo Unterhalt – uzdržavanje Vermögensaufteilung –podjela imovine Umgangsrecht – pravo na kontakt	✓	✓
Gemini	N/A	Vermögensaufteilung – podjela imovine; Obsorge – skrbništvo; Rechtsmittelfrist – rok za žalbu; Verfahrenswert – vrijednost predmeta spora; Gerichtsgebühren – sudske takse	✓	✓
Microsoft Copilot	https://sl.bing.net/cr9qOOF59KS	Trennungszeit – razdoblje odvojenog života; Obsorge – skrbništvo; Ehewohnung – bračni stan; Rechtskraft des Urteils – pravomoćnost presude; Ehepartner – supružnici	✓	✓
Perplexity.ai	https://www.perplexity.ai/search/Antworte-als-Experte-GBsKgERUWqil2GvA15Bw?s=c	Einvernehmliche Scheidung – sporazumni razvod; Bezirksgericht – općinski sud; Elternberatung – savjetovanje roditelja; Ehevertrag – bračni ugovor; Aufenthaltsbestimmungsrecht – pravo određivanja prebivališta	✓	✓

5. Results and Discussion

Basic information, such as documents, participants, and background information, was provided by all chatbots. Better research capabilities are achieved through interactive dialogue. The interactive functionality allows users to access additional information and references during the work process. This supports interpreters' terminology research skills in different languages and contributes to improved terminology preparation. In this regard, prompt engineering has become crucial, showing that without linguistically adequate and goal-oriented input, meaningful output cannot be anticipated. This step is new compared to traditional terminology work and requires an understanding of how chatbots work. It shows that digital competencies are essential.

When it comes to searching for and filtering data and terminology related to the assignment, Microsoft Copilot provided a dialogic representation of the court

hearing, unlike other chatbots that primarily provided summaries of court proceedings. This distinction can prove advantageous in the preparation phase of an interpreting assignment.

In the context of analysing data and terminology pertinent to interpreting assignments in legal contexts, it is essential to consider the legal framework, as demonstrated by the application of Austrian law in this case study focusing on interpreting at a divorce hearing. The Austrian legal language needs to be incorporated into terminology research. However, challenges arose when conducting terminology work, both in the source language German and in the target language Croatian. For instance, ChatGPT-4 failed to provide meaningful suggestions for terms like *Zerrüttungsprinzip* (irretrievable breakdown of marriage as grounds for a divorce), which could be described in Croatian as *teški i trajno poremećni bračni odnosi*. Moreover, ensuring transparency of sources is crucial. Sources are not consistently provided, which complicates the process of verifying information accuracy and adequacy.

This issue is particularly prominent in legal contexts, where reliance on authoritative sources is indispensable. Perplexity.ai and Microsoft Copilot provided authoritative sources (at least in German), while ChatGPT-4 and Gemini did not. The lack of authoritative sources can pose a problem for terminology work, especially in legal language, where instead of consulting authoritative sources, general language datasets are used. This hinders serious terminology work and ultimately accuracy.

The analysis of four different chatbots has demonstrated that each one produces fundamentally similar but distinct results upon closer examination. Selecting an appropriate terminology and data management system and integrating it into the interpreting process (including preparation and follow-up) is a crucial aspect of comprehensive terminology research.

To ensure future usability, the results of terminology research must be exportable in a structured format and able to be imported into a terminology database. However, certain limitations were encountered; while Perplexity.ai lacked export capability altogether, the paid version of ChatGPT-4 was hindered by usage caps, preventing file export. Generally, ChatGPT-4 can export to Excel sheets, and this functionality is also available in Gemini and Microsoft Copilot. Although all chatbots allow for data and terminology extraction, the choice of chatbot may be influenced by factors such as export capabilities and usage limitations. Furthermore, all chatbots facilitate collaborative terminology work by offering the option to share a public link.

6. Conclusion

The objective of this case study was to assess the transformative power of chatbots on the efficiency of the terminological preparation process, adequacy and accuracy of results, and their user-oriented features. However, it is essential to acknowledge the study's limitations stemming from its reliance on a single example, although this approach can identify fundamental trends. The benefits of efficiency gains

and time savings are clear, as the use of chatbots can speed up term extraction. User-friendliness is maintained as long as the tools remain free, without restrictions, and offer straightforward export capabilities. Nonetheless, a notable concern is the uncertainty in term equivalence and the occasional inability of source queries to produce results, highlighting areas for further research to enhance adequacy and accuracy of results in this promising integration of AI tools in terminology research.

References

AlZaabi, Adhari; ALAmri, Amira; Albalushi, Halima; Aljabri, Ruqaya; AAlAbdulsalam, Abdulsalam. 2023. "ChatGPT Applications in Academic Research: A Review of Benefits, Concerns, and Recommendations." *bioRxiv*. doi: https://doi.org/10.1101/2023.08.17.553688

Ahn, Sangzin. 2023. "A Use Case of ChatGPT in a Flipped Medical Terminology Course." *Korean journal of medical education* 35(3). 303–307. doi: https://doi.org/10.3946/kjme.2023.269

Bishop, Lea. 2023. "A Computer Wrote This Paper: What ChatGPT Means for Education, Research, and Writing." *SSRN Journal*. doi: https://doi.org/10.2139/ssrn.4338981

Braun, Sabine. 2019. "Technology and Interpreting." In O'Hagan, M. (Ed.) *The Routledge Handbook of Translation and Technology*. Routledge handbooks in translation and interpreting studies. London: Routledge, Taylor & Francis Group. 271–288.

Chiocchetti, Elena; Lušicky, Vesna; Wissik, Tanja. 2023. "The Role of ISO/TC 37 Standards for Translators, Interpreters, Terminologists and Beyond." *DT* 10(2). 156–179. doi: https://doi.org/10.1075/dt.00006.chi

Corpas Pastor, Gloria; Dúran-Muñoz, Isabel (Eds.). 2018. *Trends in E-Tools and Resources for Translators and Interpreters*. Approaches to translation studies, 45. Leiden, Boston: Brill Rodopi. doi: https://doi.org/10.1163/9789004351790

Engberg, Jan. 2023. "Frame Approach to Legal Terminology: What May Be Gained from Seeing Terminology as Manifestation of Legal Knowledge?" In Biel, L. & Kockaert, H. J. (Eds.) *Handbook of Terminology*. Amsterdam: John Benjamins Publishing Company. 16–36. doi: https://doi.org/10.1075/hot.3

Fantinuoli, Claudio. 2018. "Computer-Assisted Interpreting: Challenges and Future Perspectives." In Corpas Pastor, G. & Dúran-Muñoz, I. (Eds.) *Trends in E-Tools and Resources for Translators and Interpreters*. Approaches to translation studies, 45. Leiden, Boston: Brill Rodopi. 153–174.

Google. 2024. "Gemini - Chat to Supercharge Your Ideas." Accessed February 9, 2024. https://gemini.google.com/app

Halaweh, Mohanad. 2023. "ChatGPT in Education: Strategies for Responsible Implementation." *CONT ED TECHNOLOGY* 15(2): ep421. doi: https://doi.org/10.30935/cedtech/13036

Haleem, Abid; Javaid, Mohd; Singh, Ravi Pratap. 2022. "An Era of ChatGPT as a Sig-

nificant Futuristic Support Tool: A Study on Features, Abilities, and Challenges." *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 2(4). doi: https://doi.org/10.1016/j.tbench.2023.100089

Havelka, Ivana; Iacono, Katia; Pöllabauer, Sonja. 2021. "Konsekutives Ferndolmetschen: Audio- und Videodolmetschen am Beispiel des Asylwesens." In *Trainingshandbuch für DolmetscherInnen im Asylverfahren*. 2. Auflage. 210–236. Linz: Trauner Verlag + Buchservice.

Hsu, Mei-Hua. 2023. "Mastering Medical Terminology with ChatGPT and Termbot." *Health Education Journal.* doi: https://doi.org/10.1177/00178969231197371

Iacono, Katia; Havelka, Ivana; Sinclair, Katerina. 2021. "Working Paper – Digitalisierung und deren Auswirkungen auf Dolmetscherinnen und Dolmetscher. Ergebnisse der Umfrage zum Audio- und Videodolmetschen." doi: https://doi.org/10.25365/phaidra.301

Jurafsky, Daniel; and Martin, James H. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Third Edition draft. Stanfort: Stanfort University. https://web.stanford.edu/~jurafsky/slp3/

Kelly, Nataly; Pöchhacker, Franz. 2015. "Telephone Interpreting." In Pöchhacker, F., Grbić, N., Mead, P. & Setton, R. (Eds.) *Routledge Encyclopedia of Interpreting Studies*. London: Routledge. 412–415.

Kluzer, Stefano; Priego, Laia Pujol; Cabrera, Marcelino; O'Keeffe, William. 2018. "DigComp into Action, Get Inspired Make It Happen: A User Guide to the European Digital Competence Framework.". Luxembourg: Publications Office of the European Union.

KÜDES. 2018. "Empfehlungen Für Die Terminologiearbeit." Accessed February 6, 2024. https://www.bk.admin.ch/dam/bk/de/dokumente/terminologie/kuedes_empfehlungenfuerdieterminologiearbeit2003.pdf

Lušicky, Vesna. 2019. "Dolmetscher*innen als Wissens-Und Terminologiemanager*innen." In Kadric, M. (Ed.) *Besondere Berufsfelder Für Dolmetscherinnen*. Wien: facultas. 67–90.

Microsoft. 2024. "Microsoft Copilot." Accessed February 7, 2024. https://copilot.microsoft.com/

OpenAI. 2024a. "Guide to Prompt Engineering." Accessed February 8, 2024. https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions

OpenAI. 2024b. "ChatGPT." Accessed February 7, 2024. https://chat.openai.com

Passmore, Jonathan; Tee, David. 2023. "The Library of Babel: Assessing the Powers of Artificial Intelligence in Knowledge Synthesis, Learning and Development and Coaching." *Journal of Work-Applied Management.* doi: https://doi.org/10.1108/JWAM-06-2023-0057

Perplexity. 2024. Accessed February 07, 2024. https://www.perplexity.ai

Russell, Stuart; Norvig, Peter. 2016. *Artificial Intelligence: A Modern Approach.* 3rd

ed. Harlow, United Kingdom: Pearson Education, Limited. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5831883

Seeber, Kilian; Fox, Brian. 2022. "Distance Conference Interpreting." In Albl-Mikasa, M. & Tiselius, E. (Eds.) *The Routledge Handbook of Conference Interpreting*. First published 2022. Routledge handbooks in translation and interpreting studies. London, New York: Routledge. 491–507.

Silverman, David. 2024. *Interpreting Qualitative Data.* Sage publications.

Van Dis, Eva A. M.; Bollen, Johan; Zuidema, Willem; van Rooij, Robert; Bockting, Claudi L. 2023. "ChatGPT: Five Priorities for Research." *Nature* 614(7947). 224–226. doi: https://doi.org/10.1038/d41586-023-00288-7

Internet sources

KÜDES. 2018. "Empfehlungen für die Terminologiearbeit." Accessed February 6, 2024. N/A

Lušicky, Vesna. 2019. "Dolmetscher*innen Als Wissens-Und Terminologiemanager*innen." In Kadric, M. (Ed.) *Besondere Berufsfelder Für Dolmetscherinnen*. Wien: facultas. 67–90.

Microsoft. 2024. "Microsoft Copilot." Accessed February 7, 2024. https://copilot.microsoft.com

OpenAI. 2024a. "Guide to Prompt Engineering." Accessed February 8, 2024. https://bplatform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions

OpenAI. 2024b. "ChatGPT." Accessed February 7, 2024. https://chat.openai.com/

Passmore, Jonathan; Tee, David. 2023. "The Library of Babel: Assessing the Powers of Artificial Intelligence in Knowledge Synthesis, Learning and Development and Coaching." *Journal of Work-Applied Management*. doi: https://doi.org/10.1108/JWAM-06-2023-0057

Perplexity. 2024. Accessed February 7, 2024. https://www.perplexity.ai

Ramos, Fernando Prieto; Guzmán, Diego. 2023. "Measuring the Quality of Legal Terminological Decisions in Institutional Translation: A Comparative Analysis of Adequacy Patterns in Three Settings." In Biel, L. & Kockaert, H. J. (Eds.) *Handbook of Terminology*. Amsterdam: John Benjamins Publishing Company. 375–396.

Russell, Stuart; Norvig, Peter. 2016. *Artificial Intelligence: A Modern Approach*. Harlow, United Kingdom: Pearson Education, Limited. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5831883

Seeber, Kilian; Fox, Brian. 2022. "Distance Conference Interpreting." In Albl-Mikasa, M. & Tiselius, E. (Eds.) *The Routledge Handbook of Conference Interpreting*. London, New York: Routledge. 491–507.

Silverman, David. 2024. *Interpreting Qualitative Data*. Sage publications.

Van Dis, Eva A. M.; Bollen, Johan; Zuidema, Willem; van Rooij, Robert; Bockting, Claudi L. 2023. "ChatGPT: Five Priorities for Research." *Nature* 614(7947). 224–226. doi: https://doi.org/10.1038/d41586-023-00288-7

Jana Kegalj, Mirjana Borucinsky

Faculty of Maritime Studies, University of Rijeka, Croatia
jana.kegalj@pfri.uniri.hr, mirjana.borucinsky@pfri.uniri.hr

Data-Driven Learning for Writing Skills Development

Abstract

Exposing students to corpus-informed research is a typical example of data-driven learning. This paper reports on the ways that corpora (i.e., text collections), corpus tools (i.e., software packages), and corpus methods (i.e., techniques for analysing corpus data) can be used to develop students' writing skills, while enabling them to improve their digital competencies, which is in line with current trends in education. The authors present and discuss the ways that corpus-derived materials can be developed for teaching writing skills with the goal of engaging students during their learning process and enabling them to conduct their own linguistic research. The data-driven learning (DDL) method based on corpus search was implemented in a specialized course aimed for doctoral students at the University of Rijeka Faculty of Maritime Studies. The course, implemented within the UNIRI CLASS A2 *Digital Citizenship—Innovations in Learning and Teaching* in 2022 project line, aimed to use the corpus-based data-driven learning method to develop students' academic writing skills. This was intended to make students more independent and autonomous in their learning, to enhance their digital skills, and to stimulate them to be more involved in their own learning.

Keywords: data-driven learning, corpora, academic writing, language resources

1. Introduction

The rapid progress and increasing availability of language technologies have greatly influenced various areas of language production and language teaching. Language technologies comprise various software and tools that include natural language processing, lexical computing, and speech technologies. They can be broadly divided into language resources, language tools, and commercial products. The most prominent among language resources are corpora, which enable the processing of large amounts of linguistic data. The application of corpora has had a strong influence on teaching and learning of foreign languages (cf. Campoy-Cubillo et al. 2010; Hunston 2022), and now there are many researchers who suggest ways to use corpora in the classroom or to create classroom materials, dictionaries, glossaries, or other useful resources. Corpora provide multiple ways of usage to address different learning needs, but in particular they enable students to become more autonomous

in their learning, to self-direct their learning, and to gain life-long learning skills.

First of all, users of corpora, both teachers and students (Kilfarriff and Kosem 2012), have to distinguish among different language technologies offered nowadays, namely language resources, language tools, and commercial products (Tadić 2003), to be able to understand what each of them has to offer. Language resources represent linguistic material that has been digitally systematized and can be used to perform various searches. They include corpora, linguistic collections, and digital dictionaries. Language tools are specialized programs, developed on the basis of language resources, which enable the processing of existing resources or the creation of new ones, such as Sketch Engine,[1] LancsBox,[2] AntConc,[3] etc. Commercial products include dictionaries, spell checkers, grammar checkers, style checkers, machine translation tools, computer-assisted translation tools, etc., which offer linguistic checks for a fee.

Johns (1991) was one of the first to suggest that corpora can be used in the classroom as an effective way to engage students and make them active participants in the learning process, i.e., to allow them to discover language and language patterns. This was the basis for the DDL method, which uses large amounts of data as input for students to observe, analyse, interpret, explore, compare, hypothesize, and draw their own conclusions. Since then, corpora have been introduced in teacher education, translator training, teaching literature, and assessment (Flowerdew 2012). The use of corpora in teaching has been attested by various studies (cf. Kennedy and Miceli 2001; Kennedy and Miceli 2010; Cheng et al. 2003; Chambers and O'Sullivan 2004; Gaskell and Cobb 2004; Lee and Swales 2006; Boulton 2012; Chujo et al. 2012; Boulton and Cobb 2017; Vyatkina 2016, 2020). This type of teaching stimulates student motivation, develops critical thinking and lifelong learning skills. The role of the teacher changes significantly as the teacher becomes more of a guide, a mentor, an advisor. The approach of using corpora as a source of information is problem-based as students are confronted with a problem that they need to investigate. It is also a form of self-directed learning as students can take initiative, identify their problems and sources, and then apply the strategies they have acquired to solve the problem.

Having in mind all the contemporary challenges posed by the digital revolution and the need to adapt to new circumstances, the authors wanted to provide a platform for developing writing skills using available language technologies. To be more specific, at the beginning of their studies, students are are expected to write a scientific paper and later their thesis. They frequently encounter problems with structuring their papers, gaining appropriate linguistic knowledge, and modifying their style of writing. The new course, conceived as an online course, was set up to

1 https://www.sketchengine.eu/

2 http://corpora.lancs.ac.uk/lancsbox/

3 https://www.laurenceanthony.net/software/antconc/

address this problem by introducing students to corpora and the data that can be extracted from them.

2. Applications of Corpora in Teaching

Corpora provide large amounts of authentic data in a particular form of output. Although they present a useful resource in education, corpora are still not widely implemented, especially at lower proficiency levels, particularly when it comes to direct application in class (cf. Vyatkina 2020; Hunston 2022).

They can be used directly by teachers and learners in class or indirectly in the production of textbooks and teaching materials (Römer 2011). Indirect applications of corpora include their use by researchers and educational material writers as resources for designing teaching syllabi, selecting the material for teaching in class, using word lists of key vocabulary in a particular register, using corpus data to design teaching materials or create textbooks (so-called corpus-informed materials). Direct applications of corpora in teaching refer to hands-on experience with corpora where teachers and learners search corpora in order to learn about language patterns, words, or phrases "in an autonomous way" (Bernardini 2002: 165). In class, students can work indirectly with corpus results in the form of filtered and printed concordances, or they can work directly with corpora, which has become possible owing to the availability of software and resources (e.g., AntConc, Skell,[4] etc). Students' work on corpora can also be classified as guided (also called direct learning) or unguided (referred to as data-driven learning).

The most commonly used corpus functionality in classrooms are concordances, that is, lines of immediate context around a searched keyword. Johns (1991) proposed using them as a way for students to infer meanings and study functions of words and thus become aware of some typical combinations. Other data and activities that can be used in teaching include various frequency lists and lists of collocations that can provide insight into word meaning, typical grammatical patterns, stylistic preferences, conceptual framework, and semantic network. Flowerdew (2009) also mentions that corpora can provide useful information on the behaviour of words, multi-word units, grammatical patterns, semantic prosody and semantic preference, as well as pragmatic and textual features. The author also emphasized that corpora are not only research tools but also pedagogic tools, as they raise students' language awareness, engage their interest, and develop autonomy in learning.

Flowerdew (2012) also notes that despite all the advantages of using corpora in the classroom, direct applications of corpora have not been implemented that much. She attributes this to the problem of authentication, i.e., demonstrating the usefulness of corpora to students, and the problem of simplification, as unedited corpus data

4 https://skell.sketchengine.eu/#home?lang=en

might be too demanding for students. This is corroborated by Boulton (2017) who notes that DDL has an "impressive pedigree" but remains on the margins in practice.

Ädel (2010) states that little attention has been paid to the possibilities of implementing corpora in teaching writing skills, apart from teaching vocabulary and collocations. She also emphasizes that the few examples of hands-on application of corpora in teaching writing were mainly one-time experiments. Farr and Karlsen (2023) note that DDL has mostly been introduced in higher education, with academic writing and teacher education as the most prominent areas. Here we will present a selection of case studies in which DDL was implemented to develop writing skills. The courses that used corpora directly in class focused mainly either on rhetorical functions (to gain fluency and wider knowledge about a specific genre or linguistic functions) or on specific lexical or grammatical structures (for accuracy and error correction).

2.1 DDL Focusing on Lexical and Grammatical Items in Writing

Focusing on teaching vocabulary, Thurstun and Candlin (1998) developed corpus-based learning materials that would cater for the needs of students from different disciplines, following the hypothesis that students generally do not struggle with discipline-specific vocabulary but have problems with mid-frequency vocabulary (cf. Li and Pemberton 1994; Coxhead 2000), or so-called "semi-technical" or "academic vocabulary" (Nation 1990). Thurnstun and Candlin (1998) focused on a restricted set of vocabulary items and the use of concordancing techniques to expose students to the authentic use of these items. They used key words from identified rhetorical functions (e.g., stating the topic, referring to the research literature, reporting the research of others, expressing opinions tentatively, explaining procedures in the study, drawing conclusions) to focus on specific vocabulary items.

Yoon and Hirvela (2004) implemented a corpus-based approach to learning writing in an intermediate and an advanced ESL class. The corpus approach was gradually introduced into the ESL course with the aim to enrich the content and teaching methodology. They took a gradual approach to corpus use, that is, from explanation and demonstration to the independent use of primarily concordances and frequencies. The use of corpus tools was aimed at enhancing students' knowledge of vocabulary and grammar. The authors found that the level of student proficiency should be one of the key factors when considering how to incorporate corpus search into an ESL course.

Chambers and O'Sullivan (2004) introduced corpus consultation work into a master's course to assist students in error correction of their own written work. A special part of the course focused on training students in corpus consultation skills. The lecturer provided feedback on a piece of writing produced by the students, and then the students had to correct and improve their text using corpus tools. Similarly, Gilmore (2009) intergrated corpus tools in the redrafting stage of students'

writing where they were required to make their own hypotheses on how to improve their writing based on corpus data.

Ädel (2010) also reports on a small-scale experiment conducted with beginner-level students aimed to improve their writing skills. The author used concordancing tools to target specific research questions, mostly in the form of guided learning. However, just like Lee and Swales (2006), Ädel's attempt did not become part of regular writing instruction.

Another attempt to introduce corpora in writing instruction was made by Kennedy and Miceli (2010). They used corpus work as an aid to writing, achieving accuracy, and solving a specific grammatical issues. As their students had an intermediate knowledge of English, they decided to apply a more guided approach to using corpora, calling it "corpus apprenticeship" and advocating for "corpus-consultancy literacy", as the knowledge gained on such a course would be potentially used in future writing tasks. The authors aimed to teach students to use concordances to enrich the content and language and to edit their text.

Flowerdew (2012) also reported on a corpus-informed course aimed at writing reports. Students were gradually introduced to corpus consultation strategies during the course and not in separate sessions. The learning process was guided, with teacher-directed tasks until the last stage when students had to work with corpora on their own. The author adopted a "guided inductive approach" to assist students in interpreting concordance results as this posed a particular challenge to them. The author reports that the search queries mainly focused on lexical and grammatical elements, but they were also encouraged to observe other (e.g., phraseological or genre-based) properties.

Bruce et al. (2016) reported on conducting a series of workshops aimed at improving academic writing skills of chemistry students using the DDL approach. The authors compiled the FOCUS corpus of academic texts produced by students and regarded as strong examples of academic work by various departments at Durham University. Their workshops focused on academic voice, reporting verbs, nominalisation, punctuation, and connectives, all of which were analysed using examples from the corpus.

2.2 DDL Focusing on Rhetorical Functions in Writing

Bernardini (2004) described an approach she calls "corpus-aided discovery learning" which relied on autonomous learning with corpora as resource. In Bernardini's unguided approach, the learner is seen as a researcher who makes hypotheses, poses questions, and finds ways to combine corpus tools to solve them. The author also distinguishes between "learning *from* corpora" and "learning *with* corpora", the latter being based on discovery learning. In the course, students worked with the instructor as their guide and explored language items and patterns in context to make inferences about their function in the text. Tribble (2004) uses a corpus

of business correspondence to show how keyword analysis can be an efficient tool in understanding linguistic patterns in professional surroundings. The keywords were then related to social functions in the texts. Lee and Swales (2006) reported on a corpus-informed writing course English for Academic Purposes (EAP) for doctoral students which relied mostly on concordancing and was focused on discourse features of academic texts. During the course, students consulted specialized corpora of academic writing and had to compile two additional corpora. Their final report had to focus on how corpus linguistics techniques raised their level of rhetorical consciousness. Cortes (2007) reported on a corpus-based genre-oriented EAP course aimed at raising students' awareness of different linguistic features typical of a genre and stimulateing them to use this knowledge in the future.

Charles (2007, 2011) also described a writing course offered at the Oxford University Language Centre based on rhetorical functions in which concordancing was used to focus on specific lexicogrammatical units in which a particular function was realized. The corpus used consisted of theses written by native speakers. In this course, the author combines the top-down discourse-based approach with the bottom-up corpus-based approach to make students aware of the key features of academic discourse. Hyland (2012) similarly emphasized the usefulness of corpora in identifying lexicogrammatical regularities in academic texts.

Flowerdew (2015) designed a workshop for science and engineering students in which corpus tools were used to identify typical move structures in the discussion section of a thesis. The students were familiarized with search strategies and then they identified useful lexical and grammatical items for specific rhetorical functions.

Aull (2017) conducted a course using context-informed corpus analysis for first-year college students with the aim of identifying the key structures used in the macro-genres of argumentative and explanatory writing. Chen and Flowerdew (2018) held a series of DDL workshops for PhD students focusing on discourse features of EAP and the language structures used to perform specific functions in a scientific text. Similarly, Dong and Lu (2020) reported on a corpus-based and genre-based EAP course set out to teach rhetorical structures in a discipline-specific academic writing course. Similarly, Wong (2019) reports on postgraduate writing courses which use a corpus-based multidisciplinary thesis writing support resource developed for this purpose.

All of these examples (which are by no means exhaustive) demonstrate how prolific the area of DDL implementation in teaching writing actually is. The researchers also report an increase in their students' proficiency, autonomy in learning, and greater motivation, which was also the goal of the course described here.

2.3. Methodology of Corpus-Based Teaching of Writing Skills

In the course designed as part of the UNIRI CLASS A2 *Digital Citizenship—Innovations in Learning and Teaching* project, the students work directly with corpora. As

different studies on the application of corpora in class have shown (cf. Thurnston and Candlin 1998; Gaskell and Cobb 2004; Kennedy and Miceli 2010), learners benefit more when the course progresses from more indirect, guided tasks towards direct, unguided tasks, which was adopted in this course as well, with the aim of reducing the students' dependence on the instructor's help. In this way, it is assumed that the students will gain basic knowledge on corpora and then have more control of their learning and become more independent. The goal was to demonstrate the usefulness of corpus tools in everyday writing tasks of PhD students, stimulate them to further explore different linguistic variations, promote autonomy in online surroundings, and use corpus tools to improve their own previously written works. Besides that, such learning ultimately promotes "tailor-made" learning, as it does not determine exactly what should be learnt, thus stimulating what Bernardini calls "serendipitous learning".

The design and topics included in the course were based on corpus studies of academic writing, which has significantly contributed to our understanding of this register. Specifically, corpus research has singled out a range of lexical and grammatical features typical for academic discourse, as well as described rhetorical functions used in different genres. As the course is intended for PhD students from different fields, the examples were taken from academic texts from different areas, while the focus was on general features of academic texts and so-called academic vocabulary which is common to all disciplines. The structure of the course adhered to Charles's (2011) proposed three-stage process: first, corpus awareness to introduce students to corpora and the data that can be extracted from them; second, corpus literacy to make the students perform simple searches, to understand concordance data, and to be able to make their own queries; third, corpus proficiency to make the students build their own (however small) corpus and perform queries in that corpus.

In the first, guided part of the course, the students work on a ready-made corpus of academic writing compiled in the AntConc program (Anthony 2023).[5] The students are introduced to the AntConc program through demonstration of its features and opportunities to use the program directly. After that, they are assigned various guided tasks with frequency lists, concordances, collocation lists from other programs (e.g., Sketch Engine, LancsBox) to familiarize them with other applications and different display options that these programs offer. In the second, unguided part of the course, the students are encouraged to compile a corpus by themselves, collecting texts from their area of interest. After compiling the corpus, they are tasked with writing a journal article abstract, which is divided into several steps. Thurnston and Candlin's chain of activities (1998: 272), namely "look at concordances", "familiarize yourself with the patterns", "practice key terms", and "create your own writing", were modified here into more research-directed activities:

5 The corpus used was AmE06_Learned, consisting of 80 files, a total of 161469 tokens.

analyse your data, deduce general rules, produce your work based on previous findings. The tasks were designed to combine lexical and grammatical features with discourse functions they realize in the text.

3. A Sample of Practical DDL-Based Tasks Employing Corpora

One module of the course focuses on writing journal abstracts, which is a challenging form, as it needs to provide the reader with a short summary of the research topic, methodology, and results. The goal was to enable students to write an abstract for their own written article by analysing the lexical and grammatical items used for a particular purpose in order to raise their awareness of the rhetorical functions found in these short texts. The students worked on the corpus of abstracts that they had compiled. The course was not designed to be comprehensive and all-inclusive, but the primary objective was to develop the students' DDL skills so that they themselves can continue using this approach for their future learning needs.

The module starts with a series of guided tasks to stimulate the students to use corpus tools to distinguish between words "paper", "article", "work", "study", and "research". In order to solve the tasks, the students need to use different functionalities of corpus processing tools (e.g., concordances, frequency analysis, list of collocates, n-grams). Some research (cf. Thurnstun and Candlin 1998) has shown that over-exposure to concordance lines might be tiring to students. Hence, a variety of tasks was designed ranging from closed exercises such as the "one-item multiple contexts" exercise (Johns 2000) shown in Figure 1 to open tasks such as the one shown in Figure 2.

Tactics, an article discussing the political direction Russia should take after the revolution.</s><s>In this	he advocated a 'revolutionary democratic dictatorship of the proletariat and peasantry', a term that Marx
sconceptualisations.</s><s>The final chapter will be devoted to a direct critique of Cruickshank's recent	A tale of two ontologies: an immanent critique of critical realism (The Sociological Review (54), 2004).</
ıction<s>Canaan Banana, the former president of Zimbabwe, is quoted by Michael Edwards in his 1989	entitled, "The irrelevance of development studies" as follows;</s>"Whereas an armchair intellectual of ru
utledge, London/New York, 1999) Wyatt, Mark, White Knuckle Ride.(Salamander Books, London, 1996)	Bennett, Peter, Blackpool Pleasure Beach, Burk, John, Tour of Blackpool, Auge, M, Orienting, Chapter 1
as a scientific method of observation.</s><s>James presented his theories by writing books and journal	, teaching them as a lecturer, and giving a series of 'Talks to Teachers'.</s><s>He is known for trying to
crux of his argument.</s><s>The Levellers set out their franchise proposals to be debated at Putney in	l from An Agreement of the People, the first Agreement, which stated:</s>'That the people of England, t
such beliefs on the movement.</s><s>On the latter point Walwyn responded unequivocally, referring to	XXX of the final Agreement:</s><s>Morton, The World of the Ranters, p. 183</s>'That it shall not be in
ı jedem Reglement' - here we see a slight indication that, although the press prevents the publication of	against the state's wishes, it cannot stop the writing or thinking of such ideas and this point offers a sligh
ct those rights - a philosophy which went against what the Ancien Régime represented.</s><s>The first	of the Déclaration states that: 'Les hommes naissent et demeurent libres et égaux en droits."</s><s>an
quired shift in attitude was to take place.</s><s>This can be seen in the language used for the eleventh	of the Déclaration; whereas others are a list of rights, this is almost a call to reason, hinting that without
mplete' and that it merely 'took the form of a short, concise preamble to the constitution'.</s><s>Indeed,	16 in the document itself refers to a constitution directly hinting that this document does not have the po
on des droits de l'homme was, in some ways, a machine de guerre contre l'Ancien Régime.</s><s>The	themselves attempt to destroy all that was unjust in French Society at that time concerning: national sov
·Cambridge Law Journal,</s><s>vol. 60, no.</s><s>2</s><s>52 Human Rights Act 1998 s.6</s><s>53	2, Protocol I, European Convention on Human Rights</s><s>54 De Freitas v Permanent Secretary, Min
·e, R (S) v Chief Constable of South Yorkshire, this declaration of incompatibility is sought with regard to	64(1A) of the Police and Criminal Evidence Act 1984 (PACE) which allows Police to retain DNA samples
open question'.</s><s>However, retention of photographs and fingerprints were seen not to contravene	8(1) of ECHR in Kinnunen v.</s><s>Finland in relation to fraud.</s><s>Therefore, according to Lord Ste
ıent between Appellants and people in an 'identical situation'.</s><s>Lord Steyn then underscores that	14 is 'not limitless' and going beyond the prescribed boundaries is not its goal. If this is discrimination, A
ɔf private information by State Institutions" as an interference with Article 8(1) of ECHR. She fears that if	8(1) is not engaged in this case then neither is Article 14 and therefore the State becomes free to retain
ılf of the model is also erroneous and is in need of further refinement.</s><s>There is no doubt that this	is the most controversial of the four, taking what Brown said out of context and too the extreme (column
› of the Main Developmental Theories in Your Discussion Review A Recent 'Child Development' Journal	and Discuss How the Results of This Article Support or Challenge This Developmental Theory<s>The d
OF LEGISLATION:Article 226 EC (ex. Article 169) Treaty of Amsterdam (EC Treaty) Article 227 EC (ex.	170) Treaty of Amsterdam (EC Treaty)Article 228 EC (ex. Article 171) Treaty of Amsterdam (EC Treaty)»

Figure 1. Example of a one-item multiple contexts exercise.

In this particular task, the students are asked to write down the different meanings of the word "article" that can be extracted from the concordance lines (the

concordance lines were taken from Sketch Engine). Different tasks target various problem-solving techniques by the student.

Tactics, an article discussing the political direction Russia should take after the revolution.</s><s>In this	article	he advocated a 'revolutionary democratic dictatorship of the proletariat and peasantry', a term that Marx
sconceptualisations.</s><s>The final chapter will be devoted to a direct critique of Cruickshank's recent	article	A tale of two ontologies: an immanent critique of critical realism (The Sociological Review (54), 2004).</
ıction<s>Canaan Banana, the former president of Zimbabwe, is quoted by Michael Edwards in his 1989	article	entitled, "The irrelevance of development studies" as follows;</s>"Whereas an armchair intellectual of ru
utledge, London/New York, 1999) Wyatt, Mark, White Knuckle Ride.(Salamander Books, London, 1996)	Articles	Bennett, Peter, Blackpool Pleasure Beach, Burk, John, Tour of Blackpool, Auge, M, Orienting, Chapter 1
as a scientific method of observation.</s><s>James presented his theories by writing books and journal	articles	, teaching them as a lecturer, and giving a series of 'Talks to Teachers'.</s><s>He is known for trying to
crux of his argument.</s><s>The Levellers set out their franchise proposals to be debated at Putney in	Article	I from An Agreement of the People, the first Agreement, which stated:</s>'That the people of England, t
such beliefs on the movement.</s><s>On the latter point Walwyn responded unequivocally, referring to	Article	XXX of the final Agreement:</s><s>Morton, The World of the Ranters, p. 183</s>'That it shall not be in
ı jedem Reglement' - here we see a slight indication that, although the press prevents the publication of	articles	against the state's wishes, it cannot stop the writing or thinking of such ideas and this point offers a sligh
ct those rights - a philosophy which went against what the Ancien Régime represented.</s><s>The first	article	of the Déclaration states that: 'Les hommes naissent et demeurent libres et égaux en droits."</s><s>an
quired shift in attitude was to take place.</s><s>This can be seen in the language used for the eleventh	article	of the Déclaration; whereas others are a list of rights, this is almost a call to reason, hinting that without
mplete' and that it merely 'took the form of a short, concise preamble to the constitution'.</s><s>Indeed,	article	16 in the document itself refers to a constitution directly hinting that this document does not have the po
on des droits de l'homme was, in some ways, a machine de guerre contre l'Ancien Régime.</s><s>The	articles	themselves attempt to destroy all that was unjust in French Society at that time concerning: national sov
-Cambridge Law Journal,</s><s>vol. 60, no.</s><s>2</s><s>52 Human Rights Act 1998 s.6</s><s>53	Article	2, Protocol I, European Convention on Human Rights</s><s>54 De Freitas v Permanent Secretary, Min
e, R (S) v Chief Constable of South Yorkshire, this declaration of incompatibility is sought with regard to	Article	64(1A) of the Police and Criminal Evidence Act 1984 (PACE) which allows Police to retain DNA samples
open question'.</s><s>However, retention of photographs and fingerprints were seen not to contravene	Article	8(1) of ECHR in Kinnunen v.</s><s>Finland in relation to fraud.</s><s>Therefore, according to Lord Ste
nent between Appellants and people in an 'identical situation'.</s><s>Lord Steyn then underscores that	Article	14 is 'not limitless' and going beyond the prescribed boundaries is not its goal. If this is discrimination, A
ɔf private information by State Institutions" as an interference with Article 8(1) of ECHR. She fears that if	Article	8(1) is not engaged in this case then neither is Article 14 and therefore the State becomes free to retain
alf of the model is also erroneous and is in need of further refinement.</s><s>There is no doubt that this	article	is the most controversial of the four, taking what Brown said out of context and too the extreme (column
ɜ of the Main Developmental Theories in Your Discussion Review A Recent 'Child Development' Journal	Article	and Discuss How the Results of This Article Support or Challenge This Developmental Theory<s>The d
OF LEGISLATION:Article 226 EC (ex. Article 169) Treaty of Amsterdam (EC Treaty) Article 227 EC (ex.	Article	170) Treaty of Amsterdam (EC Treaty)Article 228 EC (ex. Article 171) Treaty of Amsterdam (EC Treaty),

Figure 2. Example of an open task based on concordance lines.

The series of tasks in the following example focuses the students' attention on different features of the given words, aiming to clarify their use. However, the deeper goal is to demonstrate to the students through their own hands-on experience with DDL how they can solve a different problem in their future work.

Example 1. A sample of tasks related to the use of the following words: 'paper', 'article', 'work', 'study', 'research'.

1. Open AntConc and upload the corpus AmE06_j_learned, which is a corpus of academic texts (see the AntConc tutorial if you do not know how to do that). Type in the word 'paper' and 'article' and analyse their concordances, plots, collocations, and clusters. What can you see?
What is the frequency of 'paper' vs the frequency of 'article'?

2. Take a closer look at concordance lines and write down some of the main findings regarding the use of 'paper' and the use of 'article' in context. What are the similarities/differences?

3. What is the most frequent word to the left/right of the words 'paper' and 'article'?

4. Now type the word 'work' in AntConc in the same corpus (AmE06_J_learned). Which meaning does the word 'work' NOT refer to?
   a) labour
   b) research
   c) article
   d) effort

5. Study vs research
These two nouns seem similar, but they actually refer to different things.
Take a look at this list of collocates.
What do you think—are they collocates of the word 'study' or the word 'research'?

6. Try searching for the word 'study' and the word 'research' in AntConc and find out which prepositions follow them.
Which is the most frequent preposition after the word 'research'?

7. Now take a closer look at the word 'research' in the corpus. Which verbs follow this noun?

8. Write a sentence using the word 'research' and one of the verbs you have found in the corpus. The sentence should refer to your own current research, the research that you are currently reading about, or the research that you have read about.

One of the steps of the module focuses on general features of abstracts. The students analyse the length of the abstracts and arrangement of rhetorical moves (e.g., background, present research, methods/materials, results, conclusion, recommendations, etc.) in the corpus that they have compiled on their own. This is followed by a series of tasks linking lexical and grammatical structures to these moves, e.g., they have to analyse the tenses used in abstracts and then link the use of tenses with particular rhetorical moves. Another example of a corpus-based task is the analysis of the use of "I" and "we" and their frequencies in their corpus, as well as of instances of self-referring (e.g., "this article") to make the students aware of this feature so they can use it in their writing. It is important to emphasize that the students have previously been familiarized with the corpus query software and have performed basic searches.

Example 2. A sample of tasks aimed at studying the general features of abstracts using a tailor-made corpus.

1. Take a look at the abstracts that you have collected in the previous step.
Can you calculate how many words they have? How many sentences do they have? What is the average number of words per sentence?

2. Take a look at your abstracts again. Which tense is mostly used?

3. Now that you have found the tense that is mostly used in your abstracts—can you say why do you think this tense is used more than other tenses? Write your opinion.

4. Using your corpus of abstracts in AntConc, find out about the use of pronouns 'I' and 'we' in the corpus. Which tool would be best to find that out?
Which pronoun is used more, if any? Is perhaps the use of expressions like 'present authors' used instead?

5. Now take a look at another feature of abstracts, which is self-referring. This implies that there is reference to the article/study/paper itself in the text, e.g., 'this article'. Are there such references in your corpus? Which tool would you use to find this out?

6. Go back to File view in AntConc and read a few of the abstracts that you have collected. Can you find a pattern? Is there a sequence of information that the authors are providing? What do they present first: purpose, background, methods, some results, or findings?
Write the structure of your abstracts here.

7. Let's now take a closer look at the opening sentences in the abstracts. Go to AntConc again and click on File view.
Go through the first sentences of your abstracts. What do they state: purpose? some standard practice? present action of the researchers? problem?
Write down 2-3 most common functions of opening sentences in your corpus.

8. Now go back to the same abstracts. Focus on the second sentence in each abstract. How is it linked with the previous one?
Research suggests that there are several ways to do this:

1. keeping the same subject,
2. putting the information from the second half of the first sentence in the subject position at the beginning of the second,
3. using a new previously unmentioned topic.

Which one of these was most common in your corpus of abstracts?

9. Researchers have established that it is quite common to introduce a new topic in the second sentence because articles have a limited number of words, they are highly compressed texts, and the authors expect the readers to have the relevant knowledge, so they do not go into further detail.
Study your corpus in AntConc. Take a look at n-gram results and single out those that would be used to connect two sentences. Write them down here and specify their meaning (contrast, addition, cause-effect, sequence, conclusion, purpose, etc).

At the end of this module, the students are expected to produce an abstract of their own. The abstract is then peer-reviewed by other participants in the course and by the instructor as well. The goal is for the students to produce a text with the assistance of DDL, i.e., corpus tools, and thus gain life-long skills they can use in any future writing project.

4. Conclusion

In this case study, corpus data were used as input for guided writing tasks and as the basis for productive tasks. The case study also took into account that students need to be trained in corpus research before they can engage in autonomous

corpus-informed learning. The process of material design highlighted some possible pitfalls and challenges of such an approach, mainly related to the compilation of a suitable corpus, the selection of appropriate material, the time required for students to master the skills and query the corpus, and the method of training learners to use corpora. The main goal of the course is to provide students with useful new skills that can be applied in any kind of writing task.

The course aims to use two types of corpora: ready-made corpora of academic texts at the beginning of the course and tailor-made corpora compiled by students at the more advanced level of the course. Ready-made corpora of academic texts are used to engage students in studying the frequent expressions, structures, and style of academic texts. After gaining some knowledge about corpora and how they can be studied, students compile their own corpus on a topic of their interest. The process of upgrading their writing skills is research-based and task-based using the data from the corpus. Such teaching poses many challenges. First, students do not possess the meta linguistic knowledge about corpora or linguistics so this aspect can be time-consuming and students might get discouraged by such an approach. Second, the availability of corpus analysis tools and ready-made specialized corpora is restricted (cf. Anthony 2019). Third, there are few resources and materials developed for corpus-based teaching of writing. This paper aims to address these challenges and bridge the said gaps by demonstrating the kind of corpus-based tasks that can be designed for learners.

The future endeavours in this sense will be aimed at determining how students have benefited from this approach, how they perceived this way of learning, and what advantages they found in such learning. Also, as a follow-up after they finish the course, a further goal is to establish whether they continue using these tools in their academic writing.

References

Ädel, Annelie. 2010. "Using Corpora to Teach Academic Writing: Challenges for the Direct Approach." In Campoy, M. C., Bellés-Fortuno, B. & Gea-Valor, M. L.. (Eds.) *Corpus-Based Approaches to English Language Teaching*. London and New York: Continuum International Publishing Group. 39–55.

Anthony, Lawrence. 2023. AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, Laurence. 2019. "Tools and Strategies for Data-Driven Learning (DDL) in the EAP Writing Classroom." In Hyland K. & Wong, L. L. C. (Eds.) *Specialised English: New Directions in ESP and EAP Research and Practice*. London and New York: Routledge. 179-194.

Aull, Laura. 2017. "Corpus analysis of argumentative versus explanatory discourse in writing task genres." *Journal of Writing Analytics* 1. 1–45.

Bernardini, Silvia. 2002. "Exploring new directions for discovery learning." In

Kettemann, B. & Brill, G. (Eds.) *Teaching and learning by doing corpus analysis. Proceedings of the fourth international conference on teaching and language corpora*. 165–182.

Bernardini, Silvia. 2004. "Corpora in the classroom: An overview and some reflections on future developments." In Sinclair J. (Ed.) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins. 15–36.

Boulton, Alex. 2009. "Corpora for all? Learning styles and data-driven learning." In Mahlberg, M., González-Díaz, V. & Smith C. (Eds.) *Proceedings of 5th Corpus Linguistics Conference.*

Boulton, Alex. 2012. "Hands-on/hands-off: Alternative approaches to data-driven learning." In Thomas, J. & Boulton, A. (Eds.) *Input, process, and product: Developments in teaching and language corpora*. Masaryk University Press. 152–168.

Boulton, Alex; Cobb, Tom. 2017. "Corpus use in language learning: A meta-analysis." *Language Learning* 67(2). 348–393. doi: 10.1111/lang.12224

Bruce, Megan Laura; Coffer, Philippa; Rees, Simon; Robson, Jacquie. 2016. "Write on the edge: using a chemistry corpus to develop academic writing skills resources for undergraduate chemists." *Chemistry Education Research and Practice* 17. 580–589. doi: 10.1039/c6rp00005c

Campoy-Cubillo, Mari Carmen; Bellés-Fortuño, Begoña; Gea-Valor, Maria-Lluïsa. 2010. "Introduction to Corpus Linguistics and ELT." In Campoy-Cubillo, M. C., Bellés-Fortuño, B. & Gea-Valor, M.-L. (Eds.) *Corpus-Based Approaches to English Language Teaching*. London and New York: Continuum International Publishing Group. 3–17.

Chambers, Angela; O'Sullivan, Íde. 2004. "Corpus consultation and advanced learners' writing skills in French." *ReCALL* 16(1). 158–172.

Charles, Maggie. 2007. "Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions." *Journal of English for Academic Purposes* 6. 289–302.

Charles, Maggie. 2011. "Using hands-on concordancing to teach rhetorical functions: evaluation and implications for EAP writing classes." In Frankenberg-Garcia, A., Flowerdew, L. & Aston, G. (Eds.) *New Trends in Corpora and Language Learning*. London: Continuum. 26–43.

Chen, Meilin; Flowerdew, John. 2018. "Introducing data-driven learning to PhD students for research writing purposes: A territory wide project." *English for Specific Purposes* 50. 97–112.

Cheng, Winnie; Warren, Martin; Xun-feng, Xu. 2003. "The language learner as language researcher: putting corpus linguistics on the timetable." *System* 31. 173–186.

Chujo, Kiyomi; Laurence, Anthony; Oghigian, Kathryn; Uchibori, Asako. 2012. "Paper-based, computer-based, and combined data-driven learning using a web-based concordance." *Language Education in Asia* 3(2). 132–145.

Cortes, Viviana. 2007. "Exploring genre and corpora in the English for academic

writing class." *The ORTESOL Journal* 25. 8–14.

Coxhead, Averil. 2000. "The Academic Word List: A Corpus-based Word List for Academic Purposes." In Kettemann, B. and Marko, G. (Eds.) *Proceedings of the Fourth International Conference on Teaching and Language Corpora*.

Cresswell, Andy. 2004. "Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course in Corpora in the Foreign Language Classroom." In Hidalgo, E., Quereda, L. & Santana, J. (Eds.) *Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*. Amsterdam and New York: Rodopi. 267–288.

Farr, Fiona; Hagen Karlsen, Petter. 2023. "DDL Pedagogy, Participants, and Perspectives." In Jablonkai, R. R. & Csomay, E. (Eds.) *The Routledge Handbook of Corpora and English Language Teaching and Learning*. London and New York: Taylor & Francis. 329–343.

Flowerdew, John. 2009. "Corpora in Language Teaching." In Long, M. H. & Doughty, C. J. (Eds.) *The handbook of language teaching*. Wiley-Blackwell. 327–350.

Flowerdew, Lynne. 2012. "Corpora in the Classroom: An Applied Linguistic Perspective." In Hyland, K., Huat, C. M. & Handford, M. (Eds.) *Corpus Applications in Applied Linguistics*. Continuum International Publishing Group: London and New York. 208–224.

Flowerdew, Lynne. 2015. "Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis." *Journal of English for Academic Purposes* 20. 58-68. doi: http://dx.doi.org/10.1016/j.jeap.2015.06.001

Gaskell, Delian; Cobb, Tom. 2004. "Can learners use concordance feedback for writing errors?". *System* 32(3). 301–19.

Gilmore, Alex. 2009. "Using online corpora to develop students' writing skills." *ELT Journal* 63(4). 363–372.

Hyland, Ken. 2012. "Corpora and Academic Discourse." In Hyland, K., Huat, C. M. & Handford, M. (Eds.) *Corpus Applications in Applied Linguistics*. Continuum International Publishing Group: London and New York. 30–46.

Hunston, Susan. 2022. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johns, Tim. 1991. "Should you be persuaded: Two samples of data-driven learning materials." In Johns, T. & King, P. (Eds.) *Classroom concordancing: ELR Journal*, 4.Birmingham: Centre for English Language Studies, University of Birmingham. 1–16.

Johns, Tim. 2000. "Data-driven Learning: The Perpetual Challenge." In Kettemann, B. & Marko, G. (Eds.) *Proceedings of the Fourth International Conference on Teaching and Language Corpora*. Amsterdam and New York: Rodopi. 107–118.

Kennedy, Claire; Miceli, Tiziana. 2001. "An evaluation of intermediate students' approaches to corpus investigation." *Language Learning & Technology* 5(3). 77–90.

Kennedy, Claire; Miceli, Tiziana. 2010. "Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource." *Language*

Learning & Technology 14(1). 28–44.

Kilgarriff, Adam; Kosem, Iztok. 2012. "Corpus tools for lexicographers." In Granger, S. & Paquot, M. (Eds.) *Electronic Lexicography*. 31–56. doi: https://doi.org/10.1093/acprof:oso/9780199654864.003.0003

Lee, David; Swales, John. 2006. "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora." *English for Specific Purposes* 25(1). 56–75.

Li Siu-Leung, Edward; Pemberton, Richard. 1994. "An investigation of students' knowledge of academic and subtechnical vocabulary." In Flowerdew, L. & Tong, A. (Eds.) *Entering text*. Hong Kong: Language Centre, Hong Kong University of Science and Technology. 183–196.

Nation, Paul. 1990. *Teaching and learning vocabulary*. New York: Newbury House.

Römer, Ute. 2008. "Corpora and language teaching." In Lüdeling, A. and Kytö, M. (Eds.) *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter. 112–130.

Römer, Ute. 2011. "Corpus research applications in second language teaching." *Annual review of applied linguistics* 31. 205–225.

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Zagreb: Exlibris.

Todd, Richard Watson. 2001. "Induction from self-selected concordances and self-correction." *System* 29(1). 91–102.

Thurstun, Jennifer; Candlin, Christopher N. 1998. "Concordancing and the teaching of the vocabulary of academic English." *English for Specific Purposes* 17(3). 267–280.

Tribble, Christopher. 2004. "Managing relationships in professional writing." In Hidalgo, E., Quereda, L. & Santana, J. (Eds.) *Corpora in the Foreign Language Classroom Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*. Amsterdam and New York: Rodopi. 289–308.

Vyatkina, Nina. 2016. "Data-driven learning of collocations: Learner performance, proficiency, and perceptions." *Language Learning & Technology* 20(3). 159–179.

Vyatkina, Nina. 2020. "Corpus-informed pedagogy in a language course: Design, implementation, and evaluation." In Kruk, M. & Peterson, M. (Eds.) *New technological applications for foreign and second language learning and teaching*. Hershey, PA: IGI Global. 306–335.

Wong, Lillian L. C. 2019. "Implementing disciplinary data-driven learning for Postgraduate thesis writing." In Hyland, K. and Wong, L. L. C. (Eds.) *Specialised English: New Directions in ESP and EAP Research and Practice*. London and New York: Routledge. 195-213.

Yoon, Hyunsook and Hirvela, Alan. 2004. "ESL student attitudes toward corpus use in L2 writing." *Journal of Second Language Writing* 13. 257–283.

Frane Malenica

Department of English Studies, University of Zadar, Croatia
fmalenica@unizd.hr

Picking up the Scraps—Analyzing Video Game Reviews Using Web-Scraping Tools

Abstract

The methods for creating corpora from websites have been in use for almost two decades (Baroni and Ueyama 2006; Baroni et al. 2009), and numerous tools for extracting textual data and metadata from websites have been developed since either as standalone programs, browser extensions, or as packages and libraries in programming languages such as Python and R (cf. Bradley and James 2019; Diouf et al. 2019; Kumar and Roy 2023). The widespread availability of these tools has allowed scholars to create custom corpora on a wide array of very specific topics, such as song lyrics (Kreyer and Mukherjee 2009; Werner 2012; Motschenbacher 2016), comics (Dunst et al. 2017; Unser-Schutz 2011), video games (Heritage 2020), and video game reviews (Guzsvinecz 2022; Arik 2022; HaCohen Kerner et. al. 2020). Previous research in this domain, conducted by Cho et al. (2020), has also demonstrated the effectiveness of NLP methods in extracting and identifying the main themes of video games. In this paper, I will present the results of research conducted on a corpus of video game reviews collected from the GameSpot website (www.gamespot.com) using the rvest package (Wickham 2021) for web scraping in R, and analysed using a combination of traditional corpus linguistic (CL) methods and Natural Language Processing (NLP) methods available in the quanteda package (Benoit et al. 2018). The main aims of this paper are to: i) identify words and phrases typical for different genre of video game reviews; ii) test the applicability of web scraping and NLP methods for linguistic research. While frequency-based analysis is good for a cursory glance at words and phrases typical for this register, the keyword analysis offers more useful results. The results of the sentiment analysis show statistically significant correlation between polarity and ratings, further highlighting the usefulness of these methods.

Keywords: corpus collection, keyword analysis, n-grams, sentiment analysis, specialized corpora, video game reviews, web scraping

1. Introduction

The increased popularity of online discourse and communication and the availability of digital linguistic materials in the past several decades have paved the way for the creation of corpora based on texts collected from the internet. The first corpora collected under this web-as-corpus paradigm (first envisaged in Kilgarriff 2001)

were based on some of the most studied languages in the world, such as English, German, Italian, and Japanese (Baroni and Ueyama 2006; Baroni et al. 2009), but the same methodology was soon applied to a whole variety of other languages and registers (Brezina 2018: 18; Biber and Reppen 2015: 37), thus confirming Kilgarriff's (2001: 345) prophetic claim: "The corpus of the new millennium is the web". While certain aspects of usage of linguistic data collected online have been questioned in terms of several criteria, such as authenticity, representativeness, size, topics, etc. (cf. Gatto 2011, 2014), the pervasiveness of digital communication has made them an indispensable source of information two decades after the initial idea.

This is reflected in the number of various tools for extracting textual data and metadata from websites that have been developed since as standalone programs, browser extensions, or as packages and libraries in programming languages such as Python and R (cf. Bradley and James 2019; Diouf et al. 2019; Suchomel 2020; Kumar and Roy 2023). Diouf et al. (2019) provide a comprehensive overview of different approaches, tools, and areas of application for web scraping, and their list of ready-made tools, such as browser extensions, different software packages and platforms, and libraries in programming languages includes more than 20 different tools that were available at the time the article was published.[1] Bradley and James (2019) provide a detailed tutorial for using the *rvest* package in R to extract and store data from webpages and entire websites, supplementing their tutorial with example R scripts made available via the Open Science Framework system, while Kumar and Roy (2023) describe a similar technique using the Python programming language.

This paper provides a brief demonstration of contemporary methods for creating and analysing corpora using specialized libraries in programming languages like R and their usefulness for analysing specialized registers, such as video games and video game reviews. The main two aims of the research component of this paper are to see whether the methods in question can be used to identify the words and phrases typical for the genre in question and whether the NLP methods, such as sentiment analysis, can be used productively in linguistic research. The paper is outlined as follows—in Section 2, I provide a brief overview of previous research on custom corpora that utilized tools similar to those described in the paper; in Section 3, I discuss the importance of video games and video game reviews in recent linguistic research; in Section 4, I describe the research questions and the methodology of the paper; in Section 5, I present the results of the research, and in Section 6, I provide conclusions, an outlook concerning the potential use of the methods in question for future studies.

1 The list of available tools has probably increased since, but the number of tools available at that point in time should certainly suffice for the purposes of linguistic research.

2. Studies on Custom Corpora

The widespread availability of tools for extracting textual data and converting them into corpora has coincided with the tendency of scholars to create custom corpora on a wide array of very specific topics, such as song lyrics (Kreyer and Mukherjee 2009; Werner 2012; Motschenbacher 2016), literary texts (Fischer-Starcke 2009; Moustafa 2022), football commentaries (Merullo et al. 2022), comics (Dunst et al. 2017; Unser-Schutz 2011), and video games (Heritage 2020).

In his study of American and British pop songs, Werner (2012) analyses the lexico-grammatical and morphosyntactic features of commercially successful songs, and the diachronic development of pop song lyrics. The corpus for this study was collected using the Songtext website and annotated using the CLAWS tagger. While relatively small in size (less than 400,000 tokens in total), Werner's corpus reveals some interesting patterns of pop song lyrics, such as low lexical density, high number of contractions, and first and second person pronouns, as well as some different tendencies between the two subcorpora, e.g., the UK corpus leaning towards a more "Americanized" flavour. On a somewhat more traditional note, Fischer-Starcke (2009) analyses the features of Jane Austen's Pride and Prejudice (P&P), by looking at the keywords and the most frequent phrases via the Wordsmith tool. By comparing the P&P corpus with the two reference corpora (the corpus containing Jane Austen's work minus P&P, and the corpus containing 30 novels published from 1740 to 1859 by different authors), she identifies the keywords patterns for every pairwise comparison. By comparing the P&P corpus to the rest of Jane Austen's works, she identifies five main patterns—family and family relationships, women, men, personal pronouns, military, while the keyword comparison with other novels of the same period reveals six keyword patterns—mental concepts and emotions, women, love, courtship and marriage, family and family relationships, communication, men (ibid: 498). According to Fischer-Starcke, the use of corpus linguistic methods allows the research to uncover the patterns which cannot be perceived intuitively and as such provides a useful tool for supplementing traditional literary critical analyses.

The usefulness of computational tools for revealing the hidden patterns in texts is even more prominent in research dealing with more contemporary genres of texts. Merullo et al. (2019) use the *spaCy* library in Python to analyse the data from the corpus of American football commentaries from over 1400 games and over 6 decades and investigate potential racial biases in player description. While their research, according to their own admission, is faced with statistical and linguistic confounds (e.g., different racial distribution across different player positions), it does illustrate some indicative patterns in sports commentaries, such as increased first-name reference for non-white players and different positive terms used to describe white and non-white players in different positions (Merullo et al. 2019: 6358-6359). Even more importantly, it highlights the capability of corpora and NLP methods to identify various biases in language use. In a similar vein, Heritage

(2020) uses the corpus tools WordSmith and AntConc to investigate representation of gender in video games. Specifically, he uses the keyword analysis to look at the most representative words of the video game corpus and looks at the collocations of masculine and feminine pronouns (he and she) to see whether they co-occur with different sets of lexemes. His results indicate that the female characters tend to be described differently in the sense of their physical capabilities being less pronounced and more emphasis being placed on their mental abilities, which is not the case with the male characters. A similar corpus based on multimodal sources is presented by Dunst and his associates who present "the first digital corpus of graphic novels, memoirs, and non-fiction written in English" (Dunst et al. 2017: 15). Their corpus is annotated using the Graphic Narrative Markup Language (GNML) to include the interrelations between textual and visual information in the corpus and is supplemented with eye-tracker data of several readers for the first chapter of each graphic narrative. The data collected using the eye-tracker show a degree of consistency in terms of attention that the readers give to a certain part of the page, such as text in captions, faces, hands, and objects relevant for the story. According to Dunst et al. (2017: 19), the tools for the creation of this corpus can be of use to multiple disciplines in the humanities and social sciences, such as linguistics, media and literary studies, and psychology.[2]

What this brief and certainly non-exhaustive overview of custom corpora-based research shows is that the advances in technology for collecting, annotating, and analysing texts has allowed us to explore such diverse genres and modalities of language use that the limits seem virtually non-existent. Thus, the main concern for every researcher should not be whether there are any tools to do the job, but to find the right domain in which to use those tools.

3. Video Games as Sources of Linguistic Data

Parallel with the development of corpora based on digital texts described in Section 2, the past two decades have also witnessed increased interest in video games as valuable sources of linguistic data, primarily focusing on their effect on language learning (inter alia Gee 2003, 2013; deHaan 2005; Sylvén and Sundqvist 2012; Zhonggen 2018; Camacho Vásquez and Ovalle 2019). In one of the first comprehensive studies on significance of video games on culture and learning, Gee argues: "Video games are a new form of art. They will not replace books; they will sit beside them, interact with them, and change them and their role in society" (2003: 204). Although one of the first empirical studies on the correlation between language acquisition and video games showed a facilitation effect of video games (deHaan 2005), multiple drawbacks of the study did not allow for any broad-sweeping

2 A more comprehensive list of papers from this project can be found on the web page of The Hybrid Narrativity Project: https://groups.uni-paderborn.de/graphic-literature/wp/?lang=en.

generalizations to be made. As deHaan (2005: 282) himself notes, the study was limited to a single participant, it took the participant too much time to learn how to play the game, the majority of the data comes from self-report questionnaires, and the same test was used to test the acquisition before and after playing the video game. However, subsequent studies have remedied these issues by taking a larger sample. For instance, Sylvén and Sundqvist (2012) look at the data collected from 86 students aged 11-12 using a questionnaire, a language diary, and a three-part vocabulary test. Their results show correlation between the amount of time spent gaming and the results of the vocabulary test. A similar result is reported by Chen and Yang (2013), whose two studies show that playing adventure games has a positive effect on EFL students learning new L2 vocabulary items and that language learners have a positive opinion on the benefits of video games in language learning, despite some potential issues with the games themselves.[3] The significance of video games for language acquisition and learning in general is perhaps best reflected in the establishment of the Game-Based Learning paradigm (Burmester et al. 2006; Santos 2017; Kasemap 2017).

3.1. Studies on Video game Reviews

Beyond the scholarly interest in video games from the perspective of language acquisition and gender representation (Heritage 2020, 2022a, 2022b), video game reviews and similar texts have also become a valuable object of inquiry in the past several years. For instance, Guzsvinecz (2022) looked at the corpus of 993,932 Steam reviews of 21 games belonging to the so-called "Souls-like" genre[4] collected using the *steam_reviews* library in Python[5] and found a slight-to-moderate positive correlation between time spent playing and positive reviews and identified some of the most liked aspects of video games mentioned in positive reviews (e.g., medieval setting, drawn graphics, 2D graphics) and some of the least liked aspects (pixel graphics and futuristic setting). Using a similar methodology with the *PRAW* (Python Reddit API Wrapper) library, Arik (2022) created a corpus of comments from the social media platform Reddit and conducted a sentiment analysis of the comments, along with frequency analysis of the words used. Meanwhile, HaCohen Kerner et. al. (2020) carried out a sentiment analysis of Steam comments in Brazilian Portuguese using the *Steam API* (Application Programming Interface). In both cases, the combination of textual data scraped from social networks (Reddit) or videogame distribution service (Steam) and sentiment analysis enables the identification of the most positive aspects of video games, such as graphics, gameplay, soundtrack, and storytelling, and the most negative aspects, such as online

3 A more comprehensive overview of similar studies is provided by Yudintseva (2015).

4 Video games resembling the Souls video game franchise.

5 A similar package is also available for R (Fox et al. 2023).

gaming-related issues, DLCs, and various bugs, as reported by users themselves (HaCohen Kerner et. al. 2020: 401).

In addition to sentiment analysis, topic modelling is another potentially productive area of applying NLP methods to corpora based on video game reviews. Wang and Goh (2020) use text analysis methods to automatically generate topics from online user reviews and compare their effect on user satisfaction, identifying narrative and achievement as having the strongest correlation with satisfaction. Cho et al. (2020) compare the feasibility of qualitative human-based analysis of video game reviews and automated text mining analysis for identifying topics in video game reviews. While the human annotations showcase better understanding of the plot and narrative, the results provided by the machine-based methods show they can also be successfully used for identifying the main topics of games, especially when dealing with large databases.

4. Towards the Present Study—Research Questions, Methodology

As can be seen from the brief overview of previous research in Section 3, video games and video game reviews represent an interesting and a highly relevant object of linguistic inquiry. Thus, the main goal of this paper is to showcase how the contemporary methods for creating and analysing corpora using the specialized libraries in programming languages like R can be applied to this fast-evolving register. In order to build on the previous research and extend the analysis to other types of video game reviews, the main aims of this paper are: i) to identify words and phrases typical for the gaming genre and the individual subgenres; and ii) to test the applicability of web scraping and NLP methods for linguistic research. In order to achieve these aims, the paper will address the following research questions:

1) What words and phrases are typical for different genres of video game reviews?
2) Are typical words and phrases more accurately captured by frequency lists or keyword analysis?
3) Is there a correlation between the value obtained by Sentiment Analysis of video game reviews and the ratings provided by writers of reviews?

For this purpose, a corpus of video game reviews was collected from the GameSpot website (www.gamespot.com) using the *rvest* package (Wickham 2021) for web scraping in R.[6] The corpus was then analysed using the combination of traditional corpus linguistic (CL) methods, such as frequency lists, n-grams, and keywords and NLP methods, such as sentiment analysis, available in the *quanteda*

6 The method used in the paper was based on the tutorial on web scraping in R by John Little from Duke University, while a similar methodology is also described in Bradley and James (2019).

package (Benoit et al. 2018). Specifically, frequency lists for individual words and n-grams and keyword analysis were used to address RQ1 and RQ2, while the Augmented General Inquirer Positiv and Negativ dictionary in the quanteda.sentiment package was used to answer RQ3.

5. Results and Analysis

The data collected for the purpose of this paper were gathered from July 19 to 26, 2022 and include a total of 5243 reviews written by 213 different authors, distributed across three video game genre—Adventure, First Person Shooter (FPS), and Strategy. The oldest reviews were from May 1996 for games such as Star Trek: The Next Generation—A Final Unity and Crusader: No Remorse, while the most recent review was from July 2022 for the game Stray. For every review, five variables were automatically scraped from the web (title, text, author, date, rating), and two variables were manually added/derived (genre, sentiment). After the relevant data had been scraped from the web, they were cleaned to remove all the parts of the texts that were not part of the review (e.g., references to other games, galleries, warnings like "You need a javascript-enabled browser", etc.) and turned into a corpus using the corpus() function in the quanteda package. The corpus with the starting size of 6.68 million words was then tokenized, and all punctuation symbols and stopwords (function words like 'and', 'or', 'but', as well as the word 'game'[7]) were deleted from it, resulting in the final corpus size of 3.63 million words.

The most frequent words and n-grams from the resulting corpus were then extracted using the *dfm*[7] function, the results of which are shown in Table 1 and Table 2 below. As we can see from Table 1, raw frequency count is not particularly helpful in identifying the main topics of video game reviews at first glance, as a vast majority of the top 20 most frequent single words are gaming-nonspecific words (e.g., can, one, also, just) with sporadic lexical items relevant for the subcorpus in question (e.g., units and battle for the Strategy subcorpus). While this observation is particularly true for single words in Table 1, the 2-grams in Table 2 offer a bit more substance in terms of the identified review topics (e.g., sound effects, artificial intelligence, single-player campaign, turn-based strategy for Strategy, frame rate, team deathmatch, level design, multiplayer mode, capture flag for FPS), along with the names of certain video games (e.g., command conquer, rainbow six, tomb raider, metal gear). For the sake of brevity, only frequency lists for single words and 2-grams are provided here. The lists for 3-grams is provided in the Figure 3 in the Appendix, while an illustration of frequency lists for 4-grams can be seen in Figure 1.

7 Document feature matrix.

Figure 1. The most frequent 4-grams in the entire corpus

Table 1. The most frequent words in the corpus and the three subcorpora

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
can	39848	can	13394	can	8694	can	17760
one	20820	one	5521	one	4855	one	10444
like	19612	units	5419	like	4809	like	9564
also	15845	like	5239	also	3895	just	8035
just	15254	new	4568	just	3578	time	7997
time	15208	also	4566	even	3481	even	7423
even	15050	even	4146	get	3297	also	7384
get	13812	time	3999	time	3212	get	6985
new	13408	games	3692	new	3194	much	6329
much	12491	just	3641	though	2966	though	6085
though	11775	get	3530	enemies	2859	story	5834
make	11130	strategy	3463	play	2793	way	5818

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
games	10991	much	3409	much	2753	make	5731
way	10660	play	3303	weapons	2698	new	5646
play	10361	make	2948	multiplayer	2641	game's	5490
game's	10216	well	2757	make	2451	Characters	5235
well	10009	though	2724	way	2450	around	5232
good	9580	battle	2700	good	2433	games	5187
enemies	9129	good	2549	well	2405	enemies	4914
around	9113	enemy	2533	game's	2334	well	4847

Table 2. The most frequent 2-grams in the corpus and the three subcorpora

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
can also	1867	real-time strategy	1158	first-person shooter	775	resident evil	866
feel like	1426	strategy game	1070	call duty	580	can also	797
sound effects	1333	strategy games	692	can also	483	feel like	762
real-time strategy	1226	can also	586	single-player campaign	441	voice acting	730
voice acting	1159	sound effects	373	feel like	404	adventure game	728
frame rate	1110	original game	339	sound effects	350	frame rate	644
feels like	1056	game can	331	first-person shooters	338	sound effects	610
even though	1049	can get	304	rainbow six	329	feels like	599
can get	945	artificial intelligence	291	frame rate	327	tomb raider	554
can use	933	even though	278	world war	312	action game	535
first-person shooter	886	can take	278	team deathmatch	296	even though	534
resident evil	876	world war	270	pc version	296	playstation 2	519
takes place	803	feel like	260	xbox 360	289	splinter cell	505
xbox 360	797	can play	259	far cry	282	can use	497
single-player campaign	758	can use	248	feels like	281	metal gear	462
playstation 2	758	units can	248	2 s	276	star wars	457

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
pretty much	753	single-player campaign	234	level design	270	can get	433
can take	749	command conquer	232	multiplayer mode	251	xbox 360	431
pc version	743	turn-based strategy	226	capture flag	240	action adventure	431
strategy games	724	pretty much	214	artificial intelligence	238	first time	425

The next step in the analysis is to look at the keywords for every subcorpus and compare its successfulness in identifying relevant topics to those obtained by the frequency-based analysis. As one of the aims of the paper is to examine the usefulness of the different methods for identifying topics specific for a particular genre, the method which allows us to identify more terms specific for the genre (or subgenre) in question is in principle the one better suited for the task. Specifically, this means identifying terms which refer to various elements of the game like graphics, units, items, sounds, gameplay, etc.

The typical procedure for conducting the keyword analysis includes comparing a particular corpus (or subcorpus) of interest with a larger and more general reference corpus (Brezina 2018: 80). However, as the results in Fischer-Starcke (2009) indicate (cf. Section 2), it is also feasible to conduct this analysis by comparing the focus subcorpus against the whole corpus (i.e., conduct the whole analysis by using a single corpus). This option is utilized in the *quanteda_textstats* package, which uses the log-likelihood ratio as a measure of keyness. As we can see from Figures 2-4, the keyword analysis yields significantly better results as it identifies far more key terms (blue bars) specific for all three subcorpora. This is particularly true for the Strategy subcorpus where the top 10 keywords are the key concepts in the games in question (e.g., units, strategy, build), followed by names of various videogame franchises (e.g., worms, tycoon,[8] etc.). The same is true, although to lesser extent, for the remaining two subcorpora where the top 10 keywords also include genre-specific topics, such as shooter, deathmatch, and weapons for the FPS subcorpus, and puzzles, story, character for the Adventure subcorpus.[9]

8 The first one refers to the Worms game series (*Worms Battlegrounds, Worms Armageddon, Worms 3D*), the second to the Tycoon series (*RollerCoaster Tycoon, Zoo Tycoon, Railroad Tycoon*).

9 A more detailed list of keywords is available in Table 5 in the Appendix.

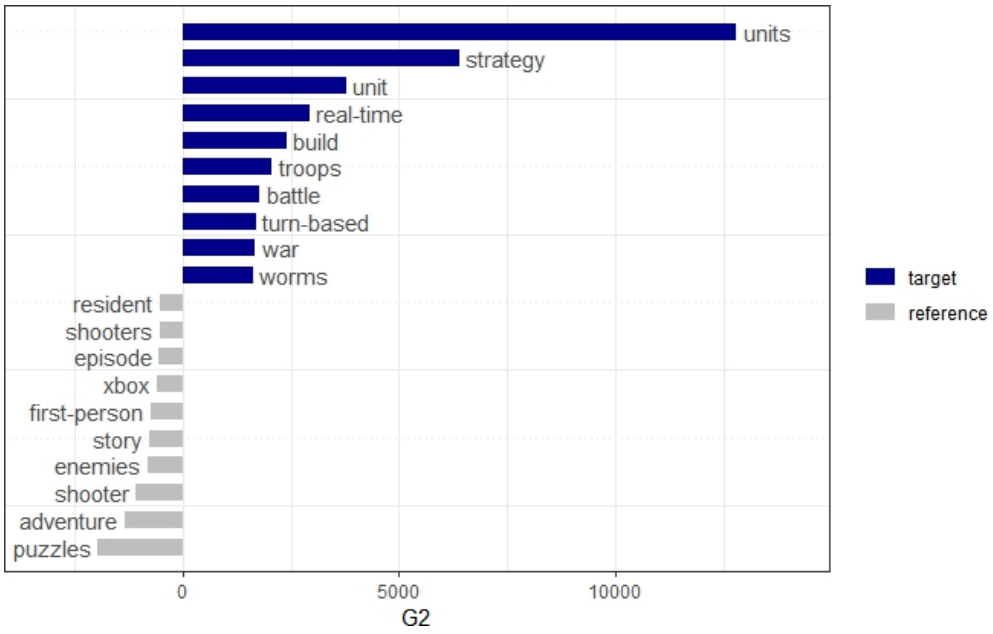

Figure 2. Keywords for the Strategy subcorpus

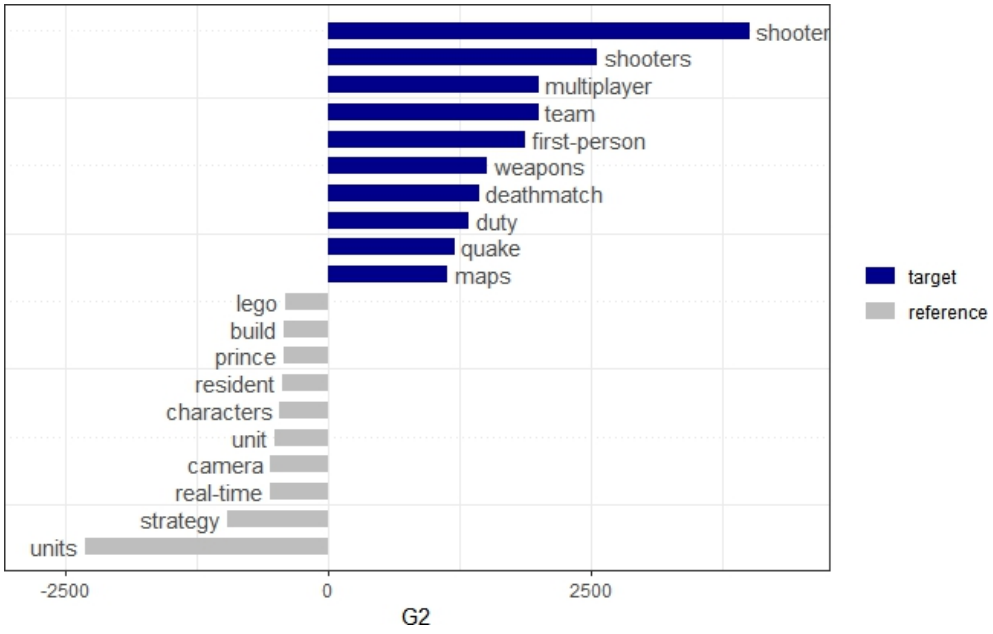

Figure 3. Keywords for the FPS subcorpus

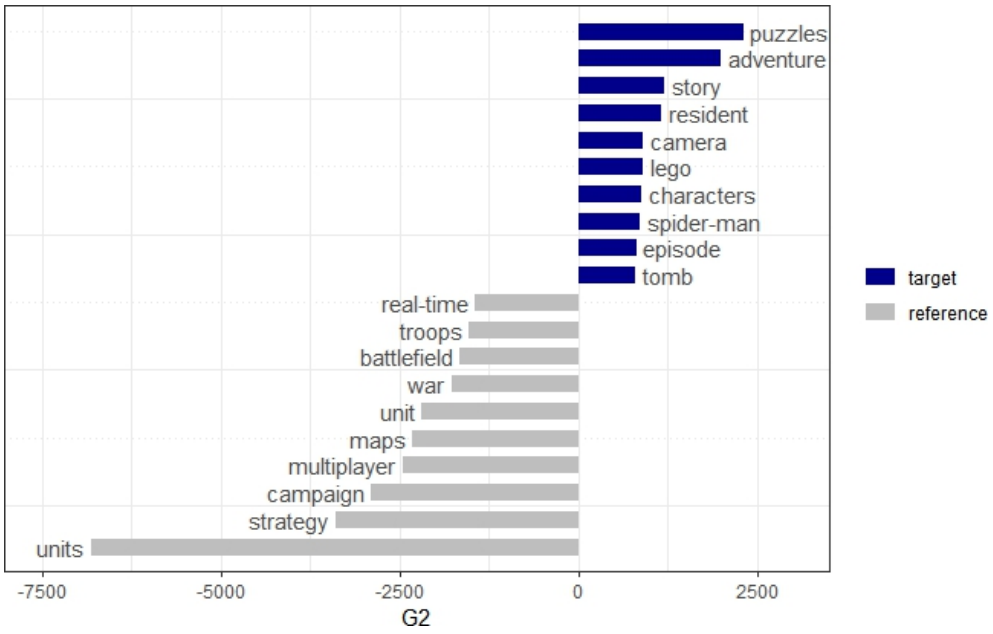

Figure 4. Keywords for the Adventure subcorpus

The last analysis, conducted to answer RQ3, is Sentiment Analysis (SA) for which the Augmented General Inquirer Positiv and Negativ dictionary in the *quanteda.sentiment* package was utilized. This SA tool is based on the presence of terms with either positive (1653 potential lexemes) or negative polarity (2010 lexemes) and assigns either a positive or a negative value. Polarity of the reviews in the corpus ranged from 1.946 (highest) to -0.765 (lowest). The obtained polarity values and ratings scraped from the website were averaged across author (N=213) and were tested for potential correlation. Pearson's correlation revealed a statistically significant (albeit relatively weak) positive correlation ($r= 0.23$, CI: 0.098 – 0.353, $p<.001$), as can be seen in Figure 5.[10] This shows that sentiment analysis can be used as a reliable tool for automatic identification and quantification of positive or negative polarity of a particular text.

10 An even stronger correlation was obtained using the *Lexicoder Sentiment Dictionary (2015)* in the same package, but is not reported here for the sake of brevity.

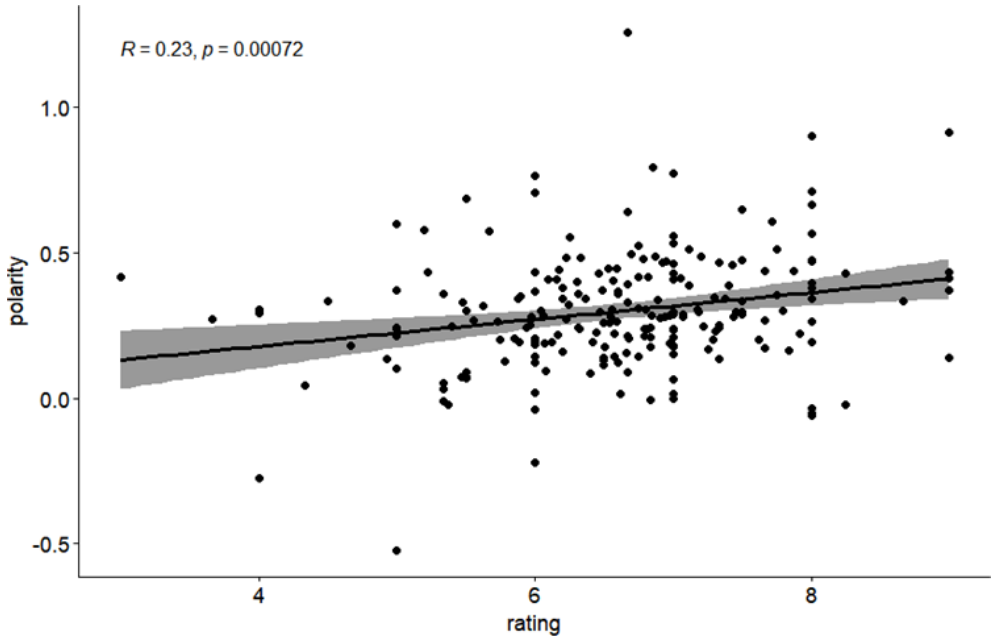

Figure 5. Correlation between polarity and ratings of the reviews

6. Conclusion

The main two aims of this paper were identifying words and phrases typical for the register of video game reviews and testing the applicability of web-scraping and NLP methods for linguistic research. As we can see from the results in Section 5, the conclusions are mixed—the most basic lists of the most frequent words and 2-grams provide only a handful of genre-specific units, while including a lot of noise in terms of general-purpose words (e.g., can, one, like in Table 1). That being said, the keyword analysis does much better since it provides a more substantial list of genre-specific lexical items, as we can see in Figures 2–4 and Table 5. Undoubtedly, this method is not without flaws, as there are again some items that do not serve the purpose of identifying the main aspects of video games, such as names of individual characters and videogame franchises. However, the benefits of such a method still outnumber the downsides shown here, which makes it quite a useful tool for identification of key topics in a particular text or register. Obviously, comparable if not even better results for identification of key topics could also be obtained by the topic modelling methodology using the *stm* package in R, but these were not employed here for the sake of brevity. Sentiment analysis, the last method used in this paper, proved to be a valuable tool as it relatively consistently recognized whether the review leaned more towards the positive or the negative end of the spectrum. The practical application of the sentiment analysis tools in linguistic research is manifold—for instance, they could be used to detect whether newspapers or politicians of different political orientation depict different social groups or minorities in more

positive or negative terms, or to examine whether there is a difference in positivity/negativity between descriptions of the same event by different individuals, that is, whether different groups perceive the same event in the same manner.

One might argue that the first two methods of the analysis used in this paper (frequency lists and keyword analysis) are available in most (if not all) software solutions for corpus management, such as WordSmith or Sketch Engine. However, what those solutions do not offer is the seamless integration of the module for collecting data (like the *rvest* package used here), the module for the basic data analysis, and the module for the more advanced NLP methods, which are all conveniently available in R. Another potential drawback of the methods used in the paper is the fact that the corpus was not lemmatized or POS-annotated. For instance, one can notice that the list of keywords for the Strategy and the FPS subcorpora both include singular and plural forms (unit and units, shooter and shooters, respectively). Integrating lemmatization and POS-tagging into the analysis using packages like *udpipe* or *spacyr* would have remedied these issues, but the idea behind this paper was to see how far one can get by using the most easily available tools, which is why this avenue was not pursued.

Ultimately, as Kilgarriff's (2001) claim about the importance of the internet for corpus linguistics remains uncontested, it is up to the linguists to embrace the new technological and analytical capabilities of the emerging tools to harness the power of the data available in online texts. Hopefully, this paper will motivate future use of tools such as those described here to incorporate new types of data into linguistic research.

References

Arik, Kaan. 2022. "Social Media Content Review of MMORPG Games: Reddit Comment Scraping and Sentiment Analysis". *Journal of Emerging Computer Technologies* 2(1). 13–21.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano; Zanchetta, Eros. 2009. "The WaCky wide web: A collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43(3). 209–226. doi: https://doi.org/10.1007/s10579-009-9081-4

Baroni, Marco; Ueyama, Motoko. 2006. "Building general- and special-purpose corpora by Web crawling". *Proceedings of the NIJL International Workshop on Language Corpora*. Tokyo: NIJL.

Benoit, Kenneth; Watanabe, Kohei; Wang, Haiyan; Nulty, Paul; Obeng, Adam; Müller, Stefan; Matsuo, Akitaka. 2018. "quanteda: An R package for the quantitative analysis of textual data". *Journal of Open Source Software* 3(30). 774. doi: https://doi.org/10.21105/joss.00774

Biber, Douglas; Reppen, Randi. 2015. *The Cambridge handbook of English corpus linguistics*. Cambridge University Press.

Bradley, Alex; James, Richard J. E. 2019. "Web Scraping Using R". *Advances in Methods and Practices in Psychological Science* 2(3). 264–270. doi: https://doi.org/10.1177/2515245919859535

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press. doi: https://doi.org/10.1017/9781316410899

HaCohen Kerner, Yaakov; Miller, Daniel; Yigal, Yair. 2020. "The influence of preprocessing on text classification using a bag-of-words representation". *SBC { Proceedings of SBGames*. https://dx.plos.org/10.1371/journal.pone.0232525

Burmester, Michael; Gerhard, Daniela; Thissen, Frank (Eds.). 2006. *Digital game-based learning: Proceedings of the 4th International Symposium for Information Design*. Stuttgart Media University, Universitätsverlag.

Camacho Vásquez, Gonzalo; Ovalle, Joan Camillo. 2019. "The Influence of Video Games on Vocabulary Acquisition in a Group of Students from the BA in English Teaching". *Gist Education and Learning Research Journal* 19. 172–192.

Chen, Howard; Yang, Ting-Yu Christine. 2013. "The impact of adventure video games on foreign language learning and the perceptions of learners". *Interactive Learning Environments* 21(2). 129–141. doi: https://doi.org/10.1080/10494820.2012.705851

DeHaan, Jonathan W. 2005. "Acquisition of Japanese as a Foreign Language Through a Baseball Video Game". *Foreign Language Annals* 38(2). 278–282. doi: https://doi.org/10.1111/j.1944-9720.2005.tb02492.x

Diouf, Rabiyatou; Sarr, Edouard Ngor; Sall, Ousmane; Birregah, Babiga; Bousso, Mamadou; Mbaye, Sény Ndiaye. 2019. "Web Scraping: State-of-the-Art and Areas of Application". *IEEE International Conference on Big Data (Big Data)*. 6040–6042. doi: https://doi.org/10.1109/BigData47090.2019.9005594

Dunst, Alexander; Hartel, Rita; Laubrock, Jochen. 2017. "The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities". *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 15–20. doi: https://doi.org/10.1109/ICDAR.2017.286

Fischer-Starcke, Bettina. 2009. "Keywords and frequent phrases of Jane Austen's Pride and Prejudice: A corpus-stylistic analysis". *International Journal of Corpus Linguistics* 14(4). 492–523. doi: https://doi.org/10.1075/ijcl.14.4.03fis

Fox, Nathan; Van Berkel, Derek; Verge, Ramiro Serrano; Lindquist, Mark. 2023. "vGameReviews: An R package for harnessing video game reviews for scientific research". *SoftwareX* 23, 101423. doi: https://doi.org/10.1016/j.softx.2023.101423

Gatto, Maristella. 2011. "The 'body' and the 'web': The web as corpus ten years on". *ICAME Journal* 35. 35–88.

Gatto, Maristella. 2014. *The web as corpus: Theory and practice*. Bloomsbury Publishing.

Gee, James Paul. 2004. *What video games have to teach us about learning and literacy*. Palgrave Macmillan.

Gee, James Paul. 2013. *Good video games and good learning: Collected essays on video*

games, learning and literacy. Peter Lang.

Heritage, Frazer. 2020. "Applying corpus linguistics to videogame data: Exploring the representation of gender in videogames at a lexical level". *Game Studies* 20(3).

Heritage, Frazer. 2022a. "Magical women: Representations of female characters in the Witcher video game series". *Discourse, Context & Media* 49, 100627. doi: https://doi.org/10.1016/j.dcm.2022.100627

Heritage, Frazer. 2022b. "Politics, pronouns and the players: Examining how videogame players react to the inclusion of a transgender character in World of Warcraft". *Gender and Language* 16(1). doi: https://doi.org/10.1558/genl.20250

Kasemap, K. 2017. "The Fundamentals of Game-Based Learning". In Kidd, T. & Morris, L. R. Jr. (Eds.) *Handbook of Research on Instructional Systems and Educational Technology*. IGI Global. 174–185. doi: https://doi.org/10.4018/978-1-5225-2399-4

Kilgarriff, Adam. (2001). *Web as corpus*. In Rayson, P., Wilson, A., McEnery, T., Hardie, A. & Khoja, S. (Eds.) *Proceedings of the Corpus Linguistics*. 342–344.

Kreyer, Rolf; Mukherjee, Joybrato. 2007. "The Style of Pop Song Lyrics: A Corpus-linguistic Pilot Study". *Anglia - Zeitschrift Für Englische Philologie* 125(1). doi: https://doi.org/10.1515/ANGL.2007.31

Kumar, Sumit; Roy, Uponika Barman. 2023. "A technique of data collection". In Goswami, T. & Sinha, G. R. (Eds.) *Statistical Modelling in Machine Learning*. Elsevier. 23–36. doi: https://doi.org/10.1016/B978-0-323-91776-6.00011-7

Merullo, Jack; Yeh, Luke; Handler, Abram; Grissom Ii, Alvin; O'Connor, Brendan; Iyyer, Mohit. 2019. "Investigating Sports Commentator Bias within a Large Corpus of American Football Broadcasts". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP). 6354–6360. doi: https://doi.org/10.18653/v1/D19-1666

Motschenbacher, Heiko. 2016. "A corpus linguistic study of the situatedness of English pop song lyrics". *Corpora* 11(1). 1–28. doi: https://doi.org/10.3366/cor.2016.0083

Moustafa, Basant S. M. (2022). "A comparative corpus stylistic analysis of thematization and characterization in Gordimer's My Son's Story and Coetzee's Disgrace". *Open Linguistics* 8(1). 46–64. doi: https://doi.org/10.1515/opli-2020-0183

Santos, Antonio. 2017. "Instructional Strategies for Game-Based Learning". In Kidd, T. & Morris, L. R. Jr. (Eds.) *Handbook of Research on Instructional Systems and Educational Technology*. IGI Global. 164–173. doi: https://doi.org/10.4018/978-1-5225-2399-4

Suchomel, Vít. 2020. *Better Web Corpora for Corpus Linguistics And NLP*. Doctoral Thesis. Masaryk University. Brno.

Sylvén, Liss Kerstin; Sundqvist, Pia. 2012. "Gaming as extramural English L2 learning and L2 proficiency among young learners". *ReCALL* 24(3). 302–321. doi: https://doi.org/10.1017/S095834401200016X

Unser-Schutz, Gianclarla. 2011. "Developing a text-based corpus of the language of Japanese comics (manga)". In Newman, J., Baayen, H. & Rice, S. (Eds.) *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. BRILL. 213–238. doi: https://doi.org/10.1163/9789401206884

Wang, Xiaohui; Goh, Dion Hoe Lian. 2020. "Components of game experience: An automatic text analysis of online reviews". *Entertainment Computing* 33. 100338. doi: https://doi.org/10.1016/j.entcom.2019.100338

Werner, Valentin. (2012). "Love is all around: A corpus-based study of pop lyrics". *Corpora* 7(1). 19–50. doi: https://doi.org/10.3366/cor.2012.0016

Wickham, Hadley. 2021. *rvest: Easily Harvest (Scrape) Web Pages*. R Package Version 1.0.2.

Yudintseva, Anastassiya. 2015. "Game-Enhanced Second Language Vocabulary Acquisition Strategies: A Systematic Review". *Open Journal of Social Sciences* 3(10). 101–109. doi: https://doi.org/10.4236/jss.2015.310015

Appendix

Table 3. The most frequent 3-grams in the corpus and the three subcorpora

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
world war ii	403	real-time strategy game	444	world war ii	191	metal gear solid	360
metal gear solid	372	real-time strategy games	352	deathmatch team deathmatch	176	grand theft auto	274
real-time strategy games	369	world war ii	187	rainbow six 3	94	lego star wars	181
grand theft auto	299	heroes might magic	99	deathmatch capture flag	91	devil may cry	179
xbox 360 version	227	turn-based strategy game	83	far cry 3	90	action adventure game	174
deathmatch team deathmatch	202	full spectrum warrior	82	team deathmatch capture	82	resident evil 4	142
lego star wars	181	final fantasy tactics	76	left 4 dead	75	xbox 360 version	132
devil may cry	180	romance three kingdoms	72	xbox 360 version	72	gear solid 2	131
playstation 2 version	153	age empires ii	68	call duty 3	72	right analog stick	108
make feel like	147	turn-based strategy games	56	rainbow six vegas	70	blood omen 2	108
resident evil 4	143	red alert 2	54	red faction ii	66	game takes place	101
spend lot time	136	railroad tycoon ii	43	bad company 2	63	playstation 2 version	98

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
right analog stick	134	game boy advance	41	call duty 4	61	game boy advance	92
gear solid 2	131	jagged alliance 2	40	quake iii arena	59	new york city	91
new york city	123	steep learning curve	39	far cry 2	57	star wars ii	80
spend much time	119	game takes place	37	duke nukem 3d	53	spend lot time	79
point point b	114	traditional real-time strategy	37	enemy artificial intelligence	49	resident evil 2	75
can also use	110	red alert 3	37	make feel like	46	third-person action game	72
enemy artificial intelligence	110	real-time strategy genre	35	call duty 2	46	spend much time	67
blood omen 2	108	sid meier's pirates	34	big red one	43	tomb raider legend	65

Table 4. The most frequent 4-grams in the corpus and the three subcorpora

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
metal gear solid 2	131	many real-time strategy games	25	team deathmatch capture flag	73	metal gear solid 2	131
lego star wars ii	80	starfleet command volume ii	21	deathmatch team deathmatch capture	71	lego star wars ii	80
team deathmatch capture flag	76	heroes might magic iii	20	call duty black ops	39	tom clancy's splinter cell	50

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
deathmatch team deathmatch capture	74	real-time strategy game set	16	standard deathmatch team deathmatch	35	grand theft auto iii	50
xbox 360 playstation 3	72	command conquer red alert	15	battlefield bad company 2	27	xbox 360 playstation 3	45
tom clancy's splinter cell	58	final fantasy tactics advance	15	far cry 2 s	27	devil may cry 3	42
grand theft auto iii	54	robin hood defender crown	15	medal honor allied assault	26	devil may cry 4	37
goes long way toward	47	war ii real-time strategy	14	call duty 4 modern	24	prince persia sands time	36
devil may cry 3	42	3d real-time strategy games	13	duty 4 modern warfare	24	metal gear solid 3	33
call duty black ops	39	typical real-time strategy game	12	xbox 360 playstation 3	24	resident evil code veronica	31
devil may cry 4	37	world war ii real-time	12	submachine guns assault rifles	23	goes long way toward	30
standard deathmatch team deathmatch	37	traditional real-time strategy games	12	ghost recon advanced warfighter	23	tomb raider angel darkness	30
prince persia sands time	36	3d real-time strategy game	12	tom clancy's rainbow six	22	grand theft auto games	29
metal gear solid 3	33	might magic heroes vi	11	left 4 dead 2	21	grand theft auto series	28
get point point b	31	red alert 3 s	11	far cry 3 s	19	harry potter sorcerer's stone	26

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
resident evil code veronica	31	harvest moon friends mineral	11	world war ii combat	19	like metal gear solid	24
playstation 3 xbox 360	30	moon friends mineral town	11	call duty world war	19	lego harry potter years	23
long way toward making	30	traditional real-time strategy game	10	call duty modern warfare	17	60 frames per second	22
grand theft auto games	30	age empires ii age	10	modes deathmatch team deathmatch	17	grand theft auto iv	22
tomb raider angel darkness	30	heroes might magic series	10	tom clancy's ghost recon	17	xbox playstation 2 versions	22

Table 5. Top 50 keywords for all three subcorpora

Strategy				**FPS**				**Adventure**			
Keyword	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*
units	12791.63	5419	157	shooter	4014.768	2117	458	puzzles	2306.533	3489	603
strategy	6382.148	3463	486	shooters	2568.759	1154	137	adventure	1994.823	3264	633
unit	3780.281	1877	180	multiplayer	2008.116	2641	2397	story	1204.104	5834	2738
real-time	2933.624	1523	179	team	2004.8	1819	1115	resident	1168.203	946	19
build	2407.141	1769	553	first-person	1886.801	1468	725	camera	913.6884	2243	678
troops	2072.831	1215	220	weapons	1523.494	2698	3079	lego	903.9804	821	36
battle	1778.051	2700	2106	deathmatch	1441.937	795	195	characters	890.7015	5235	2678
turn-based	1688.705	759	39	duty	1346.864	624	85	spider-man	861.5379	644	4
war	1680.435	2182	1479	quake	1206.947	471	22	episode	814.7044	1046	135

Strategy				**FPS**				**Adventure**			
Keyword	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*
worms	1626.313	690	20	maps	1145.582	1501	1357	tomb	799.1362	687	22
battles	1562.245	2066	1425	modes	1049.383	1164	898	film	777.4658	1123	180
rts	1493.901	625	15	single-player	1048.418	1346	1195	prince	756.8535	747	47
resources	1405.063	1000	293	halo	1005.054	484	77	raider	690.3584	558	11
building	1333.066	1392	745	teammates	975.9379	511	108	evil	667.7697	1594	470
ships	1273.965	873	236	rifle	919.7942	579	195	batman	620.9776	525	15
map	1270.758	2114	1773	online	861.2664	1283	1290	lara	620.0901	492	8
strategic	1220.25	828	219	half-life	843.5262	320	11	solve	588.5413	909	162
armies	1201.858	612	66	sniper	843.4046	568	219	splinter	550.6774	572	43
command	1170.965	972	377	doom	789.4749	487	157	harry	535.476	453	13
tycoon	1130.417	451	4	rainbow	771.0926	379	65	boss	514.4711	1325	417
interface	1125.971	1059	495	call	768.5167	890	717	moves	506.795	1507	530
heroes	1117.458	905	337	gun	720.2128	1019	983	zelda	505.9009	388	4
empire	1081.608	750	208	guns	695.3923	880	771	movie	479.4775	1239	391
scenarios	1076.796	902	355	shooting	693.5467	1007	992	button	478.4417	1955	840
resource	1067.193	597	93	ops	682.4686	376	92	puzzle	463.5554	1124	336
campaign	1016.328	2460	2641	unreal	679.5289	305	36	platforming	456.2773	531	55
campaigns	990.7091	676	181	players	677.9108	2016	3072	sequences	451.063	1459	544
factions	935.3527	643	175	f.e.a.r	661.134	234	2	creed	447.2823	379	11
army	911.6773	880	426	weapon	647.8509	1245	1498	character	446.0561	3738	2178
civilization	909.6756	520	87	battlefield	634.895	849	781	assassin's	438.9373	408	20
empires	895.2129	353	2	assault	616.9774	635	452	stealth	427.3437	1211	411
tactical	825.9579	961	582	levels	591.976	1939	3084	horror	411.4611	668	128

Strategy				**FPS**				**Adventure**			
Keyword	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*
conquer	809.142	401	38	turok	578.3071	211	4	cell	392.9027	611	110
can	794.9819	13394	26454	crysis	568.2212	210	5	fisher	380.6659	321	9
buildings	782.1325	980	641	grenades	560.7733	499	298	potter	362.2752	284	4
skirmish	778.7961	424	60	medal	557.9524	261	37	kain	341.3627	243	0
terrain	717.5223	617	253	campaign	549.3755	1933	3168	crime	340.234	544	102
infantry	714.1497	571	208	bots	539.8228	291	67	objects	337.1887	1319	553
age	703.5445	703	356	duke	531.9984	274	55	combos	336.0768	398	43
cities	702.2771	565	208	firefights	518.7839	270	56	solving	324.3613	544	109
victory	685.059	576	228	mode	504.4413	1994	3421	ninja	316.6469	452	71
research	683.68	527	180	3	503.184	980	1189	dialogue	299.1989	1395	642
tactics	682.8009	705	372	cover	497.4774	838	923	arkham	299.0238	221	1
expansion	624.9209	655	352	shoot	489.3064	836	930	legend	293.7579	429	70
structures	624.7087	578	264	counter-strike	484.9129	176	3	sword	288.2717	708	214
wargame	623.3512	248	2	vehicles	474.9474	785	853	hitman	281.9814	259	12
historical	614.7302	434	125	grenade	470.9992	365	179	kong	279.9012	275	17
forces	601.1434	972	797	soldier	467.043	457	307	dead	273.3179	1307	609
scenario	591.8282	530	231	pc	461.3926	1227	1772	persia	272.1509	244	10
management	568.418	428	140	wolfenstein	455.2862	163	2	detective	271.3996	346	44

Ana Ostroški Anić

Institute for Croatian Language and Linguistics, Zagreb, Croatia
aostrosk@ihjj.hr

Definitional Patterns in Specialized Resources for Schoolchildren

Abstract

The paper describes the analysis of frequent definitional patterns in two specialized corpora, English and Croatian, compiled of educational texts in the field of climate change written for schoolchildren and young adults. The analysis is based on extracting knowledge-rich contexts from the corpora, followed by determining common lexical markers used to identify definitions, lexical knowledge patterns, and lexical-semantic relations between terms. Definitions and definitional patterns will be used in developing a children-oriented dictionary of climate change in Croatian.

Keywords: definitions, knowledge patterns, knowledge-rich contexts, lexical markers, specialized corpora

1. Introduction

When schoolchildren first encounter novel concepts in any specialized domain, e.g., in the field of biology or chemistry, they are often presented with a two-fold challenge: first, they need to understand the meaning of the concepts and put them into relation with previously acquired knowledge, and second, they need to memorize their definitions in order to be able to reproduce them later when needed. As expected, educational experts creating specialized resources for children must be aware of different linguistic skills children possess at different age, as well as their general cognitive development and categorization skills, which often hinder their understanding of very abstract concepts (Casadevante et al. 2019).

The issue of climate change and all related phenomena have been vastly discussed in all aspects and domains of modern life from various perspectives. The topic is therefore well known to children, as climate change topics are also part of the curricula of school subjects (e.g., science and geography), and are even approached in activities for pre-school children. However, climate change educational materials in Croatian are still very scarce, and mostly designed for teachers as additional instructions in preparing classroom activities.

This paper describes the process of identifying the most frequent definitional patterns in two specialized corpora, English and Croatian, consisting of educational

texts in the field of climate change written for schoolchildren and young adults. The method used to identify and formulate definitional patterns is corpus-driven, and it encompasses both the analysis of the lexical level of terms, the lexical markers used to identify definitions, and the lexical-semantic relations between terms, as well as the conceptual analysis applied to identify concept characteristics expressed in definitions. The motivation for the research is explained in the next section, followed by the detailed presentation of the methodology applied in the compilation of corpora, their analysis, and the annotation and analysis of knowledge patterns in the extracted definition. After the discussion of the definitional patterns in the analysed examples, a few recommendations for future research are given in the final section.

Results will be used in preparing material for experimental research investigating children's categorization development, especially in the context of developing classification schemes for children, as mentioned by Murray and Reuter (2005). In practical terms, definitions and definitional patterns will be used in developing a children-oriented dictionary of climate change in Croatian.

2. The Importance of Context in Children's Categorization

When creating educational or instructional resources for children, whether those are textbooks or leaflets and other similar short text formats used to educate children, it is important to consider the structure and format of the resource, as well as the language used to introduce and define new concepts. Using clear and age-appropriate language, e.g., formulating ideas in shorter sentences and avoiding the use of complex terms, ensures that both content and form resonates with children's language development and cognitive abilities. Murray and Reuter (2005) argue that many experimental researchers in children categorization mistake a difference in language use for a difference in ability to understand categories, whereas children can be fluent in less structured language, but not in applying adult-like categories (2005: 9). The ability to comprehend abstract concepts falls within that domain.

Research shows that children aged six to nine acquire abstract concepts to which they are more emotionally connected (Vigliocco et al. 2018). The emotional experience remains an important factor in concept acquisition even beyond the critical age, alongside the engagement of the sensorimotor system (Reggin et al. 2021). Moreover, Gelman and Meyer (2011) emphasize the contextual nature of children's categorization processes, indicating that the surrounding context strongly influences how they categorize information. Previous research has already shown that context is the dominant distinguishing element in eliciting conceptual relations by abstract and concrete concepts. Caramelli et al. (2004) report on the different development of concrete and abstract conceptual knowledge in children aged eight, ten, and twelve, i.e., on the different pattern of information elicited by concrete and

abstract concepts in schoolchildren. Their findings confirm that the conceptual information specific for abstract concepts refers to situations and events in which they are experienced, as opposed to concrete concepts that convey mostly information on the properties of the objects they refer to (2004: 31). As opposed to a wider range of conceptual relations elicited by concrete concepts, abstract concepts elicit thematic information that is contextually dependent. These results, as many other experimental findings, lead to doubts on the suitability of conventionally formulated definitions, e.g., intensional terminological definitions, in resources designed for schoolchildren.

3. Knowledge-Rich Contexts and Knowledge-Driven Approach to Definitions

The concept of knowledge-rich context is a well-known construct in terminology literature.[1] Starting from the Meyer's seminal paper in 2001, the idea of a knowledge-rich context (KRC) as a linguistic context providing information useful for conceptual analysis has seen numerous applications both in practical terminology work and in theoretical investigations. In order to detect these contexts, knowledge patterns are applied as recurrent linguistic structures "indicative of semantic relations" (Barrière 2004). Knowledge patterns usually include two elements linked by a relationship, e.g., a concept and one of its characteristics, a concept and its superordinate concept, etc. On the linguistic level, they are realized by terms or other linguistic expressions and a marker of their relationship (Marshmann 2006). Marshmann investigated lexical knowledge markers, which she defined as markers that take the "form of a lexical unit or sequence of lexical units" (2006: 2) in knowledge patterns for the cause-effect and association relations in medical texts. For our analysis, both relations are relevant as they often relate to the conceptual relations of events and activities, specific for the climate change domain.

A more recent example of using KRCs in text-mining and terminological analysis is Ramos and Costa (2023), who extracted definitional contexts as knowledge-rich contexts to establish lexical markers, lexical-semantic relations between terms, and interpret the term relations, which was followed by a conceptual analysis. A similar procedure of extracting definitional information from specialized corpora of unstructured texts (Fišer et al. 2010) revealed uncertainty in distinguishing between definitions and non-definitions, especially in texts where new concepts are introduced and described in many different manners (2010: 2934). The same was expected to be a distinguishing feature of semi-specialized texts prepared for children, which formed the corpora designed for this study.

1 For an overview of the use of knowledge-rich contexts in terminology and knowledge engineering, see Condamines (2022).

When Croatian terminological literature is concerned, Grčić Simeunović and Vintar presented a list of sixteen lexico-syntactic knowledge markers to extract definition candidates from the corpus on karst (2015: 257–258). The markers are roughly grouped into three categories: those that represent typical intensional definitions (e.g., [noun in Nominative] + [verb *be*] + [noun in Nominative], markers comprising structures used to introduce new concepts, including the verbs *represent*, *define*, and finally, markers that are specific for the analysed domain, e.g., [noun] + *process*, *consists of* + [noun in Genitive]. Grčić Simeunović and Vintar (2015) applied the lexico-syntactic patterns in order to identify definitions and extract knowledge-rich contexts from the corpus. The aim of this paper could be regarded as quite opposite: to start from definitional and knowledge-rich contexts to determine lexico-syntactic markers used in defining concepts of climate change and identify common knowledge patterns in definitional contexts.

4. Methodology

In order to identify the patterns of definitions in educational resources on climate change for schoolchildren, the analysis was carried out both in English and Croatian. Given the overall dominance of English materials in the domain, we started from English as the dominant language, and then applied the same methodology to Croatian. Two small, specialized corpora were initially compiled, both consisting of popular educational or instructional resources written for schoolchildren as the target audience. We understand popular educational resources to include all popular scientific books as well as glossaries, brochures or leaflets, which excludes textbooks used as the primary teaching source in regular school settings. Since the analysis described here presents the preliminary work that is part of a larger study, we focus on instructional materials only. The English corpus consists of 421,756 tokens. The lack of resources for the population of schoolchildren and young adults is evident in the size of the Croatian corpus, which consists of only 123,905 tokens.

First, lists of keywords and terms were extracted from both corpora. The list of keywords (or single unit terms) did not yield many results due to too much noise in the corpus because of the varieties of resources included, some of which are glossaries and not actual running texts. Therefore, we decided to focus on the list of terms to choose the candidates for the list of seed words for corpus querying. Additionally, as a type of validation, we manually analysed two popular educational resources in English on the topic of climate change, *Climate glossary for young people* (Cognuck González and Numer 2020) and *What's the issue? Climate change* (Jackson and Guitian 2020) to make sure that certain key concepts were not omitted from the list of candidates. The Croatian translation of Jackson and Guitian (2020) was analysed as well.

Having analysed the first one thousand term candidates from the corpus, we obtained the list of terms used for extracting definitions. We decided to omit the

terms for concepts that are not limited to the domain of climate change as well as those that are borderline with words in general language or those that, by being the core terminology of science, can be defined from several perspectives, e.g., *water*, *oxygen*, *atmosphere*, *polar ice*, etc. A list of forty-four key terms was eventually compiled on the basis of terms validated in other resources, e.g., *acid rain*, *air pollution*, *biodiversity*, *carbon footprint*, *climate*, *climate action*, *fossil fuel*, *global warming*, *renewable energy*, etc.

Definitions and knowledge-rich contexts of these key concepts have been extracted and first divided according to the type of the category of the *definiendum*, i.e., into events, activities, and entities, which were further categorized into concrete and abstract concepts. The same procedure was then applied to the Croatian corpus, starting from the Croatian terms for the firty-four English key concepts identified in the English corpus. Certain terms did not yield any definitions or knowledge-rich contexts (e.g., *klimatske akcije*, *klimatski uvjeti*, *klimatska pravda*, and *klimatska politika* in the Croatian corpus), and were therefore omitted from the final list. Being a much larger corpus, we had anticipated that the English corpus was to yield a more representative and interesting corpus examples.

Having extracted all relevant definitions and defining contexts, we analyzed the examples of verb patterns, as well as nominal and adjectival phrases in the sentences. All examples were then annotated for lexical markers, lexical knowledge patterns, and lexical-semantic relations expressed in the sentences in the same manner for English and Croatian (Appendix 1 and Appendix 2). Strategies used to define the concepts are then discussed according to the type of definition applied. A common strategy, especially in defining activities or processes, is to describe the cause or the most prominent aspect of the process, rather than rely on the intensional definition as the prototypical terminological definition. Different definitional patterns are then presented according to the types of concepts they identify, with the aim of suggesting definition templates, including various types of terminological definitions.

5. Results and Discussion

Given the register, style, and format of the texts comprising both corpora, especially the English one, it is difficult to determine whether certain extracted contexts could be classified as definitional contexts or as proper definitions. To illustrate this issue, we provide three extracted examples describing the concept of climate change:

(1) Climate change is the global climate variation of the earth.

(2) The adverse effects of climate change have resulted in impacts that people have not been able to cope with or adapt to, and that can lead to loss or damage, such as loss of biodiversity and ecosystem services, loss of income and damage to infrastructure.

(3) Climate change may be caused by natural internal processes or by external forces, such as volcanic eruptions or persistent anthropogenic actions.

Example (1) is the only example of the three that could be considered an intensional definition, modelled according to the traditional Aristotelian principle of defining a concept by referring to a superordinate concept and its delimiting characteristics. Regardless of its structure, it is still not the best example since both nouns, *change* and *variation*, are abstract nouns, and could be considered synonyms. Example (2) is an evident case of a knowledge-rich context because it contains conceptual information relevant for placing the concept in relation to other concepts in the network. However, it is not a proper definition as it elaborates on the effects of climate change, rather than defining climate change itself. Sentences like this one are classic examples of describing a concept by putting it in a broader context after it is initially identified. Example (3), on the other hand, although it does not define what climate change is, explains the causes of it and could therefore be considered a definitional context.

There is an overlap in the understanding of knowledge-rich contexts and definitional contexts. Having already defined KRCs as linguistic contexts that provide information useful for conceptual analysis, we understand *definitional context* in a narrower sense, as the "discursive context where relevant information to define a term could be found" (Sierra et al. 1998: 77). In unstructured texts such as popular scientific resources, it is a valid procedure to identify a specialized concept by a definitional context that provides the concept characteristics necessary to understand it.

In order to determine the most frequent definitional patterns used in educational resources on climate change, we first started from the English corpus. Appendices 1 and 2 list common lexical markers used to link the defined concept to other related concepts or its characteristics, among which the verb *be* is most often found in lexical knowledge patterns both in English and Croatian extracted examples. We distinguish between lexical markers and knowledge patterns in the sense that lexical markers are used as linguistic signals that introduce terms and/or link them to other terms or linguistic elements in the sentence, whereas lexical knowledge patterns are linguistic structures pointing to different conceptual or semantic relations and are therefore relevant for the conceptual analysis. Lexical markers are linguistic expressions, e.g., *is called*, *identified by*, *caused by*, or *attributed to*, while knowledge patterns are expressed as formulas, where NP stands for noun phrases expressing terms. Lexical-semantic relations expressed by the knowledge patterns are also identified for both languages, while the fourth column in each appendix contains examples of definitional contexts extracted from corpora.[2]

As is expected, the most typical lexical-semantic relation expressed between the two terms in definitional contexts in both languages is the relation between a

2 Lexical-semantic relations are named after those in Marshmann (2006), whereas the lexical knowledge patterns are defined following Grčić Simeunović and Vintar (2015).

hyperonym and a hyponym, as in the following definition of *fossil fuels*, where *energy sources* is the hyperonym of *fossil fuels*:

(4) Fossil fuels are energy sources that are generated when plant and animal matter biodegrades.

The lexical-semantic relation between a term and its superordinate term is in the English corpus lexically signalled by markers such as *also known as*, *is called*, *refer to*, and by a not so common but interesting marker *should be seen as*. In addition to *is called* and *refer to*, the Croatian equivalents of markers for the same lexical-semantic relation include *represents* and the variants *the name for* and *the term for*.

As Table 1 shows, the *cause-effect* relation is another commonly detected lexical-semantic relation in the English data, as is *synonymy* and the *association* between the two terms expressed in a sentence. Using exemplification—signalled here by the lexical marker *such as*, but often also by markers *for example* or *that is*—is a useful approach used to further describe the defined concept in intensional definitions by means of enumerations (Nilsson 2015), such as in Example (5):

(5) Climate change may be caused by natural internal processes or by external forces, such as volcanic eruptions or persistent anthropogenic actions.

When enumeration is used as the governing principle to describe the concept, as in the definition "The three key greenhouse gases in our atmosphere are methane, water vapor, and CO_2" it is no longer the case of an intensional definition.

Table 1. Lexical markers and knowledge patterns used in English definitional contexts

Lexical marker	**Lexical knowledge pattern**	**Lexical-semantic relation**
also known as	[NP] [be] also known as [NP]	synonymy hyperonymy/hyponymy
attributed to	[NP] [be] attributed to [NP]	association
is (a/the), are (a/the)	[NP] [be] ([article]) [NP] (of)	hyperonymy/hyponymy
caused by	[NP] [be] caused by [NP] [NP] [can/may] [be] caused by [NP]	cause-effect
is called, can be called	[NP] ([can]) [be] called [NP]	hyperonymy/hyponymy
due to	[NP] [be] due to [NP]	cause–effect
identified as/by, can be identified as/by	[NP] ([can]) [be] identified as/by [NP]	association
made from	[NP] [be] made from ([article]) [NP]	result
mean	[NP] means [NP]/ [VP]	cause–effect
refer to (as)	[NP] refers to [NP] (as)	hyperonymy/hyponymy
should be seen as	[NP] should be seen as [NP]	hyperonymy/hyponymy

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation
such as	[NP] ([punctuation mark]) such as [NP]	exemplification
that is	[NP] [punctuation mark] that is ([comma]) [NP]	synonymy

Although the Croatian corpus is significantly smaller than the English one, and not all knowledge-rich contexts of the seed terms could have been used for the extraction of definitional contexts, it nevertheless provided a significant number of lexical markers useful for reaching conclusions about the dominant lexical-semantic relations present on the term level. Apart from the mentioned relations of *hyperonymy* and *hyponymy*, as well as *synonymy* and the relation of *association*, the *causative* relation is interesting because of the variation in knowledge patterns, i.e., the valence patterns of the verbs used as lexical markers. Because the spatial preposition *do* 'by' in Croatian requires the Genitive case of nouns, the phrasal verbs *dovesti do* 'lead to' and *doći do* 'occur' always stand before a noun in Genitive, while the verb *uzrokovati* 'cause' asks for a noun in Nominative. Table 2 lists other markers and knowledge patterns, some of which also have variants, such as those that have the verb *be* in the copulative position.

Table 2. Lexical markers and knowledge patterns used in Croatian definitional contexts

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation
biti 'be'	[NP] [biti] [NP]	hyperonymy/hyponymy
činiti 'make'	[NP] čine [NP]	meronymy
dolazi do 'occur'	do [NPGen] dolazi zbog [NPGen] do [NPGen] dolazi jer [NPNom]	cause–effect
dovesti do 'lead to'	[NP] [dovesti] do [NPGen]	cause–effect
ime za 'name for'	[NP] [ime] za [NP]	hyperonymy/hyponymy
koristiti 'use'	[NP] [koristiti] [NPAcc]	purpose
nastati 'form'	[NP] [nastati] [NPIns]	result
nastati od 'made of'	[NP] [nastati] od [NPGen]	meronymy
naziv za 'term for'	[NP] [biti] naziv za [NP] / [NP] naziv [biti] za [NP]	hyperonymy/hyponymy
nazivati se 'called'	[NP] [nazivati] se [NP] / [NP] se [nazivati] [NP]	synonymy hyperonymy/hyponymy
naziva se još 'also called'	[NP] [nazivati] se još (i) [NP] / [NP] se još [nazivati] (i) [NP]	hyperonymy/hyponymy
odnositi se na 'refer to'	[NP] [odnositi] se na [NP]	hyperonymy/hyponymy

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation
posljedica 'consequence of'	[NP] ([biti]) [posljedica] [NPGen] / [NP] [posljedica] ([biti]) [NPGen]	result
predstavljati 'represent'	[NPAcc] [predstavljati] [NP]	hyperonymy/hyponymy
uzrokovati 'cause'	[NP] [uzrokovati] [NP]	cause-effect

The Croatian data show the use of the relations of *purpose* and *result*, which were not detected in the English examples. However, these lexical-semantic relations could be broadly grouped as general causative relations, so it could certainly not be claimed that they are not present in the English data.

The function of lexical-semantic relations at the term level corresponds to the conceptual relations between concepts at the concept level or the level of knowledge organization. Therefore, at the concept level, the relations *type_of* and *part_of* are hierarchical concept relations used to express whether a subordinate concept is a type of or part of a superordinate concept. In addition to these relations, which make up the largest number of all concept relations in extracted contexts, non-hierarchical or associative relations are used to refer to the cause-effect, purpose, result, origin, attributes, or function of the concept. In the domain of climate change, where many activities and processes are linked by causal and temporal relations, non-hierarchical concept relations can offer more knowledge information than the traditional ontological relations hierarchically organized and presented.

6. Conclusion

The analysis of the definitional contexts of key terms in the domain of climate change began with extracting knowledge-rich contexts from English and Croatian corpora of popular educational resources written for schoolchildren aged eight or nine to approximately sixteen. After the initial analysis and downsizing the data to only those examples that could be used to define specialized concepts in case no definitions are available, a list of examples of definitional contexts was prepared for each language. The lexical markers and lexical knowledge patterns marked in them point to the conclusion that similar markers are used in both languages, with the difference that the knowledge patterns in Croatian have a more fixed structure in terms of valence patterns of certain causal verbs and overall a slightly greater variation in the form of lexical markers.

Causal relations are certainly the dominant type of semantic relations, which is not surprising given the format and register of the texts, as well as the target audience. *Cause-effect*, *purpose*, and *result* relations are to be expected in educational materials, the purpose of which is not merely to identify specialized concepts

(e.g., as a database or a glossary would), but also to explain how and why these concepts occur. What is not covered by this analysis, but could be elaborated in a future study, is the function of terminological variation on the text level. Alongside this, the comprehension of abstract versus concrete concepts is certainly a line of research worth investigating, focusing on different age groups of young users and applying different methods to test the categorization and classification skills of schoolchildren and young adults.

References

Barrière, Caroline. 2004. "Knowledge-Rich Contexts Discovery." In *Proceedings of the Seventeenth Canadian Conference on Artificial Intelligence*. London, Ontario: CSCSI. 187–201.

Caramelli, Nicoletta; Setti, Annalisa; Maurizzi, Donatella D. 2004. "Concrete and abstract concepts in school age children." *Psychology of Language and Communication* 8(2). 18-34.

Casadevante, Cristina; Romero, Miriam; Fernández-Marcos, Tatiana; Hernández, José Manuel. 2019. "Category Learning in Schoolchildren. Its Relation to Age, Academic Marks and Resolution Patterns." *The Spanish Journal of Psychology* 22. 1–13.

Cognuck González, Sara; Numer, Emilia. 2020. *Climate glossary for young people*. Panama: UNICEF.

Condamines, Anne. 2022. "How the Notion of "Knowledge Rich Context" Can Be Characterized Today." *Frontiers in Communication* 7, 824711. doi: 10.3389/fcomm.2022.824711

Fišer, Darja; Pollak, Senja; Vintar, Špela. 2010. "Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10). Valletta, Malta: European Language Resources Association (ELRA).

Gelman, Susan A.; Meyer, Meredith. 2011. "Child categorization." *Wiley Interdisciplinary Review: Cognitive Science* 2 (2). 95–105.

Grčić Simeunović, Larisa; Vintar, Špela. 2015. "Modeliranje znanja: korpusna analiza definicijskih stilova." In Bratanić, M., Brač, I. & Pritchard, B. (Eds.) *Od Šuleka do Schengena*: *Terminološki, terminografski i prijevodni aspekti jezika struke*. Zagreb: Institut za hrvatski jezik i jezikoslovlje. 251–266.

Jackson, Tom; Guitian, Cristina. 2020. *What's the issue? Climate change*. Mission Viejo, CA: QED Publishing.

Jackson, Tom; Guitian, Cristina. 2021. *Klimatske promjene – O čemu je riječ*? Zagreb: Školska knjiga.

Marinellie, Sally A. 2009. "The content of children's definitions: The oral-written distinction." *Child Language Teaching and Therapy* 25(1). 89–102.

Marshmann, Elizabeth. 2006. "Lexical Knowledge Patterns for Semi-Automatic

Extraction of Cause–effect and Association Relations from Medical Texts: A Comparative Study of English and French." PhD thesis, University of Montreal.

Meyer, Ingrid. 2001. "Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework." In Bourigault, D., Jacquemin, C. & L'Homme, M-C. (Eds.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins. 279–302.

Murray, Craig G.; Reuter, Kara. 2005. "Children's acquisition of categories and the implications for research in the development of classification schemes." *ASIST* 1–13.

Nilsson, Henrik. 2015. "Enumerations count: Extensional and partitive definition." In Kockaert, H. J. & Steurs, F. (Eds.) *Handbook of Terminology. Volume 1*. Amsterdam: John Benjamins. 82–100.

Ramos, Margarida; Rute Costa, Rute. 2023. "Extracting knowledge-rich information from definitions. A corpus-based approach to building a conceptual-based terminological resource." In Di Nunzio, G., Costa, R. & Vezzani, F. (Eds.) *Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023)*. RWTH Aachen University. 1–14.

Reggin, Lorraine D.; Muraki, Emiko J.; Pexman, Penny M. 2021. "Development of Abstract Word Knowledge." *Frontiers in Psychology* 12.

Sierra, Gerardo E.; Alarcón, Rodrigo; Aguilar, Cesar Antonio; Bach, Carmen. 2008. "Definitional verbal patterns for semantic relation extraction." *Terminology* 14. 74–98.

Vigliocco, Gabriella; Ponari, Marta; Norbury, Courtenay. 2018. "Learning and Processing Abstract Words and Concepts: Insights From Typical and Atypical Development." *Top Cogn Sci* 10(3). 533–549.

Appendix 1

Table 3. Lexical markers and lexical knowledge patterns in English definitional contexts

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation	Definitional context
also known as	[NP] [be] also known as [NP]	synonymy hyperonymy/ hyponymy	Fossil fuels are also known as «nonrenewable» fuels because the amount of time required to transform living tissue into coal, oil, or natural gas is tens of millions of years at a minimum. TROPICAL RAINFOREST Also known as jungles, these forests grow wherever it is warm and wet all year round. Carbon monoxide, sulfur dioxide, ozone, methane, fluorocarbons and black carbon are also known as shortlived climate pollutants.
attributed to	[NP] [be] attributed to [NP]	association	Climate change is attributed to human activities that may alter the composition of the atmosphere.
be is (a/the), are (a/the)	[NP] [be] ([article]) [NP] (of)	hyperonymy/ hyponymy	A carbon footprint is the measure of the amount of greenhouse gases that are produced directly or indirectly from your activities. Fossil fuels are energy sources that are generated when plant and animal matter biodegrades. Greenhouse gases are the gaseous component of the atmosphere.
caused by	[NP] [be] caused by [NP] [NP] [can/may] [be] caused by [NP]	cause-effect	Industrial pollution is caused by factories and industry, such as brick kilns, manufacturing companies or power generators, releasing pollutants into the air. Climate change may be caused by natural internal processes or by external forces, such as volcanic eruptions or persistent anthropogenic actions.
called, can be called	[NP] ([can]) [be] called [NP]	hyperonymy/ hyponymy	Energy from sources like wind, water, and the sun is called renewable energy because there is an endless supply of it. Every country is responsible for how much greenhouse gas it releases, which can be called its “carbon footprint.”
due to	[NP] [be] due to [NP]	cause–effect	Current climate change is due to global warming, which is caused by the increase in GHG emissions as a result of human activities.

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation	Definitional context
identified as/by, can be identified as/by	[NP] ([can]) [be] identified as/by [NP]	association	Climate change is identified by variability in climate properties that persists for a prolonged period
made from	[NP] [be] made from ([article]) [NP]	result	Fossil fuels are made from the remains of living things which have been squeezed and heated underground for millions of years.
mean	[NP] means [NP]/ [VP]	cause–effect	A bigger carbon footprint means more emissions of carbon dioxide and methane, and therefore a bigger contribution to the climate crisis.
refer to (as)	[NP] refers to [NP] (as)	hyperonymy/ hyponymy	Climate change refers to the dramatic warming of the planet caused by increased levels of carbon dioxide in the atmosphere.
should be seen as	[NP] should be seen as [NP]	hyperonymy/ hyponymy	Climate governance should be seen as a "multi-level" process that includes the 12 following levels:
such as	[NP] ([punctuation mark]) such as [NP]	exemplification	It is produced from the burning of fossil fuels (such as coal and diesel) and the smelting of mineral ores that contain sulfur. One symptom of climate change is the increased number of extreme weather events, such as intense hurricanes, extended and wildfires.
that is	[NP] [punctuation mark] that is ([comma]) [NP]	synonymy	When we burn fossil fuels –that is, long-dead plants–inside our car engines Many scientists were eager to learn whether the total amount of these gases in the atmosphere (particularly carbon dioxide, that is CO_2) was lower in 2020 compared to 2019.

Appendix 2

Table 4. Lexical markers and lexical knowledge patterns in Croatian definitional contexts

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation	Definitional context
biti 'be'	[NP] [biti] [NP]	hyperonymy/hyponymy	Učinak staklenika je proces kojim zračenje Zemljine atmosfere zagrijava površinu na višu temperaturu nego što bi ona bila kada ne bi postojala atmosfera.
be'			"The greenhouse effect is the process by which radiation from the Earth's atmosphere heats the surface to a higher temperature than it would be if the atmosphere were not present."
činiti	[NP] čine [NP]	meronymy	Staklenik čine staklenički plinovi ugljik dioksid i metan.
'is composed of'			"The greenhouse is composed of greenhouse gases such as carbon dioxide and methane."
dolazi do	do [NPGen] dolazi zbog [NPGen]	cause–effect	Do klimatskih promjena dolazi zbog povišenja temperature Zemlje (globalnog zatopljenja) izazvanog dodavanjem neprirodno velike količine stakleničkih plinova u atmosferu.
'occur'			"Climate change occurs due to the Earth's rising temperature (global warming) caused by the addition of an unnaturally large amount of greenhouse gases into the atmosphere."
	do [NPGen] dolazi jer [NPNom]		Do efekta staklenika dolazi jer atmosfera sadrži plinove kao što su vodena para, ugljikov dioksid, metan i dušikov oksid.
			"The greenhouse effect occurs because the atmosphere contains gases such as water vapor, carbon dioxide, methane, and nitrous oxide."
dovesti do	[NP] [dovesti] do [NPGen]	cause–effect	Klimatske promjene dovode do većeg broja vremenskih ekstrema.
'result in'			"Climate change leads to an increased number of weather extremes."
			Klimatske promjene dovest će do povećanja broja toplijih dana i smanjenja broja hladnih dana.
			"Climate change will result in an increase in the number of warmer days and a decrease in the number of colder days."

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation	Definitional context
ime za	[NP] [ime] za [NP]	hyperonymy/hyponymy	Uragani, tajfuni i cikloni različita su imena za nasilne oluje koje nastaju iznad toplih tijela vode, kao što su Tihi ocean ili Karipsko more, kada postoji mnogo toplog, vlažnog zraka u atmosferi.
'name for'			"Hurricanes, typhoons, and cyclones are different names for violent storms that form over warm bodies of water, such as the Pacific Ocean or the Caribbean Sea, when there is a lot of warm, moist air in the atmosphere."
koristiti	[NP] [koristiti] [NPAk]	purpose	Klimatski modeli koriste matematičke jednadžbe kako bi opisali ponašanje elemenata Zemljina sustava koji utječu na klimu.
'use'			"Climate models use mathematical equations to describe the behavior of Earth's system elements that influence the climate."
nastati	[NP] [nastati] [NPIns]	result	Ti dodatni staklenički plinovi uglavnom nastaju spaljivanjem fosilnih goriva radi proizvodnje energije.
'produce'			"Those additional greenhouse gases are mainly produced by burning fossil fuels for energy production."
nastati od	[NP] [nastati] od [NPGen]	meronymy	Fosilna goriva nastala su od ostataka biljaka i životinja koji su živjeli prije milijun godina.
'formed from'			"Fossil fuels are formed from the remains of plants and animals that lived millions of years ago."
naziv za	[NP] [biti] naziv za [NP] /	hyperonymy/hyponymy	Globalno zagrijavanje je naziv za povećanje prosječne temperature zemljine atmosfere i oceana.
'term for'	[NP] naziv [biti] za [NP]		"Global warming is the term for the increase in the average temperature of the Earth's atmosphere and oceans."
nazivati se	[NP] [nazivati] se [NP] /	synonymy or hyperonymy /hyponymy	To se naziva "efektom staklenika" jer atmosfera funkcionira poput stakla u stakleniku – grije unutrašnjost.
'called'	[NP] se [nazivati] [NP]		"It is called the 'greenhouse effect' because the atmosphere functions like glass in a greenhouse, warming the interior."

Lexical marker	Lexical knowledge pattern	Lexical-semantic relation	Definitional context
naziva se još, nazivaju se još	[NP] [nazivati] se još (i) [NP] /	hyperonymy /hyponymy	Ugljen, nafta i prirodni plin nazivaju se još i fosilna goriva, a trenutačno predstavljaju glavni izvor energije u svijetu.
'also called'	[NP] se još [nazivati] (i) [NP]		"Coal, oil, and natural gas are also called fossil fuels and currently represent the primary source of energy in the world."
odnositi se na	[NP] [odnositi] se na [NP]	hyperonymy /hyponymy	Klimatske promjene odnose se na mnoge različite učinke globalnog zagrijavanja na klimatski sustav Zemlje.
'refer to'			"Climate change refer to many different effects of global warming on the Earth's climate system."
posljedica	[NP] ([biti]) [posljedica] [NPGen] /	result	Ono je posljedica emisije ugljikovog dioksida i metana, tzv. stakleničkih plinova, u atmosferu većinom iz industrijskih postrojenja.
'consequence of'	[NP] [posljedica] ([biti]) [NPGen]		"It is a consequence of the emission of carbon dioxide and methane, known as greenhouse gases, into the atmosphere, mainly from industrial facilities."
predstavljati	[NPAk] [predstavljati] [NP]	hyperonymy/ hyponymy	Klimu predstavljaju prosječni vremenski uvjeti na nekom mjestu tijekom relativno dugih vremenskih razdoblja (npr. 30 godina).
'represent'			"Climate represents the average weather conditions in a location over relatively long periods (e.g., 30 years). The greenhouse is composed of greenhouse gases, carbon dioxide, and methane."
uzrokovati	[NP] [uzrokovati] [NP]	cause-effect	Utvrđeno je kako klimatske promjene već sada uzrokuju veće zagrijavanje.
'cause'			"It has been determined that climate change is already causing greater warming."

Kaja Mandić

Faculty of Health Studies, University of Mostar, Bosnia and Hercegovina
kaja.mandic@fzs.sum.ba

Nursing Corpus and the Academic Collocation List

Abstract

Knowledge of frequent collocations is very important in the scientific discourse and writers should be familiar with collocations valued in a particular scientific discipline. Collocations enable English language learners to put words commonly used by native speakers to appropriate use in appropriate contexts. The study is based on the comparison of collocations from a corpus of articles written by native speakers of English and collocations from a corpus of English articles written by native speakers of Croatian, Bosnian, and Slovenian with the Academic Collocation List (ACL). The objective of the study is to extract the most frequent two-word collocations from the two nursing corpora and compare them with the ACL. The study will produce two corpora that are classified as specialized corpora. Both native and non-native corpora will include only English scientific articles from the field of nursing science. The results will be useful to generate a field-specific academic vocabulary list and teaching materials in order to strengthen learners' academic reading and writing proficiency.

Keywords: collocations, corpus linguistics, non-native speakers, data analysis, nursing

1. Introduction

English is the language of dissemination of academic texts and therefore of academic knowledge in all parts of the world. Its dominant role as the language of higher education is reflected in the number of scientific articles published in English. Using English as the main medium to transfer new research is the only way for scientists to become recognized and successful in their work and be cited by their peers. It is challenging to define the exact number of indexed articles published in English, but it is now estimated that more than 90% of scientific articles are written in English, regardless of the author's native language (Hamel 2007; Lillis, and Curry 2010; Montgomery 2013). According to Baji et al. (2022), nearly 80% of all indexed journals and the world's top fifty journals are in English. However, non-native writers face difficulties in the production of fluent academic texts and adequate vocabulary knowledge (Crystal 2012; Montgomery 2013). English for Specific Academic Purposes (ESAP) plays a very important role in educating existing and new users

who will join the international network of scientific English (Wood 2001: 71-73). The corpus-based approach is recognized to be particularly suitable for research and teaching of English for Specific, Professional, or Academic Purposes. Lindquist (2009) points out that if one is interested in the workings of a particular language, one efficient way to do this is by using corpus methodology. This means that corpus linguistics allows us to see how language is used today in different contexts, thus enabling us to teach and learn the language in a different way (Bennett 2010).

2. Definition of Collocation and Review of Previous Studies

Collocations are seen as a necessary component of foreign language spoken and written competence and are considered an essential component in developing native-like fluency (Nesselhauf 2005). This means that they strongly contribute to language efficacy, namely language comprehension and production. Collocations are of particular importance for learners striving for a high degree of competence in the second or foreign language; however, they are also important for less ambitious learners, as they enhance both accuracy and fluency (Nesselhauf 2003).

It is a challenging task to define the term collocation. Sinclair as a pioneer of corpus-based research defines collocation as: "The occurrence of two or more words within a short space of each other in text" (1991: 170). A simple definition of the concept of collocation is of arbitrarily restricted lexeme combinations (Nesselhauf 2005). Stubbs (2001) sees a collocation as a relationship of habitual co-occurrence of words, either lemmas or word-forms. Collocations are one type of a group of expressions whose importance in language has been increasingly recognized in recent years. This group of expressions has been variously called prefabricated units, prefabs, phraseological units, (lexical) chunks, multi-word units, or formulaic sequences (Najafi and Talebinezhad 2018). Lexical collocations are the subject of this study, more precisely two-word lexical collocations, noun-noun and adjective-noun combinations. The corpora constructed are specialized corpora consisting of scientific articles from the field of nursing.

Collocations play an essential role in language learning and seem to be the basis of creative language development. They are essential for spoken and written fluency, and the availability of a large number of prefabricated units makes fluent language possible. The use of collocations supports comprehension, as the reader can understand the meaning without having to attend to every word (Aitchison 2003; Hunston and Francis 2000; Najafi and Talebinezhad 2018).

In English for Specific Purposes (ESP), learners have to learn high-priority vocabulary items, which need to be selected and included into learning materials and class activities. What is important in order to ensure their effective learning is that students turn a high proportion of input to which they are exposed into intake (Kavaliauskienë and Janulevièienë 2001). In terms of specific genre analysis and ESP, there

is a wide range of evidence on the relationship between the notion of collocation and scientific writing and style (Gledhill 2000). Knowledge of collocations is essential for English as a Foreign Language (EFL) learners since they take up a large portion of language. According to Partington (1998), registers rely largely on prefabrication, and in many genres of writing, pre-cooked expressions are vital elements. Collocations are a big concern for non-native speakers as they are challenged with the task of producing proficient texts (Seretan et al. 2003: 91), and they can become an issue for them due to the inference of the mother tongue. Hyland (2008) observes that professional collocations are familiar to readers and writers who regularly participate in a particular discourse or a given language community. This conclusion was reached by observing the absence of such clusters and lack of fluency in specific registers in novices or newcomers in these communities. According to Hill (2000), collocations estimate up to 70% of everything we say, hear, read, and write.

A number of corpus-based studies suggests that non-native writers underuse collocations when compared to native peers (Demir 2017; Durrant and Schmitt 2009; Liu and Shaw 2001). Studies on academic word lists have a longer history and greater popularity in corpus linguistics than those on academic collocation lists, but findings and methodological methods from those studies provide a tremendous support for collocation studies. These types of lists entail frequent lexical combinations that enhance English language learners' proficiency in a specialized field. The Academic Word List (AWL) by Coxhead (2000) was the first list that applied frequency, range, and specialized occurrence. It used a corpus of 3.5 million words from four disciplines: commerce, science, humanities, and law (Coxhead and Nation 2001: 255). The AWL inspired numerous studies on word lists that later proved many flaws of Coxhead's list. One of its flaws was that it did not include texts from medicine and used only a limited number of short texts from the law subcorpora (Hyland and Tse 2007). In 2015, Yang also conducted a study closely related to the current one, and created a Nursing Academic Word List. This list contains the most frequent words found in nursing research articles. These findings suggest that it is necessary to generate field specific lists for EFL nursing students and professionals in order to strengthen their academic written and spoken comprehension.

A review of collocation studies and non-native speakers of English indicates that there is a lack of collocational knowledge in EFL learners and lack of sufficient exposure to specific types of collocations (Hill 1999, Henriksen 2013, Nesselhauf 2003, 2005). An example is a corpus of 150 cancer research articles, the results showing that new science is founded on a system of preferred expressions, and that collocation is a fundamental mechanism that allows for new formulations to take place throughout the text (Gledhill 2000). Fan (2009) examined a native and non-native corpus composed of sixty essays written by Hong Kong secondary school students and sixty British essays collected in a comprehensive school in northern England. The results revealed that Hong Kong learners used a limited number of collocations compared to British students. Durrant and Schmitt (2009) investigated the

use of collocations in native and non-native student essays. Non-native essays were written by Turkish and Bulgarian first-year undergraduate and postgraduate students. The findings confirm that native writers use significantly more low-frequency collocations than their non-native peers. Demir (2017) investigated research articles from leading journals on English Language Teaching. On two small corpora, he found that native writers used a higher percentage of collocations than Turkish authors. Two Arabic authors explored the production of English adjective-noun collocations by Arab ESL (English as a Second Language) writers as compared to English native writers. The findings revealed that Arab ESL writers use collocations with a greater frequency, but in a context often judged as "not appropriate" by English language instructors (Qureshi and Nurmukhamedov 2018). Chen and Baker (2010) compared the use of frequent word combinations in academic writing by Chinese EFL university students, native English-speaking university students, and native expert writers. The study revealed differences and similarities between native and learner academic writing. The use of word combinations in non-native and native student essays was very similar, but native expert writers used a wide range of word combinations when compared to others. Navarro Gil and Martinez Caro (2019) explored the use of lexical bundles in a corpus of bachelor dissertations from linguistics and medicine written in English by Spanish native students and compared it with published research articles. The product of the study was a list of 218 different bundles with the intention of assisting ESL and EFL learners in their academic writing. Pavičić Takač and Lukač (2013) confirm that collocations are indeed a problematic area for non-native users of medical English and that teaching of vocabulary through collocations can be very useful. The study indicates that teaching of medical collocations has a significant effect on vocabulary retention, and instructors should strive towards incorporating collocations into their teaching content. According to the literature review, we can conclude that EFL and ESL writers' use of collocations equals the native writers in terms of frequency and accuracy.

3. Assembly of the Native and Non-Native Nursing Corpora

The two corpora in this study are composed entirely of scientific articles from the field of nursing and are classified as specialized. The argument is that we should use specialized corpora in order to understand academic and professional language instead of general corpora (Connor and Upton 2004). The spread of novel nursing knowledge, which stems from scientific research and is published in nursing journals, is vital for the development of the nursing profession, and can have a much higher quality and be provided much faster compared to what students and working professionals can find in nursing textbooks (Campbell-Crofts 2012: 6). We also believe that an entirely article-based corpus better supports the study and its purposes.

The corpus of English native nursing scientific articles (NNSAC) comprises

1,119,441 words from articles of high-quality journals. The articles were chosen on the basis of the NHARS Selected List of Nursing Journal from 2016. We collected 262 nursing articles from ten nursing journals (see Appendix 1 for a complete list of journals and number of articles). Accessibility and availability in electronic form were an important criterion; therefore, the number of articles from journals varies, since not all journals and articles are open access and available for download. Our second corpus, the corpus of non-native English nursing scientific articles (NNNSAC), consists of 249 nursing scientific articles from seven nursing journals and includes 930,786 words written by Croatian, Bosnian, and Slovenian authors (see Appendix 2 for a complete list of journals and number of articles). This corpus of non-native articles can be classified as learner corpora. The non-native journals represented here are the only ones available in this field of research in countries where English is not the native language of the speakers who are the subjects of this study. We collected articles from four journals from Croatia, two journals from Bosnia and Herzegovina, and one journal from Slovenia. There is no clear classification of these types of journals in the region; therefore, we selected all that were available and in which nursing scientific articles are published in the English language.

After corpora assembly, the first step was to extract and compare collocations produced by native and non-native speakers of the English language, and compile a list of English nursing collocations. The focus was on noun-noun and adjective-noun collocations. This particular combination has also been confirmed as the most frequent by the founders of the Academic Collocation List (ACL), our reference corpus. Further, we compare the list of nursing collocations (see Appendix 3 for a complete list of collocations) and the ACL (https://www.pearsonpte.com/teachers/academic-collocation) in order to establish whether there is a need for a specialized list of collocations for nurses.

3.1. Methods

The selection criterion for the differentiation of native and non-native authors was done according to Wood. In current literature, there is no exact way or agreed criteria to ascertain the first language of a writer, especially in published research articles. As Wood suggests, first authors must have names that are native to the country in question and be affiliated with an institution where this language (English) is spoken as the first language. Wood does not claim that every scientist subjected to this criterion is in fact a native or non-native speaker of English, but that the overall proportion of native and non-native speakers will be approximated by such an operational definition (2001: 79). In our non-native corpus for any author to be regarded as an L1 (native language) Croatian, Bosnian, or Slovenian writer, they must be affiliated with an institution in their home country and have the first and last name considered native to the same country. The same criteria were applied for L1 English writers. We removed all scientific research papers that did not meet Wood's criteria.

We used Shin and Nation's (2008) criteria for identifying collocation but adapted it to serve the purpose of the study. We adapted two out of six criteria: the third criterion on the reference corpus was eliminated, since our reference corpus is the Academic Collocation List, as well as the fourth criterion for frequency, in which we also adopted the normed frequency from Ackerman and Chen (2013), stating that a collocation had to occur 0.2 times per one million words.

Three software programs were used in the present study. The ABBY Fine Reader, for the conversation of PDF files into simple text files. The second software is the Corpus Builder v.2.3., which is an online software program that assembles text, html, or Word files into a single combined file. The third and most important software used in the study is TermeX version 1.0. It is a tool for automatic collocation extraction and terminology lexical construction. It is based on statistical measures called association measures (Delač 2009: 2). What differs it from the rest is that it provides a much wider range of association measures to choose from and outperforms the majority of tools in terms of processing speed.

4. Results and Discussion

The corpus-based study retrieved a total of 480 collocation combinations from the NNSAC and 272 collocations from the NNNSAC. The final product of the comparison of the two corpora is a list of 629 nursing collocations.

Adjective-noun combinations of collocations are the most frequent form of two-word collocations used by both native and non-native speakers of the English language in our study. This particular combination has also been confirmed as the most frequent one by the founders of the ACL. Since the percentages for the use of adjective-noun combinations are similar in both corpora, 53% for the NNSAC and 56% for the NNNSAC, we can conclude that the results oppose those previously found by Durrant and Schmitt (2010). In their study on the exposure to collocations non-native learners receive, with an emphasis on adjective-noun combinations and their retention, they found that there is a shortfall in the use of adjective-noun collocations by non-native learners, and that the use is limited to a certain set of collocations and is not native-like. Our study shows that both native and non-native writers use adjective-noun combinations to a larger extent than other combinations. Since the highest number of our non-native articles comes from Croatian nursing journals, the results on the most frequent collocation combinations might be under the influence of the first language and the fact that the most common collocation pairs in the Croatian language are formed by an adjective and a noun (Hudeček and Mihaljević 2012: 6). Here are some examples in Croatian: *duljina djelovanja, vijek trajanja, stopa rasta, stanje mirovanja, mjera sigurnosti*.

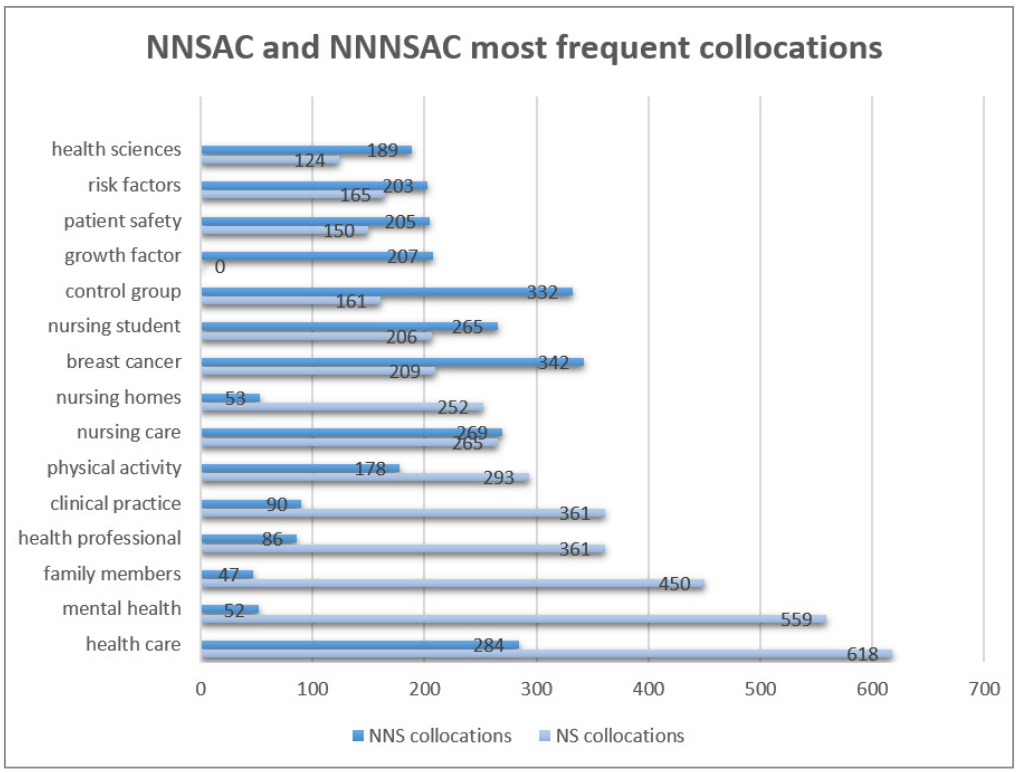

Figure 1. Frequency of the most common collocations from the NNSAC and NNNSAC

Figure 1 shows the most frequent collocations from native and non-native nursing corpora. We can see 10 colocations from the NNSAC and 10 from the NNNSAC as well normed frequency for each extracted collocation. Collocations extracted from the non-native corpus have lower frequencies than those found in the native corpus. The most frequent collocation from the native corpus, *health care*, has the normed frequency of 618, while the most frequent collocation from the non-native corpus, *breast cancer*, has the normed frequency of 342. Collocations with the highest frequencies from the NNSAC are exclusively professional nursing collocations, and three from the NNNSAC (*risk factors, growth factor,* and *control group*) are general collocations used across different scientific fields. Some other examples of professional nursing collocations from the native corpus include the following: *blood sugar, care facility, care staff, care team, clinical education, family caregivers, healthcare team, nurse leader, nurse burnout*.

These results could also be under the influence of topics and particular matters being studied in the selection of native and non-native issues and articles. They could be a by-product of the topic of breast cancer studied in the selection of non-native scientific research articles. The same argument could be true for the native corpus and its second most frequent collocation, *mental health*. This particular collocation could have a high frequency due to the topic studied, since its normed frequency in the non-native corpus is only 52.

We compared two very different corpora regarding size. When combined, our nursing corpora consists of 2,050,227 words from nursing scientific articles written

by native and non-native speakers of English. The corpus of the ACL comprises over 37 million words of academic written and spoken texts from five major English-speaking countries. The list of nursing collocations is assembled on the basis of overlapping collocations from the two corpora of nursing articles, with the addition of all other extracted collocations. The number of overlapping collocations with the ACL is only 151 or to be more accurate only 6% of collocations from the Academic Collocation List appear on the nursing list. Figure 2 represents the proportion of collocations from the ACL and the list of nursing collocations.

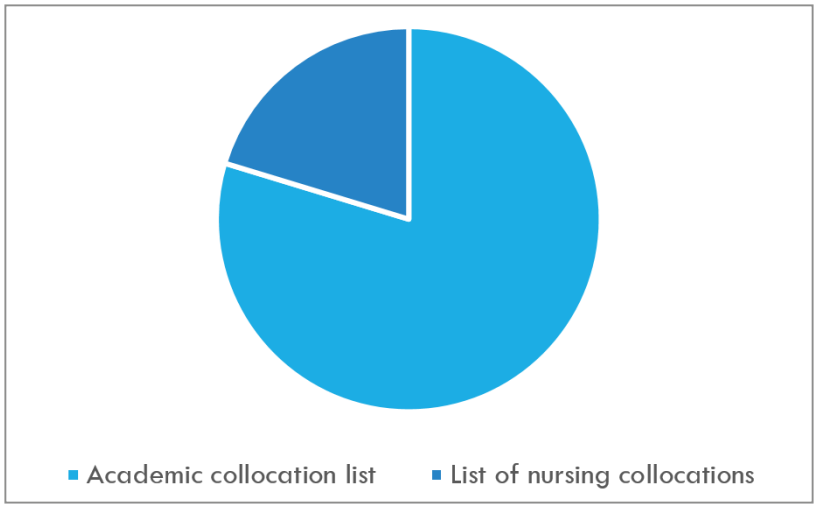

Figure 2. Proportion of overlapping collocations from the list of nursing collocations and the ACL

An interesting finding is that nursing collocations have significantly higher frequencies on the list of nursing collocations than on the ACL. Out of the 151 overlapping collocations, only six collocations (*mental health, mental illness, physical activity, physical symptom,* and *adverse effect*) are classified as professional nursing collocations. All other overlapping collocations are mostly general ones used in various scientific fields, for example: *available data, research methodology*, and *study results*. Other examples of overlapping collocations with the ACL include the following: *blood samples, clinical characteristics, community health, diagnostic test, family medicine, food intake, low dose, nursing team, speech disabilities*. We can conclude that nursing collocations have a low occurrence rate on the Academic Collocation List.

5. Limitations

A limitation to the representativeness of the study is that we only looked at noun combinations of collocation (noun-noun and noun-adjective), and our reference list, the ACL, consists of several other collocation combinations. The comparison of the two lists could be more comprehensive if we also included other types of lexical collocations found on the ACL. If this had been the case, the statistical results of this

study would have a much broader framework. Another limitation is that our corpus is assembled only from scientific articles, and the Pearson International Corpus of Academic English contains written and spoken language from lectures, seminars, textbooks, and journal papers. The two nursing corpora of this study could be additionally verified with larger corpora of other medical and healthcare genres.

6. Conclusion

Both the ACL and the list of nursing collocations contain collocations that English learners are likely to encounter during their academic education. The ACL represents the most important cross-disciplinary collocations that can help learners increase their collocation competence and thus their proficiency in academic English. The list created as part of this study comprises collocations purposely tailored for EFL nurses in order for them to better acquire English for specific and academic purposes.

There is a need for an independent list of nursing collocations that will be available for nursing students, working nurses, and other medical professionals who are in need to use English terminology in their writing and specific job situations. This is supported by Gilquin et al. (2007), who state that if we rely solely on native corpus data, the learning materials fail to provide non-native learners with the information that is most valuable to them. According to Hyland (2008), writers need a familiarity with both the clusters which characterize their disciplines and those which are valued in the particular genres of those disciplines. The existence of an independent list of nursing collocations is further supported by Mackin (1978), who claims that collocations are countless, and it is difficult for non-native speakers of English to rule out those unnecessary for them. Our results also support Hyland and Tse's (2007) claim that general vocabulary lists do not reflect the real needs of English for Specific or Academic Purposes students. Providing learners with a list of most frequent academic nursing collocations helps them to overcome the deficiency of academic nursing vocabulary competence. Another reason for an independent list of collocations is the increasing emphasis of researchers on the importance of collocation acquisition in foreign language teaching. Pawley and Syder state that the inventory of multiword combinations known to the mature speaker of English should account for hundreds of thousands (1983: 92).

The development of a discipline specific collocation list hopes to inspire and draw attention of teachers, students, learners, and nurses on the importance of this type of English vocabulary. In order to save time and improve acquisition of specific academic vocabulary, mastery of the list of nursing collocations would be a more valuable decision, since these collocations proved to have higher frequency in nursing scientific articles, which are the base for any future work and professional development in an evidence-based healthcare profession such as nursing. The study results can be used for future design of English nursing materials for reading, writing, and comprehension purposes. They are also beneficial and helpful for nursing

students, nurses, teachers, ESP material designers, and all researchers in applied linguistics and health sciences. The list can serve as a basis for developing English teaching materials for English for Academic Specific Purposes and English for (Professional) Nursing Purposes. Material developers of professional English textbooks should consider using collocations, since multi-word vocabulary items are very frequent in professional academic texts. English foreign language learners who are interested in continuing nursing studies can use the list to help them expand their vocabulary in size and practice, and use the list of nursing collocations as an aid in their academic and scientific writings.

References

Ackermann, Kirsten; Chen, Yu-Hua. 2013. "Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach." *Journal of English for Academic Purposes* 12(4). 1235-247.

Bahji, Aness; Acion, Laura; Laslett, Anne-Marie; Adinoff, Bryon. 2022. "Exclusion of the non-English-speaking world from the scientific literature: Recommendations for change for addiction journals and publishers." *Nordic Studies on Alcohol and Drugs* 40(1). 6-13.

Bennett, Gena. 2010. *Using Corpora in The Language Learning Classroom: Corpus Linguistics for Teachers.* Michigan ELT.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Byrd, Pat; Coxhead, Averil. 2010. "On the other hand: Lexical bundles in academic writing and in the teaching of EAP." *Papers in TESOL* 5. University of Sydney. 31-64.

Campbell-Crofts, Sandra. 2012. The future of nursing journals. *Renal Society of Australasia Journal* 8(1). 6-8.

Chen, Yu-Hua; Baker, Paul. 2010. "Lexical bundles in L1 and L2 academic writing." *Language, Learning and Technology* 14(2). 30-49.

Connor, Ulla; Upton, A. Thomas. 2004. *Discourse in the Profession: Perspectives from the corpus linguistics.* Amsterdam: John Benjamins Publishing Company

Coxhead, Averil; Nation, Paul. 2001. "The specialised vocabulary of English for academic purposes." In Flowerdew, J. & Peacock, M. (Eds.) *Research perspectives on English for academic purposes*. Cambridge University Press. 252-267.

Coxhead, Averil. 2000. "A New Academic Word List." *TESOL Quarterly* 34(2). 213-238.

Crystal, David. 1992. *The Cambridge encyclopaedia of language.* Cambridge: Cambridge University Press.

Delač, Davor; Krleža, Zoran; Šnajder, Jan; Dalbelo Bašić, Bojana. 2009. "TermeX: A Tool for Collocation Extraction". In Glebukh, A. (Ed.) *Computational Linguistics and Intelligent Text Processing*. 149-157.

Delač, Davor. 2009. *TermeX v1.0.* University of Zagreb, Faculty of Electrical Engineering and Computing.

Demir, Cuneyt. 2017. "Lexical Collocations in English: A Comparative Study of Native and Non-native Scholar of English". *Journal of Language and Linguistic Studies* 13(1). 75-78.

Durrant, Phillip; Schmitt, Norbert. 2009. "To what extent do native and non-native writers make use of collocations?" *International Review of Applied Linguistics in Language Teaching* 47(2). 157-177.

Fan, May. 2009. "An exploratory study of collocational use by ESL students – A task based approach". *System* 37(1). 110-123.

Gilquin, Gaëtanelle; Granger, Sylviane; Paquot, Magali. 2007. "Learner Corpora: The Missing Link in EAP Pedagogy". *English for Academic Purposes* 6(4). 1-26.

Gledhill, Christopher. 2000. "Collocations in Science Writing". *Language in Performance Series* 22. 7-20.

Granger, Sylviane; Meunier, Fanny. 2008. *Phraseology, An Interdisciplinary Perspective.* John Benjamins Publishing Company.

Hamel, Rainer Enrique. 2007. "The dominance of English in the international scientific periodical literature and the future of language use in science." *AILA Review* 20(1). 53–71.

Henriksen, Brigit. 2013. *Research on L2 learners' collocational competence and development – a progress report.* Copenhagen University.

Hill, Jimmie. 1999. "Collocational competence". *English Teaching Professional* 11. 3-6.

Hill, Jimmie. 2000. "Revising priorities: From grammatical failure to collocational success." In Lewis, M. (Ed.) *Teaching collocation: Further developments in the lexical approach*. Hove: LPT. 47-67.

Hoey, Michael. 2005. *Lexical Primings: A new theory of words and language*. Oxon: Routledge.

Hsu, Wenhua. 2013. "Bridging the vocabulary gap for EFL medical undergraduates: the establishment of a medical word list". *Language Teaching Research* 17(4). 454-484.

Hudeček, Lana; Mihaljević, Milica. 2012. *Hrvatski terminološki priručnik*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Hyland, Ken; Tse, Polly. 2007. "Is there an academic vocabulary?" *TESOL Quarterly* 41(2). 235-253.

Hyland, Ken. 2008. "Academic clusters: text patterning in published and postgraduate writing". *International Journal of Applied Linguistics* 18(1). 41-62.

Kavaliauskienë, Galina; Janulevièienë, Violeta. 2001. "Using the Lexical Approach for the Acquisition of ESP Vocabulary." *The Internet TESL Journal* 7(3). 1-6.

Kosem, Iztok; Krek, Simon; Gantar, Polona. 2020. "Defining Collocation for Slovenian Lexical Resources." *Slovenščina* 20(2). 1-27.

Lillis, Theresa; Curry, Mary Jane. 2010. *Academic writing in a global context: The politics and practices of publishing in English*. London, UK: Routledge

Lindquist, Hans. 2009. *Corpus Linguistics and the Description of English*. Edinburgh University Press.

Mackin, Ronald. 1978. *On collocations: Words shall be known by the company they keep*. In Strevenes, P. (Ed.) In honour of A.S. Hornby. Oxford University Press. 149-165.

Montgomery, Scott. 2013. *Does science need a global language? English and the future of research.* Chicago, IL: University of Chicago Press.

Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2). 223-242.

Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus*. John Benjamins Publishing Company.

Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Pawley, Andrew; Syder, Frances Hodgetts. 1983. Two puzzles for linguistic theory native like selection and nativelike fluency. In Richards, J. C. & Schmidt, R. W. (Eds.) *Language and communication*. London: Longman. 191-230.

Qureshi, Muhammas; Nurmukhamedov, Ulubek. 2018. "Use of Collocations in Freshman Composition: Implications for L1 English and Arabic ESL Writers." *International Journal of Lexicography* 30(4). 454-482.

Seretan, Violeta; Nerima, Luka; Wehrli, Eric. 2003. Multi-Word Collocation Extraction by Syntactic Composition of Collocational Bigrams. In Nicolov N., Bontcheva, K., Angelova, G. & Mitkov, R. (Eds.) *Recent Advances in Natural Language Processing III*. John Benjamins Publishing. 91-100.

Shin, Dongkwang; Nation, Paul. 2008. "Beyond single words: the most frequent collocations in spoken English." *ELT Journal* 62(4). 339-348.

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

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Wang, Jing; Liang, Shao-Lin; Ge, Guang-Chun. 2008. "Establishment of a medical academic word list." *English for Specific Purposes* 27(4). 442-458.

Wood, Alisatir. 2001. International scientific English: The language of research scientists around the world. In Flowerdew J. & Peacock M. (Eds.) *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press. 71-83.

Yang, Ming-Nuan. 2015. "A nursing academic word list." *English for Specific Purposes* 37(1). 27-38.

Appendix 1

Table 1. Journals and number of articles for The corpus of English native nursing scientific articles

No.	Journal	Number of articles
1	Journal of Clinical Nursing	83
2	Journal of Advanced Nursing	34
3	Scandinavian Journal of Caring Science	28
4	Cancer Nursing	26
5	International Journal of Nursing Practice	20
6	Journal of Nursing Management	17
7	Journal of Nursing Education	11
8	Journal of Psychiatric and Mental Health Nursing	13
9	Oncology Nursing Forum	8
10	Journal of Women's Health	22
Total		262

Appendix 2

Table 2. Journals and number of articles for The corpus of English non-native nursing scientific articles

No.	Journal	Number of articles
1.	Acta Clinica Croatica	38
2.	Bosnian Journal of Basic Medical Sciences	98
3.	Croatian Nursing Journal	22
4.	Journal of Applied Health Sciences	6
5.	Journal of Health Sciences	35
6.	Slovenian Nursing Review	15
7.	Southeastern European Medical Journal	31
Total		249

Appendix 3

Table 3. List of nursing collocations

academic life	academic year	action research	acute care	acute illness
adverse effect	adverse events	age group	age range	alcohol abuse
alcohol consumption	alcohol use	annual review	anxiety levels	assessment tools
available data	average age	average score	average value	behavior pattern
beneficial effect	bipolar disorder	blood clot	blood donor	blood flow
blood pressure	blood samples	blood sugar	blood vessels	body composition
body fat	body image	body language	body mass	body temperature
body weight	bone marrow	bone mineral	bowel disease	breast cancer
brief review	broad range	caesarean section	cancer care	cancer cell
cancer center	cancer diagnosis	cancer nurse	cancer nursing	cancer patients
cancer screening	cancer survivor	cancer treatment	cardiovascular disease	care activities
care environment	care facility	care home	care management	care plans
care providers	care provision	care recipient	care services	care staff
care systems	care team	care unit	case report	case study
cell growth	cell proliferation	cell tumor	cervical cancer	chemotherapy administration
chest injuries	children's hospice	chronic condition	chronic disease	chronic illness
chronic pain	chronic wounds	cigarette smoke	clinical care	clinical characteristics
clinical community	clinical decision-making	clinical education	clinical examination	clinical experience
clinical features	clinical nurse	clinical nursing	clinical outcomes	clinical placement
clinical practice	clinical reasoning	clinical setting	clinical skills	clinical staff
clinical study	clinical trials	cognitive deficits	cognitive functions	cognitive impairments
cognitive performance	cohort study	collective action	colon cancer	communication skills
communication strategies	community care	community health	community nurses	community setting

comparative analysis	comparative study	conceptual framework	conservative management	consultation models
continence care	continued use	continuous variable	control group	conventional methods
coping strategy	coronary artery	coworker support	critical analysis	critical thinking
cruciate ligament	cultural background	cultural differences	current research	current study
daily activities	daily living	data analysis	data collection	data extraction
data saturation	delivery room	dementia care	demographic characteristics	demographic data
dental health	dependent variable	depression scale	descriptive statistics	diabetes patients
diagnostic test	direct care	eating disorders	education level	educational programme
educational qualification	elderly cancer	elderly patients	eligibility criteria	emergency department
emergency room	emotional distress	emotional exhaustion	emotional intelligence	emotional reaction
emotional support	empirical research	empirical study	erectile dysfunction	essential role
ethical approval	ethics committee	ethnic group	exclusion criteria	exercise program
experimental design	expert nurse	eye contact	facial expression	family care
family caregivers	family medicine	family members	family support	fatty acids
final analysis	final version	focus groups	foetal abnormalities	food intake
foot care	future research	future study	gastric cancer	gastrointestinal tract
gender differences	general anesthesia	general health	general hospital	general population
general practitioner	germ cells	global health	graduate nurse	great impact
growth factor	gynecological cancer	health behavior	health care	health condition
health education	health information	health insurance	health issues	health organization
health outcomes	health policy	health problems	health professionals	health promotion
health research	health sciences	health service	health status	health system
health wards	healthcare assistant	healthcare costs	healthcare encounters	healthcare environment
healthcare institutions	healthcare organizations	healthcare personnel	healthcare professional	healthcare providers
healthcare services	healthcare settings	healthcare staff	healthcare system	healthcare team

hearing loss	heart disease	heart failure	heart rate	high concentration
home care	hormone therapy	hospital admission	hospital care	hospital environment
hospital nurse	hospital setting	hospital staff	hospital stay	hot flushes
human research	human resource	immune system	important aspects	important factor
inclusion criteria	independent variable	individual level	individual needs	informal caregivers
information sharing	information sheet	insertion pain	intellectual disability	intensive care
interdisciplinary collaboration	internal conflict	international journal	international study	interpersonal relationships
ischemic stroke	joint pain	key concept	key findings	key role
kidney disease	knee injury	knee joint	labour induction	lacrimal sac
large percentage	leading cause	learning environment	learning strategy	left ventricle
life quality	limited information	literature review	local anesthetic	logistic regression
longitudinal study	low dose	low level	low score	lower frequency
lower rates	lung cancer	magnetic resonance	main category	major cause
major change	male caregivers	male infertility	malignant tumor	management strategies
managerial competence	manual restraint	marital status	maternity leave	mean age
mean score	median age	medical care	medical centers	medical comorbidity
medical condition	medical imaging	medical record	medical research	medical sciences
medical staff	medical student	medical treatment	medical unit	medical university
medication adherence	medication preparation	medication safety	melanoma cells	mental disorder
mental health	mental healthcare	mental illness	mental retardation	metabolic activity
metabolic syndrome	mortality rate	multidisciplinary team	muscular dystrophy	myocardial infraction
negative attitude	negative effect	negative impact	nervous system	normal distribution
numerous studies	nurse assessment	nurse assistant	nurse association	nurse burnout
nurse care	nurse competence	nurse educator	nurse intervention	nurse leaders
nurse managers	nurse practitioner	nurse report	nurse research	nurse specialist

nursing activities	nursing assessment	nursing care	nursing curriculum	nursing documentation
nursing education	nursing homes	nursing interventions	nursing journals	nursing management
nursing practice	nursing profession	nursing research	nursing roles	nursing schools
nursing staff	nursing student	nursing study	nursing team	observational study
occasional tiredness	oncology nurse	oncology nursing	oncology patients	online version
open access	organ transplantation	original work	outcome measures	outpatient service
outpatient utilization	ovarian cancer	overall level	pain assessment	pain management
pain rehabilitation	palliative treatment	pancreatic cancer	parental cancer	parental role
pathological features	patient care	patient characteristics	patient education	patient experiences
patient needs	patient outcomes	patient population	patient safety	patient satisfaction
pediatric care	peer support	permanent teeth	personal care	pharmacological interventions
phenomenological analysis	physical activity	physical care	physical environment	physical exercise
physical health	physical restraints	physical symptom	pilot study	poor communication
poor prognosis	positive attitude	positive correlation	positive effect	positive experience
positive impact	positive outcome	positive value	postmenopausal women	postoperative care
postpartum depression	potential risk	prenatal care	preoperative care	preoperative protocol
present study	preterm infants	previous research	previous study	primary care
primary data	primary healthcare	professional body	professional care	professional development
professional experience	professional practice	professional status	professional support	professional training
professional work	prognostic factors	prognostic value	prospective study	prostate cancer
psychiatric hospital	psychiatric nurse	psychological distress	psychological empowerment	psychological symptoms
public health	pulmonary disease	qualitative analysis	qualitative approach	qualitative data
qualitative design	qualitative method	qualitative methodologies	qualitative research	qualitative study
quality appraisal	quality assessment	quality indicators	quality outcomes	quantitative data
random sample	rapid response	rating scale	recent decades	recent study

rehabilitation clinic	relevant factors	replacement therapy	research article	research assistant
research ethics	research evidence	research fellow	research findings	research method
research methodology	research project	research topic	residential care	resonance imagining
respiratory depression	respiratory rate	retrospective study	risk assessment	risk factors
role model	room temperature	safety vests	sample size	sampling methods
scientific research	secondary care	secondary data	secondary education	secondary school
sectional study	sedentary behavior	service delivery	service providers	service user
sexual activity	sexual health	sexual intercourse	sexual orientation	shared decision
shared experience	shift work	side effects	significant amount	significant change
significant correlation	significant difference	significant effect	significant impact	significant improvement
significant increase	significant relationship	similar results	sleep behaviors	sleep conditions
sleep disorder	sleep disturbance	sleep duration	sleep patterns	sleep quality
small percentage	smooth muscle	social activity	social care	social contact
social function	social interaction	social organization	social status	social welfare
social worker	socioeconomic status	soft tissue	speech disabilities	spinal cord
sport activities	staff members	staff training	standard deviations	statistical analysis
statistical significance	stem cells	stress level	stroke patient	student nurse
study design	study findings	study group	study limitations	study methods
study participants	study period	study programme	study report	study results
study sample	substance use	support programme	surgical patients	surgical procedure
surgical treatment	survival rate	systematic study	target audience	team members
theoretical framework	time frame	time period	total income	total number
total score	tract infections	training course	training programme	training session
traumatic stress	treatment options	undergraduate nurse	university hospital	urinary catheter
urinary incontinence	urinary tract	usual care	vaginal delivery	vast number

verbal communication	vital signs	vocational nurse	weight gain	weight loss
wide range	wide variation	work experience	work overload	work practice
working conditions	working environment	workplace violence	younger generation	

Košuta Estera Lerga

Faculty of Humanities and Social Sciences, University of Rijeka, Croatia
kosuta.lerga@gmail.com

Lucia Načinović Prskalo, Marija Brkić Bakarić

Faculty of Informatics and Digital Technologies, University of Rijeka, Croatia
lnacinovic@uniri.hr, mbrkic@uniri.hr

Adapting the Generic English-Croatian NMT Model to a Religious Domain

Abstract

Recent discoveries in the field of artificial intelligence have significantly impacted various professions, including the translation industry, leading to notable changes in translators' work processes. The study presented in this article indicates that today any translator, even those without advanced IT skills, can develop a higher quality Neural Machine Translation (NMT) system based on their own texts. This paper evaluates Google's AutoML Translation service, which enables users to train high-quality models using their own text data. Specifically, AutoML Translation integrates an additional layer that tailors the generic Translation API model to a specific domain. The training process involves providing a user-defined dataset containing aligned sentences in the source and target languages. Google's AutoML Translation service was used to adapt the base English-Croatian Google NMT model to the field of religion. Following a brief introduction to machine translation, this paper outlines the key aspects of the training and evaluation processes. Additionally, it presents two corpora employed in the training phase. The results demonstrate that a customized model outperforms the base model, as evidenced by the BLEU score.

Keywords: automatic translation, domain adaptation, neural machine translation, religious domain, aligned parallel corpora

1.Introduction

When presented with a message, there are numerous methods to effectively communicate its intended meaning. Similarly, it is highly probable that any translator among a group would offer slightly varied translations of a message originally conveyed in the source language. The inherent subjectivity of this task contributes to the inherent complexity associated with machine translation (MT) and its subsequent evaluation.

MT approaches generally fall into two categories: rule-based methods and data-driven approaches. Rule-based methods were predominant before the 2000s,

characterized by their subjective and labour-intensive nature, making them susceptible to unforeseen language phenomena and scalability issues. Rule-based systems involve linguists crafting specific rules to transform source language into target language. In contrast, data-driven approaches emerged in the 1980s, aiming to learn translation patterns by analysing numerous pairs of human-translated segments. Data-driven MT encompasses example-based MT, statistical machine translation (SMT), and neural machine translation (NMT). Example-based MT retrieves similar examples from pairs of human-translated sentences to generate translations. The concept of SMT originated in the late 1980s but gained mainstream acceptance around 2000. Neural network approaches gradually integrated into various components of SMT, reaching their full potential from 2015 onwards. Despite their existence in the previous century, the computational complexity associated with these methods hindered serious advancements beyond toy examples (Koehn 2020). Sequence-to-sequence models eventually replaced traditional phrase-based approaches in NMT systems based on the encoder-decoder paradigm (Chen et al. 2018). The first production Google's NMT system was presented by Wu et al. (2016).

Comparative quality analyses of neural machine translation systems versus statistical machine translation systems, as detailed by Koehn and Knowles (2017), indicate that neural machine translation systems often achieve lower quality on out-of-domain texts, favouring fluency over adequacy to a point of sacrificing the latter. Consequently, they are more sensitive to domain mismatches compared to SMT systems (Ruopp 2020). Furthermore, these models exhibit a "steeper" learning curve concerning data volume, resulting in reduced performance in low-resource settings (Koehn and Knowles 2017).

The datasets utilized in this research can serve valuable pedagogical purposes in training translation students and as foundational resources for interdisciplinary research in fields such as translation studies, cross-linguistic analysis, and lexical semantics. Moreover, they facilitate the refinement of large language models and the adaptation of existing machine translation services. In this study, we leverage these datasets to adapt the base English-Croatian Google NMT model to the domain of religion using Google's AutoML Translation service.

The organization of the paper is as follows: Section 2 briefly presents related work. Section 3 describes the training procedure and provides details about the two aligned parallel corpora used in this process. The results and discussion follow, which are then summarized in the concluding remarks presented in the final section of the paper.

2. Related Work

Carlson et al. (2018) utilized various versions of the Bible aligned by chapter and verse numbers to create a corpus for the style transfer task. Style transfer can be viewed as a form of monolingual translation, akin to a machine translation prob-

lem where the source and target languages differ only in terms of style. The authors incorporated thirty-four stylistically distinct Bible versions, including the archaic language of the King James Version, which dates back centuries. The study involved training and evaluating both an encoder-decoder recurrent neural network and an SMT system. The neural system outperformed the statistical approach when more substantial changes are required to the source segment.

Viswanathan et al. (2019) used Google's AutoML Translate to train a system aimed at producing more consistent translations concerning register, specifically tone and style, while still harnessing the capabilities of a general-purpose MT system. The authors specifically focused on the register associated with personal pronouns. The task can be viewed as a special case of domain adaptation. To accomplish this, they employed formality-specific datasets to train custom models that are strongly biased towards the respective registers. They repeated the training procedure multiple times on the same training dataset, replacing the model with the one obtained from the previous iteration, using Google's generic NMT model as the base model. The results indicate that fewer than 5000 sentences may be sufficient to leverage transfer learning effectively from the base model.

Ruopp (2020) utilized Google AutoML Translation to train custom NMT engine adapted to COVID-19. The study acknowledges the importance of the translation memory format in the era of adaptive, document-context aware NMT systems for preserving document context. Significant improvements in BLEU scores were achieved. However, the author also highlights the risks associated with domain adaptation for high resource language pairs, as adapting to one domain can lead to a deterioration in quality even for closely related domains. Furthermore, combining training data from different sources may blend translator-specific or organization-specific preferences embedded in the training data due to quality assurance procedures, potentially resulting in contradictions.

AutoML Translation is also employed to customize Google Translate for three different genres: song lyrics, novels, and subtitles. Higher BLEU scores are reported in all three cases, with the most significant increase in BLEU observed for subtitles (Al-Sabbagh 2024).

3. Research Study

The primary focus of this paper is to explore how Google's AutoML Translation service enables translators, including those with limited IT expertise, to create improved Neural Machine Translation (NMT) systems using their own text data. Additionally, this study aims to analyse the impact on the workflow within the translation industry, considering factors such as the BLEU score and subjective perceptions of translation quality. As this study focuses on translating religious texts, the initial step involved acquiring aligned parallel corpora of texts within the religious domain. Descriptions of the dataset are provided in the following subsection.

3.1. Data Description

The primary focus of this research centres around the translation of religious texts, necessitating the initial step of acquiring a well-matched parallel corpus of texts within this domain. The corpora used in the training procedure include selected texts by William Branham and their respective translations, as well as the King James Version of the Bible and its translation by Ivan Vrtarić from 2016.

The first corpus consists of texts in English authored by William Branham and their translations. Each text was translated by a single translator, although contributions were made by several different translators. Two versions of the dataset are utilized—one aligned based on paragraphs (BranhamTextsPars) and another based on sentences (BranhamTextsSents). Two randomly extracted excerpts are provided in Tables 1 and 2, respectively, offering insights into the translation alignment process and the structure of the dataset.

Table 1. Selected excerpts from the paragraph-level aligned corpus of William Branham texts (BT_{Pars})

Gracious Lord, we bring to Thee these little parcels, perhaps some of them look to be maybe little vests for a baby, or--or some little undershirt, or maybe a little pair of booties, or--or something, a handkerchief, that's going to the sick and the afflicted. Lord, it is according to Thy Word that we do this. For we read, in the Book of Acts, that they taken from the body of Your servant, Paul, handkerchiefs and aprons, because they believed that Your Spirit was on the man. And unclean spirits went out of people, and afflictions and diseases left them, because they believed. And now we realize, Lord, that we're not Saint Paul, but we know that You still remain Jesus. And we pray that You'll honor the faith of these people.	Dragi Gospode, donosimo ti ove male komadiće materijala, možda neki sliči na mala odijelca za bebu, ili - ili neku malu potkošulju ili možda mali par cipelica, ili - ili tako nešto, maramice koje će ići k bolesnima i napaćenima. Gospode, to činimo prema tvojoj Riječi, jer čitamo u Djelima Apostolskim da su uzimali maramice i ubruse s tijela tvoga sluge Pavla, zato što su vjerovali da je tvoj Duh bio na tom čovjeku. I nečisti duhovi su izlazili iz ljudi, a bolesti i muke su ih napuštale, zato što su vjerovali. A sada mi shvaćamo Gospode da mi nismo sveti Pavao, ali znamo da si ti i dalje Isus. I molimo da ti poštuješ vjeru ovih ljudi.
Not at the first day: dead form; second day there was a rumor (See?)--or the second day it was still dead: Luther, Wesley. At the beginning of the third day there was a rumor around. Nothing on the first day (Luther), nothing on the second day, and in the third day (the dispensation of the Holy Spirit) there was a rumor around that He was the same yesterday, today, and forever. But at the end of the third day, that's where He made Hisself known, come right among them, come among His people and said, «Look at Me; I'm the same One.»	Ne prvog dana, tada je bila mrtva forma; drugog dana bila je glasina - drugog dana još je uvijek bilo mrtvo - Luther, Wesley. Na početku trećeg dana pojavila se glasina. Ništa se nije dogodilo prvog dana (Luther), ništa se nije dogodilo ni drugog dana, a trećeg dana (etapa Duha Svetog) se pojavila glasina da je On isti jučer, danas i zauvijek će biti isti. Ali na kraju trećeg dana Se On obznanio. Došao je među njih, među Svoje ljude i rekao: «Pogledajte Me, Isti Sam.»

Table 2. Selected excerpts from the sentence-level aligned corpus of William Branham texts (BT_{Sents})

All down through the ages they received the Holy Spirit, but not in the measure that they have It now; 'cause it's a restoration of the first.	Skroz kroz doba oni su primali Duha Svetoga, ali ne u mjeri u kojoj Ga sada imaju jer je obnova prvog.
Like it must've been in our Lord when He looked over Jerusalem, His own people (See?), said, «Jerusalem, Jerusalem, how oft would I have hovered you as a hen would her brood, but you would not.»	Kao što mora da je bilo u našem Gospodu kada je gledao na Jeruzalem, Svoj vlastiti narod (Razumijete?), da je rekao: «Jeruzaleme, Jeruzaleme, koliko sam se puta htio nadviti nad vama, kao što bi kvočka nad svojim pilićima, ali ne htjedoste.»

The King James Version (KJV), also referred to as the King James Bible (KJB) or the Authorized Version (AV), is an Early Modern English translation of the Christian Bible, commissioned by King James VI and I and published in 1611. This translation holds immense cultural significance within English literature and language, shaping literary expression for centuries. Notably, the KJV is in the public domain, allowing for widespread distribution and utilization. The KJV has been automatically aligned using chapter and verse numbers. This method of alignment bypasses the pitfalls of imperfect text alignment generated by standard algorithms, which can introduce errors into the translation pipeline, ultimately compromising translation quality (Bibleverse). Furthermore, the decision to use the KJV as a foundational text was informed by its exclusive citation within the writings of William Branham. This deliberate choice underscores the importance of linguistic and contextual consistency in the study and translation of religious texts. Two randomly extracted excerpts from this aligned corpus are provided in Table 3.

Table 3. Selected excerpts from the parallel corpus of the Bible (Bib_{ver})

Who art thou that judgest another man's servant? To his own master he standeth or falleth. Yea, he shall be holden up: for God is able to make him stand.	Tko si ti da sudiš tuđega slugu? Svome gospodaru on stoji ili pada. A stajat će, jer je moćan Bog održati ga.
The Lord GOD hath sworn by his holiness, that, lo, the days shall come upon you, that he will take you away with hooks, and your posterity with fishhooks.	Gospodin BOG se zakleo svojom svetošću da će na vas, evo, doći dani kada će vas odvlačiti kukama, a potomstvo udicama.

Both procedures were straightforward for distinct reasons: the first due to translations being initially generated using appropriate Computer-Assisted Translation (CAT) tools, and the second owing to the sequential numbering of verses within the text. The descriptions of the corpora are provided in Table 4.

In addition to these efforts, a new corpus was compiled by merging the two existing corpora, taking into consideration both levels of alignment. This resulted in

the creation of the MixSent corpus for the sentence-aligned William Branham corpus and the MixPar corpus for the paragraph-aligned William Branham corpus, as described in Table 5.

A notable observation from the compiled corpora is that the number of sentences is slightly higher on the Croatian side. Conversely, all other statistics listed in Table 4 demonstrate an average increase of approximately 16% on the English side of the corpora. Furthermore, expectedly, the Croatian side of the corpora exhibits a noteworthy disparity, featuring between 50-67% more word types and between 20-35% more lemmas compared to the English side, as illustrated in Table 5. While acknowledging the possibility of errors in the automatic lemmatization process, these differences highlight the linguistic nuances and complexities inherent in the language under study.

Table 4. Corpora statistics

	# of pairs	Tokens		Words		# of sentences		Avg segment length	
		En	*Cro*	*En*	*Cro*	*En*	*Cro*	*En*	*Cro*
BibVer	24.996	947.265	776.324	790.943	620.154	35.074	37.234	23	17
BTSents	63.728	1.312.984	1.146.871	1.046.581	925.172	84.049	84.549	12	11
BTPars	36.151								

Table 5. Corpora lexicon sizes

Corpus	# of pairs	Word types		Token-type ratio		Lemmas	
		En	*Cro*	*En*	*Cro*	*En*	*Cro*
BibleVer	24.996	14.160	43.154	67	18	10.318	15.898
BTSents	63.728	24.389	50.076	54	23	15.001	18.742
BTPars	36.151						
MixSent	88724	31.543	74.047	68	25	20.663	26.932
MixPar	61147						

3.2. System and training

In this paper, we employ AutoML Translate,[1] a Google Cloud AI product designed for tailoring NMT engines to specific industries and domains. The AutoML Translation framework uses transfer learning and neural architecture search to develop

1 See https://cloud.google.com/translate/automl/docs for official Google documentation.

new models based on existing NMT models. Notably, it builds upon the Google NMT (GNMT) system, which is a sequence-to-sequence neural machine translation system featuring a deep LSTM network (Chen et al. 2018; Wu et al. 2016) as the baseline model.

This framework is particularly adept at constructing domain-specific customized models using input datasets from the target domain. It excels in not only adapting to specific industry requirements but also in its ability to generalize effectively across various tasks. By leveraging transfer learning and advanced neural network architectures, such systems offer a robust solution for developing and deploying tailored NMT models that address the unique linguistic challenges and nuances present in specific domains.

The dataset is divided into training, development, and test sets. If the dataset contains fewer than 100,000 sentence pairs, AutoML Translation automatically allocates 80% of the dataset for training, 10% for validation, and 10% for testing. The training set is the data the model "sees" during training and is used to learn the parameters of the model, namely the weights of the connections between nodes of the neural network. The validation set, also known as the "dev" set, is utilized to select the best model generated (by evaluating performance on the validation set) and to adjust the model's hyperparameters accordingly. Employing a separate dataset for fine-tuning, the model structure enhances the model's ability to generalize effectively beyond the training data. The performance exhibited by the model on the test set provides valuable insight into its expected performance on real-world data.

3.3. Evaluation

For the evaluation of models, we utilize a well-established metric called BLEU (Doddington 2002). BLEU is a precision-based metric that quantifies lexical similarity on a 0-1 scale, where 0 represents the lowest score. BLEU compares translation n-grams with n-grams from a reference translation and counts the number of matches at the sentence level, but this count is clipped to the maximum n-gram count found in the reference. These sentence counts are then aggregated over the entire test set. The matches are independent of word position within the sentence. Adequacy is reflected in word precision, while fluency is reflected in n-gram precision. Translations that are significantly shorter than the reference are penalized using a brevity penalty with an exponential decay. It is advisable not to compare BLEU scores across different corpora and languages. However, Table 6 provides an interpretation of BLEU scores that can serve as a rough guideline.

Table 6. BLEU score interpretation[2]

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

4. Results and Discussion

The evaluation results are presented in Table 7. The base NMT model achieved BLEU scores ranging from twenty to thirty-three. According to the interpretation provided in Table 6, the base model produced translations ranging from understandable to good for William Branham texts but exhibited significant grammatical errors when translating the Bible and the combined test set.

After adapting the base model, the resulting BLEU scores again fell into two categories. Specifically, translations of the Bible were generally categorized as understandable to good, whereas the translations of the other four cases could be considered high-quality.

The results indicate that the NMT base model performed poorest when translating the Bible. Consequently, adapting the model using the Bible training data resulted in the most significant BLEU score improvement. Furthermore, the NMT base model demonstrated better performance in translating pure William Branham texts compared to the combination of Bible and William Branham texts. The least improvement in BLEU score was observed when translating William Branham texts only (+10.47). The remaining three BLEU score improvements were approximately equal, suggesting that combining training data resulted in similar performance gains. However, the highest translation quality was achieved when training solely on William Branham texts with sentence-level alignment (45.1 BLEU).

Additionally, in the merged corpora scenarios, the level of alignment seemed to have less impact compared to homogeneous scenarios, where more finely-grained alignments had a more pronounced effect on quality improvement.

2 Google Cloud. 2024. *Evaluating models*. Available at: https://cloud.google.com/translate/automl/docs/evaluate

Table 7. Performance of trained models

Model	Google NMT	Google AutoML	BLEU score gain/loss
ModelVerse	20.34	36.02	**+15.69**
ModelSent	32.84	**45.1**	+12.26
ModelPar	32.86	43.33	+10.47
ModelMixSent	27.74	40.1	+12.36
ModelMixPar	27.73	39.91	+12.18

Although no formal human evaluation procedure has been conducted, informal conversations with several translators revealed that even the lowest-scoring model was deemed useful. However, there was a consensus that the highest-scoring model was indeed the best among them.

5. Conclusion

The importance of domain customization is best seen with the occurrence of different crisis such as COVID-19, which are especially challenging for MT engines.

In this paper, Google's AutoML Translation service is used to adapt the base English-Croatian Google NMT model to the field of religion. The corpora used in the training procedure is the King James Version of the Bible and its translation by Ivan Vrtarić dating back to 2016 and the selected texts of William Branham and their respective translations extracted from the translation memories provided by a group of translators. The Bible parallel corpus is aligned at the level of verses, while the texts of William Branham are aligned at the level of sentences. Using the corpora described above eliminated the need for applying text alignment algorithms and errors that might occur.

Since the provided corpora differ in genre and alignment level, several models are built by running the training procedure on each individual corpus, but also on their combination. William Branham texts were additionally aligned at a higher level, i.e., level of paragraphs. This resulted in three models trained on individual parallel corpora, and two models trained on their combination. The Bible parallel corpus was aligned at a fixed verse level in all instances. The results show that the best model, measured by the BLEU score, is obtained when training on William Branham texts alone aligned at the sentence level. The results also show that the base NMT model achieved the best score on William Branham's texts.

Even though formal human evaluation is not presented in this paper, translators who continued working on translations of religious texts assessed the custom-tailored systems positively and felt that all these systems produce translations which can be used as a starting point for human post-editing. Moreover, the best scoring system proved indeed the best in their translation practice.

References

Al-Sabbagh, Rania. 2024. "ArzEn-MultiGenre: An Aligned Parallel Dataset of Egyptian Arabic Song Lyrics, Novels, and Subtitles, with English Translations." *Data in Brief*. doi: https://doi.org/10.17632/6k97jty9xg.4

Carlson, Keith; Riddell, Allen; Rockmore, Daniel. 2018. "Evaluating Prose Style Transfer with the Bible." *Royal Society Open Science* 5(10). doi: https://doi.org/10.1098/rsos.171920

Chen, Mia Xu; Firat, Orhan; Bapna, Ankur; Johnson, Melvin; Macherey, Wolfgang; Foster, George; Jones, Llion; Parmar, Niki; Schuster, M.; Chen, Zhifeng; Wu, Yonghui; Hughes, Macduff. 2018. "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation.". doi: http://arxiv.org/abs/1804.09849

Doddington, George. 2002. "Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics." In *Proceedings of the Second International Conference on Human Language Technology Research, HLT '02*. 138–45. doi: https://doi.org/10.3115/1289189.1289273

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge University Press.

Koehn, Philipp; Knowles, Rebecca. 2017. "Six Challenges for Neural Machine Translation." doi: http://www.statmt.org/wmt17

Ruopp, Achim. 2020. "Using Contemporary US Government Data to Train Custom MT for COVID-19." In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*. doi: https://www.cdc.gov

Viswanathan, Aditi; Wang, Varden; Kononova, Antonina. 2019. "Controlling Formality and Style of Machine Translation Output Using AutoML." *Annual International Symposium on Information Management and Big Data*. https://cloud.google.com/translate/automl/docs

Wu, Yonghui; Schuster, Mike; Chen, Zhifeng; Quoc, V. Le, Norouzi, Mohammad; Macherey, Wolfgang; Krikun, Maxim; Cao, Yuan; Gao, Quin; Macherey, Klaus; Klinger, Jeff; Shah, Apurva; Johnson, Melvin; Liu, Xiaobing; Kaiser, Łukas; Gouws, Stephan; Yoshikiyo, Kato; Kudo, Taku; Kazawa, Hideto; Stevens, Keith; Kurian, George; Patil, Nishant; Wang, Wei; Young, Cliff; Smith, Jason; Riesa, Jason; Rudnick, Alex; Vinyals, Oriol; Corrado, Greg; Hughes, Macduff; Deanet, Jeffrey. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." http://arxiv.org/abs/1609.08144.

Nikolina Palašić, Klaudia Križanec

Faculty of Humanities and Social Sciences, University of Rijeka, Croatia
nikolina.palasic@ffri.uniri.hr, klaudia.krizanec@gmail.com

Translating Elements of Culture Using the Example of the Series "Only Fools and Horses"

Abstract

Since language expresses, embodies, and symbolizes cultural reality, difficulties will inevitably occur during translation when transferring culturally marked expressions from one language to another. Through translation, other cultural realities are introduced into the target text, which, over time, inevitably alter the target language as well. The issue of how to translate cultural elements has been addressed by numerous authors, so in the introduction of this paper, we describe and compare some of them and decide to apply Pedersen's taxonomy to the corpus (Pedersen 2011). For the analysis of cultural elements and translation strategies, the series *Only Fools and Horses* was chosen because it is deeply rooted in the culture of the English working class and abounds in culturally specific elements, for which the Croatian translator found some interesting solutions. Twenty episodes were randomly selected for analysis, and each analysed section of the text is accompanied by the corresponding episode and the timestamp in minutes and seconds indicating when the selected subtitle appeared on the screen.

Keywords: translation, element of culture, subtitling, translation strategies

1. Introduction

Every translator is aware that it is almost impossible to create a translation in which the meaning and form coincide symmetrically with the original text and that crisis points in translation are inevitable. One of the reasons for this is the rootedness of culture in language. The concept of culture is not simple to define,[1] but it is generally accepted that it encompasses the common experiences and creations of a human

1 The complexity of trying to define the concept of culture became apparent long before the problem of elements of culture appeared in translatology, namely forty years before that. It became apparent in an overview of the definitions of culture, which was published in 1952 by Kroeber and Kluckhohn. The authors analysed 164 definitions of the term *culture* and nevertheless failed to come to a single conclusion.The complexity of trying to define the concept of culture became apparent long before the problem of elements of culture appeared in translatology, namely forty years before that. It became apparent in an overview of the definitions of culture, which was published in 1952 by Kroeber and Kluckhohn. The authors analysed 164 definitions of the term *culture* and nevertheless failed to come to a single conclusion.

community, while language arises from the need to communicate them with one another. It is not surprising, then, that each language encompasses concepts relating to phenomena specific to the culture from which it emerged. Transferring these concepts to the language of a culture in which these phenomena do not exist or are understood differently poses a problem for the translator.

Translation studies theoreticians began to consider more seriously the importance of culture as a factor influencing the approach to translation in the 1990s after Susan Bassnett and André Lefevere defined the concept of cultural turn (Bassnett and Lefevere 1990).

This work is based on three assumptions: culture always affects meaning in language, elements of culture are inevitable, elements of culture are not insurmountable. The identification and classification of elements of culture and the procedures for their translation have been addressed by many theorists. It is interesting that they have all independently developed very similar theories and classifications. In the practical part of this paper, the taxonomy of Swedish theorist Jan Pedersen is applied in the analysis of the translation. The selected translation consists of subtitles from the British comedy series *Only Fools and Horses,* which was first broadcast in 1981. In addition to the interest in the expressions used in the series and the recognition that many of them are related to culture, the reason for choosing this series is also the combination of genre and media that will have a great influence on the translator's decisions. What is more, it can be expected that elements of culture appear more frequently in a work whose plot focuses on everyday, ordinary life, not too different from the lives of real viewers, than in works with elements of fantasy.

It is an undeniable fact that language and culture are inseparable. These two elements of humanity are created in the same way—gradually evolving, one from the common life experiences of a community and the other from the need to communicate these shared experiences with each other. Language cannot be separated from the community that has created it—it does not arise by itself, nor is it static and unchangeable. In this sense, the translator must have certain cultural competence, that is, their job is not only to translate a language, but also to adequately translate the culture integrated into that language.[2]

Linguist Claire Kramsch (1998: 3) extensively examines the relationship between language and culture[3] and argues that the two concepts are related with regard

2 In this sense, the concept of bicultural competence is also found in literature (cf. e.g., Witte 2007: 12) as a basic prerequisite for adequate translation, whereby the interpreter is advised to take an adequately distanced attitude towards their own culture and consciously learn about another culture. Some authors (e.g., Kupsch-Losereit 2000) speak of intercultural competence as a prerequisite for bridging cultural differences between the source and target language in translation.

3 In the last thirty years, numerous authors have been dealing with cultural elements in the language with special reference to the translation problems that such elements cause, and here, since the reflections on the importance of such elements in the studied literature coincide,

to three aspects. The first is that language expresses cultural reality—people express common experiences in language, that is, facts, ideas, events, as well as one's own views and beliefs that can be communicated because they share them with other people. Then, language embodies cultural reality—besides expressing their experiences through language, people also create them through language. Kramsch states that the way people use the medium of communication creates meanings that are understandable to the group of people who utilize it. The third relationship between language and culture is that language symbolizes cultural reality—language is a symbol of a group of people, allowing its members to recognize each other, while those who do not speak it are excluded from the group.

2. A Brief Overview of the Approach to Translating Culture

In 1990, theorists Susan Bassnett and André Lefevere published a collection of essays entitled *Translation, History and Culture* in which they gathered their opinions on the then state of translation as a science and for the first time defined the cultural turn in translation. The cultural turn implies a shift from previous linguistic approaches to translation that were oriented towards finding equivalence to combining extratextual cultural factors in the production and study of translations. Culture has become a new translation unit. According to Bassnett, one of the first glimpses of the cultural turn is found in the polysystem theory of Even-Zohar, published in 1970, according to which literary works should not be studied separately but as part of a broader literary, social, cultural, and historical framework (cf. Bassnett, 1998: 124). Translated literature can influence domestic literature and culture. Snell-Hornby adds that in such a system, translated works are not just a copy of the original text, but separate texts that have an impact on the target culture (cf. Snell-Hornby 1988: 24). Polysystem theory has driven the view that translatology should be a separate, interdisciplinary scientific field, not just part of linguistics and literature, and that it has a lot in common with cultural studies. Questions began to be asked about textual and extra-textual "shackles" and the norms that the translator adheres to during translation. André Lefevere proposed a theory of "patronage", in which he argues that the literary system, in addition to internal factors such as critics, teachers, and translators themselves, is influenced by external factors—patrons, that is, anyone who has the power to decide whether a literary work should be translated and in what way (cf. Bassnett 1998: XVI).

Even before the cultural turn, questions were raised about the translator's visibility, and the very concept of visibility depends in part on how the translator treats

---

we list only some of them: Carbonelli (1996); Venuti (1995, 1996, 1998, 2013); Gentzler (1998); Katan (2002); Tymoczko (2005); Apter (2009); Bandia (2009); House (2009, 2015); Pedersen (2011); Chen and Huang (2014); Reis and Vermeer (2014); Baldo (2016); Bielsa (2016); Kharina (2018); and others.

elements of culture. Although these are strategies that have been used since ancient times, Lawrence Venuti formulated the concepts of domestication and foreignization in the book *The Translator's Invisibility: A History of Translation* (1995). In domestication, the aim is to make the produced translation fit as much as possible into the target culture, thus losing the elements of the culture of the source text. In this case, the translator is invisible. In foreignization, the translation contains the elements of the culture of the source text, thus retaining the "alienity" of the original and the reader's sense of reading the translation rather than the source text, making the translator visible. The dichotomy of domestication and foreignization is a cultural issue more than a linguistic one and can only be applied to texts that have cultural elements. Venuti argues that there is also a political reason for domestication in translations—according to him, contemporary Anglo-American culture, which is more prone to domestication than foreignization, considers translation a kind of colonization. Friedrich Schleiermacher spoke about the strategies of foreignization and domestication back in 1813, when he proposed that the translator should have two choices: to bring the reader closer to the author or the author to the reader. Venuti, like Schleiermacher, was more prone to foreignization, thinking that domestication was not fair to the original culture because it erases its values (cf. Venuti 1995: 20).

Juliane House cites a dichotomy that closely resembles Venuti's: "obvious, open" (*overt*) and "hidden, invisible" (*covert*) translation. She claims that overt translation is required in texts that are in a specific way related to their original culture and community, such as historical texts or artworks that have gained "timeless" status and can be interesting for a wider audience but are still tied to a culturally specific period. Such translation tries to keep the source text in the transmission intact as much as possible. Overt translation is focused on the text and the author. Unlike it, covert translation is used in texts that are not specific to a culture. This kind of translation is aimed at the reader and is considered a separate text. Functional equivalence is possible only within covert translation. House also talks about the concept of *a cultural filter* that the translator uses in covert translation. The cultural filter applies the characteristics of the target culture to the original text (cf. House 2015: 65-68).

American linguist Eugene Nida formulated the concepts of dynamic and formal equivalence in the book *Toward a Science of Translating* (1964). These two approaches differ significantly in their treatment of elements of culture. Formal equivalence seeks to retain the form and content of the source text, and in it, loyalty to the original is essential. Dynamic equivalence, on the other hand, demands that the translation produces the same effect on the reader as the original would have produced, that is, as Nida and Taber put it: "intelligibility is not to be measured merely in terms of whether the words are understandable and the sentences grammatically constructed, but in terms of the total impact the message has on the one who receives it." (Nida and Taber 1969: 22). Nida primarily dealt with the Bible translation, and his dynamic equivalence research focuses on translating Hebrew terms

from the Bible whose conveyed meanings are very different from the literal ones. The focus of dynamic equivalence is not on the form and style of the original, and critics felt it crossed the line between translation and adaptation. In response, Nida introduces the notion of functional equivalence, which is actually an improved version of dynamic equivalence, and argues that the form of the source text should also be paid attention to because the form itself has meaning. However, consistently conveying the meaning of the source text will always be more important than conveying the form.

The scopos theory was formed by Hans J. Vermeer in 1978. Evidently in the translation of the word *skopos* itself (Greek: *purpose*), Vermeer argues that translation is motivated by purpose and goal. Translation is an action, and each action has a specific goal; the adequacy of translation is valued according to this goal. Each situation is different and requires different translation strategies depending on the purpose, and the target audience is the most important factor that determines this purpose. The translator, as the real recipient of the message of the original text, creates a translation under the conditions of their culture based on their own assumptions about the needs, expectations, and knowledge of the target audience of the translation (cf. Reiss and Vermeer 2014: 85).

Finally, it should be said that translation aimed at making one culture understandable to another inevitably involves a certain degree of violence, especially if it is a culture of the "other", for example when texts from Eastern and other cultures are translated into a Western culture. Certain discourse defaults (institutional, socio-anthropological, etc.) create certain expectations in the audience for which the text is being translated, so the translator must act according to these expectations, which leads to "violent changes" in the source text. In this process, unknown concepts and indigenous practices are transformed into something that is more familiar and closer to the target audience through a translation process, that is, they are assimilated into culturally known forms of concepts and practices. In this sense, we talk about violence against the original text (Dingwaney 1995: 4, 5). However, it is not only about "violence" in the process of translation; a similar type of selection occurs at the initial choice of what will be translated into the target culture. This selection is also related to the already mentioned cultural and discursive expectations of the target audience.

On the other hand, the translation of culture ensures the continuation of that culture in the future as well as its spread in other areas. Although it certainly takes something away, it also significantly contributes to both the original culture and the culture into which it is translated.

A similar approach can be found in the famous Schleiermacher's article "Über die verschiedenen Methoden des Übersetzens", in which he, speaking about the translation of foreign texts into German, points out that the translation into German should not sound like the original German text, but that his reader must feel

this "otherness", so for example, behind the Spanish translation, one must feel Spanish; otherwise, the original text loses a part of its identity.

3. Elements of Culture

Translation always involves elements that pose challenges for the translator. Jan Pedersen calls these elements translation *crisis points,* referring to Lorscher, who distinguishes between non-strategic and strategic translation. In the first case, translation comes to the translator's mind of "automatically", and they easily find an equivalent segment in the target language. Strategic translation refers to segments that require the conscious use of translation strategies and "problem solving" (cf. Pedersen 2011: 41). In addition to wordplay and songs, elements of culture are often crisis points in translation. In translation theory, these elements have been widely discussed—each theorist offering their own concept and definition, which often differ only in scope.

The term *realia* was defined by Bulgarian translators Vlahov and Florin. Realia are lexical units that denote objects, phenomena, and customs that are present in the source language community and absent in the target language and have no equivalent in the target language (cf. Vlahov and Florin 1986/2012, according to Kharina 2018: 12). Vlahov and Florin divided realia into three categories: geography, ethnography, and politics/society, which they classified into various subcategories.

In his research, Pedersen calls problematic elements *extralinguistic cultural references*, defining them as linguistic expressions relating to extralinguistic phenomena (e.g., places, persons, customs, food) that someone who lacks encyclopedic knowledge of the culture from which they emerged will not understand even if they understand the language in question. Although he admits that his definition is similar to that of realia, he refuses to use the term because its literal meaning can create the thought that cultural references exclude fictitious phenomena (for example, fictitious characters that are deeply rooted in culture). As the name *extralinguistic cultural references* suggests, elements of culture must meet and combine the two criteria. Extralingualism means that they refer to real-world phenomena rather than the details of language systems, but this criterion is only useful when the referent in question requires knowledge of culture. Pedersen gives an example, noting that one person can have knowledge of a language without having knowledge of the culture of the language community, and vice versa. For instance, a person who speaks English will easily discern that the word *tree* refers to a tree, a real-world phenomenon. Therefore, the word *tree* is not considered an element of culture when translated from English into Croatian. Conversely, a person who knows the meanings of the words in the term *finishing school* will be able to discern that it is a school, but without encyclopedic knowledge of English culture, they will not have access to the information that the referent is a type of school in which girls from wealthy families learn social skills suited to members of higher social classes. Pedersen lists

personal names as elements of culture that are easiest to identify as such, because in this case, knowledge of the language alone will not provide the reader with any information. Obviously, he notes that a single language can have multiple language communities, and so can multiple cultures—for example, elements of culture in British English may be incomprehensible to speakers of American English or Australian English. Although he claims that elements of culture constitute extralinguistic phenomena, they also include a "grey area" consisting of concepts that could be understood as intralingual elements but are so connected with extralinguistic culture that they cannot be ignored. As an example, he gives formal titles and expressions from slang (cf. Pedersen 2011: 44-48).

J. Franco Aixela, who calls such elements *culture-specific items* (CSI), argues that many authors avoid defining them and that they are determined by collective intuition, leading to excessive arbitrariness as well as the perception that they are static and immutable. He notes that culturally specific elements do not exist for themselves in the source text, but manifest themselves in the transfer to another language, which is why we can only talk about them when comparing two cultures. Franco Aixela gives an example of translating a figurative image of a lamb from the Bible. Due to the significance of the Bible in Western countries, in many European cultures, as well as in Croatia, lamb is associated with innocence and helplessness. If this concept were transferred from one European language to another, it would not be considered culturally specific. However, if it were transmitted into the language of a culture in which the Bible had no influence and the concept of lamb did not assume an association with innocence, then it would be an element of culture (cf. Franco Aixela 1996: 57-59). There is another problem in defining elements of culture. Although they are already defined as cases in which an equivalent phenomenon is completely lacking in one of the two cultures, there are also cases when an extralinguistic phenomenon exists in both cultures, but in a significantly different form or function, or with different connotations. For example, the concept of the education system can be very different in the two cultures observed. Although the phenomenon exists in both cultures, differences in their forms can cause communication problems.

It is also important to consider the variability over time. Elements of culture, especially in the age of globalization, can lose their status as "foreign elements" when they are domesticated in a foreign culture. One such example is the concept of *Halloween*, which somewhat lost its status as a culturally specific element when this once purely American holiday began to be celebrated among children in other countries, including Croatia. Pedersen introduces the concept of *transculturalism* as a parameter that determines how recognizable an element is in the target culture. Guided by the Diagram of R. Leppihalma, elements of culture are divided into three types based on their level of transculturality. The transcultural element is recognizable in both cultures that are compared, so it does not create problems in translation (but it is not recognizable in some other cultures). The monocultural

element is recognizable only in one of the two cultures and will be recognized by most members of the source culture, although not necessarily all. The infracultural element is tied to the original culture, but it is so specific that it will only be known to a small amount of its members. As an example of the monocultural element that eventually became transcultural, Pedersen mentions *Pocahontas*, a member of the Native American tribe whose name was unknown to many people outside North America before the release of the Disney film of the same name in 1995. He mentions a translation of the book from English into Norwegian that contains her name (the book was translated in 1989). The translator translated *Pocahontas* as *a Native American princess* because he assumed it would not be known to the target culture. In 2024, it can be assumed that most people are familiar with the name *Pocahontas*, even though not knowing everything about her life, but at least they can recognize the connotation of the Native American people (cf. Pedersen 2011: 106-109)

As for the division of extralinguistic elements of culture depending on their role in real life, Birgit Nedergaard-Larsen (1993) offers the most detailed classification, which includes geography (meteorology, biology, cultural geography), history (buildings, events, people), society (economy, social structure, politics, social conditions, lifestyle and customs), and culture (religion, education, media, leisure).

Having defined elements of culture, it is possible to discuss ways of translating them. Jan Pedersen developed a detailed taxonomy divided into strategies oriented to the source language and target language (Pedersen 2011): 75). Source-oriented strategies include retention, specification, and direct translation, while target-oriented strategies include generalization, substitution, and omission.[4] Pedersen also mentions another strategy that he did not classify in his taxonomy because he considers it a ready-made solution, and that is the use of *official equivalent*. It involves replacing the original element of culture with a pre-established equivalent, and the translator does not have to think about it extensively. Some of the elements that fall into that category are units of measurement, place names, famous fictional characters, etc. The official equivalent is often the result of direct translation.

Pedersen notes that his use of the word *strategy* is arbitrary. Different theorists use different names in their taxonomies, but Pedersen, whose taxonomy in this paper will be demonstrated in the analysis of the translation, calls problem solving at the "local" level a strategy, while in order to refer to making larger decisions, concerning the whole text, he uses the term methods. This taxonomy[5] is a product of constant improvement and upgrading and was created with subtitling in mind.

4 Due to space constraints, we will not explain each of the strategies in detail here but refer to Pedersen (2011: 75), and we will show on examples in the corpus analysis what these strategies represent in practice.

5 Pedersen's taxonomy is very similar to those of other theorists, such as Vladimir Ivir (1987) and Diaz-Cintas and Remael (2007). These strategies are basically identical, except that they are fragmented and categorized in a different way.

4. Corpus Analysis

4.1. Selected Corpus

The theory discussed in the previous chapters will be applied in the analysis of the Croatian subtitles for the comedy series *Only Fools and Horses* (1981), with an emphasis on Pedersen's translation strategies. *Only Fools and Horses* is a British comedy by John Sullivan, which follows the life and intrigues of two brothers, Del Boy and Rodney Trotter, who live in a poor part of London in the eighties. Having achieved great popularity in the United Kingdom, the comedy was soon imported to other countries, including Croatia, where it was translated by Tomislav Pisak and gained cult status. The reason for choosing this comedy is the fact that it is deeply rooted in the culture of the English working class and abounds in culturally specific elements. The Croatian translator employed interesting solutions in translating the problematic elements. Twenty episodes of the comedy were randomly selected for the analysis of cultural elements and translation strategies. For each analysed text segment, the corresponding episode and the timestamp in minutes and seconds indicating when the selected subtitle appeared on the screen will be provided. The official Croatian translation is taken from the DVD-collection published by *Jutarnji list*.

Translation in question is a subtitle, i.e., a form of audiovisual translation, in which the translation of a dialogue appears as text on the screen. Due to the popularity of the Anglophone television industry in the world, subtitling is one of the most common forms of translation, and Croatia is one of the countries that favours it over dubbing. Subtitling as a transfer from sound to written medium has specific norms and limitations that influence the translator's decisions in the application of translation strategies. For example, due to the need to preserve space and shorten the time it takes the viewer to read the text, the translator will opt for an adequate solution that they might not have chosen when translating another form of text. Other factors influencing translation include genre, publisher rules, and audience needs.

4.2. The Text Approach

Several factors influenced the translator's approach to the text. Purpose, i.e., the scopos of the comedy series is primarily humour, and the translator had to ensure that the translation had the same (or at least very similar) impact on the Croatian audience as the source dialogue had on the British one, thus orienting itself towards dynamic equivalence instead of the formal one. The translator was guided by his own assumptions regarding the Croatian audience's knowledge about British culture (and other cultures), and translating the elements of culture that he assumed would not be understandable, he used a strategy that prioritizes impact rather than the faithful transfer of information. Due to the limitation to only two lines of text, as well as the time interval in which the original dialogue is heard on the screen, the translator was forced to make the necessary impact in as few words

as possible. Due to the nature of subtitling, there is an interesting conflict between the previously established theories. It was concluded that domestication by similarity coincides with the covert type of translation (bringing the text closer to the audience and using the cultural filter). However, subtitling is an overt type of translation because the reader is always aware that they are reading the translation, at the same time having access to the original dialogue. Nevertheless, the translator often tried to bring the original text closer to Croatian cultureto the extent that in many cases he used a substitution, which replaces the original elements of culture with the Croatian ones. Due to this, dissonance can arise between what the viewer reads in the subtitle and what they hear. Domestication, however, is appropriate in the subtitling of humorous content, and for many viewers it can improve the experience. Although the translator often uses strategies oriented towards the target language, he also uses a retention strategy very frequently, which contributes to foreignness in translation.

4.3. Overview of Elements of Culture

In all twenty episodes totalling 578 minutes, a total of 477 elements of culture (EC) were identified, of which 292 originate from the source (English) culture, and 185 from various third cultures, most often American, Irish, or Indian. Oriented by the Nedergaard-Larsen classification, 59 elements fall into the category of *geography*, 37 into *history*, 133 into *society*, and 248 into the category of *culture*.

All terms from the geography category are cities, regions, or streets in England, most often in or near London, since the plot of the series takes place there. It can be assumed that the mention of geographical ECs leaves a completely different impact on someone who has been to that location or is familiar with it than on someone who hears about it for the first time. Of the 37 historical ECs, the most common are historical figures, whose mention plays a role in jokes, catchphrases, and exclamations.

In the corpus, trademarks are often mentioned, which fall into the class of *economy*, a subcategory of *society*. These are brands that in most cases originate from the United Kingdom, although many of them are also recognizable on the Croatian market. In the subcategory *social structure*, headlines from the ranks of the British police dominate, which is also appropriate because the protagonists engage in criminal activities. The majority of ECs from the category of *society*, however, belong to the subcategory *lifestyle and customs*, of which food and drink occupy by far the highest percentage.

The category *culture* includes the most identified ECs. Interestingly, three of the four identified ECs from the category of *religion* derive from Indian culture (Hinduism), while the fourth is a Catholic holiday directly associated with Irish culture. The subcategory *education* includes levels from the British education system as well as the names of British colleges. The subcategory *media* contains the names

of films, series, and shows. Although Nedergaard-Larsen does not include fictitious characters in her classification, they can also be classified within this subcategory. Popular fictional characters from British and American series, films, and comics are often the subject of humorous comparisons with characters from the series. Due to its versatility, the subcategory *leisure* contains the most ECs. Famous personalities such as actors, musicians, and athletes are the subject of many jokes, and their recognition by the Croatian viewer depends on their familiarity with British or American popular culture. In most cases, the translator left their names intact. In addition, the names of restaurants, hotels, and concert halls, predominantly those in London or near London, are also very common.

5. Analysis of Translation Strategies on Examples[6]

5.1. Retention

Retention is one of the most commonly used strategies in the translation in question and is mostly used for names of celebrities, names of geographic locations and car brands, which in most cases remain unchanged and unmarked. Untranslated food and beverage names are indicated in quotation marks or slashes, although the translator is not always consistent in choosing which ECs to mark and which not to. One of the types of cases in which the translator had to use this strategy is when the very name of the EC has an impact on humour. In example (1), Del Boy tries to convince his partner Raquel to perform at a nightclub, assuring her that it would not be her first time:

(1)

Raquel: Del, I've never sung in a real nightclub before!	Raquel: Nisam pjevala u pravom noćnom klubu.
Del Boy: You showed us that poster of when you appeared on the same bill as **Otis Redding** at the Talk of the Town, London.	Del Boy: Jesi! Imaš plakat: Ti i **Otis Redding** u Londonu.
Raquel: It was **Laurie London** at the Talk of the Town, Reading.	Raquel: Ne. Ja i **Laurie London** u **Readingu**! (S7E3, 12:30)

Although the EC *Talk of the Town,* referring to the London concert hall, was omitted, the translator decided to leave the other ECs intact. This was because the similarity between the phrases "Otis Redding in London" and "Laurie London in Reading" caused Del Boy's confusion—he thought that it was a performance with a

6 Of the total of 477 examples of the elements of culture, we have chosen 42 that in our opinion represent everything we wanted to show in this research when it comes to translation strategies and decisions the translator had to make in order to successfully convey the originally intended information and humour.

popular American soul singer, rather than a lesser known English former child star. Nevertheless, the most Croatians probably would not know who they are. Retention in these cases may be inadequate in conveying information but allowed because the information is not the scope of the text. In example (2), Rodney tries to convince a noblewoman to live in a luxurious part of the city, describing it as follows:

(2)

Rodney: It's, er, well, like a little **St John's Wood** you know, just south of the water	Rodney: To je zapravo kao **St. John's Wood** u malome. (S2E7, 12:51)

St. John's Wood is a luxurious district in London, known for many sights such as the cricket club and its association with the famous music group *The Beatles*. Croatian viewers may not be able to recognize the EC, but it is clear from the context what Rodney's intention is, and no further explanation is needed.

Another example of the use of retention is when EC information is visible on the screen. In example (3), Del Boy mentions the Vauxhall Velox, a classic model of a luxury car produced by the British company *Vauxhall Motors*. Even if viewers from the name itself cannot recognize what it is, the screen shows that it is a car, and the translator is not required to provide further clarification.

(3)

Del Boy: listen, look I've got the **Vauxhall Velox** outside haven't I, and...	Del Boy: Vani mi je moj **Vauxhall Velox**. (S1E3, 01:29)

Example (4) demonstrates retention with adaptation to the pronunciation of the Croatian language. *Lady* is the title used in the UK to refer to women of noble birth. Although it literally means *dama* in Croatian, this title is an example of Pedersen's previously mentioned "grey zones" in which the intralingual term is tightly related to culture:

(4)

Lady Ridgemere: I'm **Lady** Ridgemere!	Lady Ridgemere: Ja sam **ledi** Ridgemere! (S2E704:50)

An interesting example of retention with adaptation to the target language occurs in the case of musical groups whose name consists of one plural noun. Although grammatically incorrect, in the Croatian language, a continuation of *-i* is often added to such names to denote the plural, despite the fact that the original noun in English is already in the plural form:

(5)

Del Boy: Raquel and Tony could become the new... the new **Carpenters**. Rodney: Or plumbers or brickers...	Del Boy: Mogli bi biti novi "**Carpentersi**". Rodney: Ili „Bravari"... (S7E3, 24:27)

In the scene we are considering, Del Boy wants to convince Rodney that he can make a new popular musical duo from his girlfriend and the singer he found in a nightclub, saying that they will be as popular as the American music duo *The Carpenters*. Although this group is popular enough to be recognized by music lovers from Croatia, leaving an untranslated title completely prevents viewers who do not speak English from understanding the following joke. *Carpenter* translated to Croatian means "tesar", and Rodney in response stingily calls the hypothetical musical duo plumbers or brickers, although the translator decided to translate this as "Locksmiths" (Bravari).

While retention is the simplest translation strategy, a translator must be very careful in deciding when it is acceptable to use it. In many cases, retention can prevent the target audience from accessing information and the intended impact of the source text, particularly in translations of humorous content.

5.2. Direct translation

Direct translation is most commonly used in names consisting of common nouns, and Pedersen considers this strategy identical to the classical way of literal translation. Calque is a form of direct translation in which no shift in the type of word has occurred, as seen in example (6) in which Rodney compares his musical band to the famous English pop band from Liverpool:

(6)

Rodney: We're styling ourselves on **Frankie Goes to Hollywood**!	Rodney: Uzor nam je "**Frankie ide u Hollywood**!" (S4E4, 04:00)

The translation of the element from the third culture using calque is visible in example (7). The term *silver bird* is a direct translation of the name of the German model of the *Silbervogel* warplane used in World War II. The question in the example is addressed to a woman who travelled to England from Germany, and the term itself can be recognizable only to those who have knowledge of warplanes:

(7)

Albert: You just came in then? Gatwick airport – **silver bird**?	Albert: Sad si doputovala? Sletjela si **srebrnom pticom**? (S5E1, 03:58)

The most common form of direct translation in the text is the translation with a shift in which the proper noun becomes an adjective:

(8)

Rodney: I took her down the **Star of Bengal**.	Rodney: Jeli smo u **Bengalskoj zvijezdi**. (S5E1, 8:01)

(9)

Del Boy: It was, um, it was this – this **Victorian globe**.	Del Boy: **Viktorijanski globus**. (S1E3, 14:37 (10)

(10)

Rodney: I didn't know whether to phone the police of the **Texas Rangers**!	Rodney: Nisam znao hoću li zvati policiju ili **teksaške rendžere**! (S1E3, 05:52)

While Croatian audience is unlikely to recognize that *Star of Bengal* (8) is a restaurant serving Indian food, this is clear from the context of the scene. It can be assumed that most Croatian viewers are familiar to some level with the concept of the Victorian era. Even if they do not know what a globe made in the style of that era looks like (9), that knowledge in the context of this translation is unnecessary because the globe itself is shown on screen. *Texas Rangers* (10) is a type of police service operating in the US state of Texas. In Croatia, the term was popularized by the American television series *Walker, Texas Ranger,* which aired from 1993 to 2001. Direct translation can significantly improve the fluency of the translation, but translators must ensure not to use it in cases where there is already an established translation.

5.3. Specification

Specification is a strategy that is rarely used because it takes up extra space, which is undesirable in subtitling. In example (11), the translator decided to supplement the name of the Ford brand car:

(11)

Del Boy: There is nothing I'd like more than to see you become someone! Nice little **Capri Ghia** and all that!	Del Boy: Najviše od svega želim da postaneš netko. Da voziš lijepi "**Ford Capri Ghia**". (S4E4, 21:42)

In example (12), the complement is used together with a direct translation. The translator first determined that WI stands for *Women's Institute* and performed a direct translation, shifting the noun into an adjective:

(12)

Tony: No, it's the local **WI**. Still, you've got to keep them happy, eh?	Tony: **Ženski institut**. I to se mora. (S7E3, 27:00)

Specification is a strategy that removes ambiguity and reduces the scope of meaning of a term, the opposite of generalization, which increases it.

5.4. Generalization

Generalization is the most used strategy in subtitling, including the analysed translation. The specificity of an element of culture is often not crucial to humour and can be replaced by a more general term that the target audience will understand. Replacement through hyperonym is the most common form of generalization, especially in the case of brands:

(13)

Rodney: What are we gonna use, eh? Superglue and a bottle of **Windolene**, knowing you!	Rodney: A što ćemo mi rabiti? Superljepilo i **deterdžent**? (S2E7, 18:53)

(14)

Del Boy: I eat on the move, mobile phone in one hand, a **Pot Noodle** in the other.	Del Boy: Jedem u pokretu. U jednoj ruci mobitel, u drugoj **tjestenina**. (S7E3, 01:13) (15)

(15)

Del Boy: I mean, on one hand you've just had your hopes and dreams dashed! But on the other hand, I've got a van load of hooky **Maltesers**!	Del Boy: S jedne su ti se strane rasplinuli svi snovi. A s druge, imam furgon pun ukradene **čokolade**! (S4E4, 23:13)

In example (13), *Windolene* is a British brand of window washing detergent that is no longer produced. The specific brand is not at all relevant to the situation and it is possible that most of the audience will not recognize it, so the translator opted for its hyperonym. An identical case can be observed in examples (14) and (15), which contain *Pot Noodle* and *Maltesers*, British brands of pasta and chocolate bars.

Example (16) contains the term from cricket, a sport that is popular in the UK and Commonwealth countries, but has never reached great popularity in Croatia. *Wicket keeper* denotes a player who has the role of goalkeeper in cricket.

(16)

Del Boy: Kuvera was one of India'spremier **wicket-keepers**.	Del Boy: Kuvera je slavni indijski **igrač kriketa**. (S1E3, 09:34)

Geographical terms are also translated by generalization when what is there is more important than their name. In example (17), one of the most significant street markets in London, *Portobello Road Market*, has been replaced by a hyperonym:

(17)

Man: You can get them in **Portobello Road** for seventeen pounds each!	Man: Na **tržnici** se prodaju po 17 funta! (S1E3, 26:25)

Generalizing a personal name is seen in the following example. Esther Rantzen was the presenter of the UK-based television show *That's Life*!, which dealt with a multitude of different topics. The EC *That's Life!* is solved by substitution, which is why it makes no sense to keep the name of the presenter who builds on it in the text.

(18)

Del Boy: Here, couldn't you write to That's Life? Ram: If Lord Krishna himself couldn't help us I really don't think **Esther Rantzen** would stand much chance!	Del Boy: A da pišeš u Potrazi? Ram: Ako nam Krišna nije mogao pomoći bojim se da onda neće ni **televizija**. (S1E3, 08:38)

The following examples contain a comparison of generalizing the same term in two different ways. *Yuppie* ("young urban professional") is a derogative term used in the 1980s to refer to young people who have well-paid jobs in urban areas. *Yuppies* are entrepreneurs, and the stereotype is that they have a "dandy" dressing style, so if necessary and according to the context, it is possible to choose which connotation to use. In example (19), generalization is well used because Del Boy describes his supposedly entrepreneurial lifestyle, but in example (20), in which his friend reproaches him for being summoned to court, saying that it is not good for his *yuppie* image, it would be more appropriate to use an adjective related to professionalism.

(19)

Del Boy: I'm out there on that **yuppie** tight rope, nerves on red alert.	Del Boy: Hodam na **poduzetničkoj** žici. Živci napeto rade. (S7E3, 01:07)

(20)

Boycie: Del Boy! I hear you're in court tomorrow. Don't do a lot for your **yuppie** image, does it?	Boycie: Nije baš **šminkerski**. (S7E3, 31:31)

Paraphrasing is a form of generalization in which instead of the original EC, the translator conveys its meaning or connotations in their words and is used when the problem is too complex to be solved by the use of hyperonym. In example (21), Del Boy tries to convince Rodney that popular music is not just the one that appears on the charts. He mentions *Top of the Pops*, a music show that airs weekly in the United Kingdom and shows performances by musicians who are on the current UK charts.

(21)

Del Boy: I mean, you take that John Denver and Roger Whitaker, they never appear on **Top of the Pops** do they, but they still sell millions of records.	John Denver I Roger Whitaker, nikad nisu na **top-listi**, ali prodaju milijune ploča. (S7E3, 23:55)

In example (22), Del Boy describes the poor condition of the hotel in which he is staying, saying that it cannot be compared to the *Ritz*, a famous five-star London hotel, considered one of the most prestigious hotels in the world. The translator decided to paraphrase the connotative meaning:

(22)

Del Boy: I mean, take a look at this place, it's hardly **the Ritz** is it, eh?	Del Boy: Ovo nije baš **hotel s 5 zvjezdica**. (S1E3, 13:42)

In example (23), Rodney explains to Del Boy that he searched for him for a long time at all Indian restaurants across a large area between two locations in south London. In order to preserve the place and unnecessary naming of the ECs, which would only burden Croatian viewers, the translator opted for a simple paraphrase:

(23)

Rodney: I've been crashing through the doors of every ccurry house and take-away from **Battersea Bridge to Colliers Wood tube station**!	Rodney: Upao sam u svaki azijski restoran u **južnom Londonu**! (S1E3, 06:08)

An interesting case of generalizing is seen in example (24) in which *ruby*, a term from cockney jargon denoting *curry* or any other spicy Indian dish, is mentioned. Cockney is spoken among the working class of East London. The term was created by rhyming the word *curry* with the surname of British singer Ruby Murray:

(24)

Del Boy: D'you think a **ruby** was wise in her condition?	Del Boy: Misliš li kako je bilo pametno da jede tako **začinjenu hranu**? (S5E1, 08:08)

Generalization is a strategy that can applied to ECs of all kinds. In most cases, these are monocultural ECs that are replaced by their transcultural hyperonyms. As a target text-oriented strategy, it significantly reduces the characteristic of foreignness in the original text.

5.5. Substitution

Substitution is the most difficult translation strategy for a translator because it requires not only their adequate knowledge of the original language and culture, but also their deep knowledge of the target culture, as well as sufficient creativity that will allow them to perform a replacement that will have the same impact in translation as the original. Since the scopos of the comedy series is not exclusively the transfer of information about the original culture but humour, the translator can allow themselves some level of "infidelity" to the source text.

In cultural substitution, one element of culture is replaced by another based on similarities or roles they play in their own cultures. Example (25) contains two monocultural ECs of the original language replaced by monocultural ECs of the target language:

(25)

Del Boy: He's got two **O-Levels** and he thinks he's **Bamber Cascoigne's** vest!	Del Boy: Ima **malu maturu** i misli da je kralj **Kviskoteke**! (S1E3, 09:24)

In the UK secondary education system, O-Level (*Ordinary Level*) is a qualification that students aged fourteen to sixteen receive by passing the individual subject exam. As the lowest level of educational qualifications in the United Kingdom, the translator, based on this similarity, replaced it with "mala matura", which is taken in Croatia by pupils who have completed lower secondary school. Bamber Gascoigne was the host of the popular British television quiz *University Challenge*, which is similar to the Croatian *Kviskoteka*, which aired in Croatia from 1980 to 1995. In this situation, it would be more correct to replace the host of the British quiz with the host of the Croatian quiz, but the translator opted for the term "kralj *Kviskoteke*" (the king of *Kviskoteka*), which possibly achieves a similar effect, but it should also be said that the translator's strategy in this example compromises veridicity. In the observed scene, Del Boy teases his brother for pretending to know who the Hindu god Kuver is, saying that by achieving some level of education, he considers himself someone who would know the answers to all the questions on a quiz.

In example (26), in which the characters discuss the baselessness of their fear of being in a place where a serial killer supposedly wanders, the translator decided to replace one transcultural element with another. *Ghoul*, a monster from the legend, the one that eats people, is not known in Croatian culture, nor is there a Croatian name for it (although it is often translated as "zloduh", which is also not quite the same). Instead, the translator wrote "vukodlak" (werewolf), another type of monster, which Croatian viewers will recognize, and the effect remained the same:

(26)

Del Boy: Here you are Rodney. See what I mean, there ain't no ghosties or **ghoulies** out here!	Del Boy: Vidiš? Nema ni duhova ni **vukodlaka**! (S3E3, 11:34)

In examples (27), (28), and (29), the characters play the popular board game *Monopoly,* whose board features street names, usually including the capital of the state in which it is localized. The characters mention the streets in London where their figurine is located. The translator decided to replace them with geographical elements of Croatian language culture, albeit recklessly because he replaced the street names with the names of the city, mountain, and island:

(27)

Del Boy: **Park lane**. I think that's one of my properties Rodney.

Del Boy: **Dubrovnik!** Mislim da je to moje. (S3E3, 14:25) (28)

(28)

Rodney: No, I don't, no I don't. Look, you've got **Coventry Street**. Grandad's got the Waterworks and all that. Ah, yeah, Park Lane, with one hotel,vtwo thousand please.

Rodney: Nisam! Vidi: imaš **Velebit**, a djed ima „fast food".vTu imam hotel... Duguješ mi 2000 funta.
(S3E3, 14:32)

(29)

Rodney: **Ah, Piccadilly**. Right, that's mine and I've got a hotel, so that's twelve hundred pounds!

Rodney: **Korčula!** To je moje i imam hotel! Plati 1200!
(S3E3, 14:59)

When using substitution as a translation strategy, there can often be a dissonance between what viewers see and read. In a situation like this, it might be unusual to read that characters living in England play *Monopoly* with geographical concepts of Croatia, which, as already mentioned, compromises veridicity of the series.

A skilfully executed substitution in which the translator managed to keep the original joke intact is seen in example (30). Trying to convince the noblewoman that he was a car connoisseur, Del Boy tells her that he drove for *the John Player Special*, referring to the Formula 1 racing team *Lotus,* which was sponsored by the British cigarette brand *John Player & Sons*. When asked if it was Formula 1, Del Boy's grandfather replied that he was actually driving as a cigarette delivery man. As this specific brand of cigarettes was never sold in Croatia, the translator replaced the EC with the Marlboro cigarette brand, which is well known in Croatia, and was also a sponsor of numerous Formula 1 racing teams:

(30)

Del Boy: I used to drive for the **John Player Special** team!
Lady Ridgemere: Oh, the Grand Prix circuit?
Grandad: No, delivering fags round Lewisham.

Del Boy: Vozio sam za „**Marlboro**"! Lady Ridgemere: Formule jedan?
Djed: Ne. Dostavljao je cigarete u Lewishamu.
(S2E7, 04:29 – 04:33)

The case of substituting one EC from a third culture with another EC from the same culture is shown in the following example. *Pfennig* is the name of money used in Germany before the introduction of the euro in 2002. The translator replaced it with the *Bundesbank*, the central bank of Germany:

(31)

Albert: Best of luck darling, keep yer hand on yer **pfennig**!	Albert: Bog, zlato! I pazi na svoju **Bundesbanku**! (S5E1, 04:56)

The extralinguistic phenomena that exist in many cultures can be expressed by expressions rooted in culture. The informal English term for police *Old Bill*, whose etymology is unclear, is substituted by the term "žbiri", which generally refers to informants and spies:

(32)

Del Boy: What do you think you're playing at, inviting the bloody **Old Bill** round here?	Del Boy: Kog jarca radiš?! Što si zvao **žbire**? (S4E4, 19:04)

We can see situational substitution in example (33) in which the translator completely changed the meaning of the sentence. In the scene, Del Boy is in a cabin with a serial killer waiting for the arrival of the police. Hearing the sound of a helicopter from the outside, the killer asks him if it is the police, and Del Boy tries to lie to him saying that it is not. *Barratt Developments* is a British home construction company known for its advertising campaigns using helicopters. Since Croatian viewers would not be familiar with this, the translator has completely changed the meaning so that it fits into the situation anyway—mosquito dusting may seem like a helicopter, appropriate for the situation in question:

(33)

Del Boy: No, you're alright. It's **Barratts**!	Del Boy: Nije. **Zaprašuju komarce**! (S3E3, 26:38)

Example (34) also contains situational substitution. After visiting local nobles, Rodney reproaches Del Boy for wanting to become part of their society:

(34)

Rodney: He can't wait to get a shotgun and a retriever and go marching across the grouse moors all done up like a **ploughman's lunch**, can he?	Rodney: Jedva čeka da nabavi sačmaricu i retrivera i da krene u lov na fazane obučen k'o **reklama za paštetu**. (S2E7, 19:41)

Ploughman's lunch is an English cold dish consisting of bread, cheese, onions, ham, eggs, and salad. In the situation in the series, Rodney says that the noble way of dressing resembles the dish in question, and the translator assumed that most Croats would not know what the dish looked like and replaced it with the phrase "reklama za paštetu" (pâté advertisement), which evokes a similar mental picture of breakfast as in the original EC.

Situational substitution as a solution to the translation of the wordplay in which the EC is located is shown by the following example. *Essoldo Kilburn*, primarily called *The Kilburn Empire Music Hall*, is a concert hall in London. Due to the unfamiliarity of this EC to the Croatian viewers, the translator had to come up with another idea to build on the word "empire":

(35)

Ram: You see, our families have been engaged in a vendetta for many, many years. It goes back to the days of the Old Empire. Rodney: He means the British Empire, not the **Kilburn**!	Ram: Naše su obitelji dugo upletene u ogorčen sukob. Još od dana starog Carstva. Rodney: Britanskog, ne **životinjskog**! (S1E3, 07:51)

In the following example, the names of institutions are approached through situational substitution. *Rampton* and *Broadmoor* are strictly protected psychiatric institutions in England. The translator translated one as a correctional facility and the other as a prison, which is not true, but it can fit into the situation in the text that talks about a young delinquent.

(36)

Mickey: I've never been to **Rampton**! I've been to **Broadmoor**, once or twice, but that's not the point.	Mickey: Nikad nisam bio u **popravnom**! Bio sam u **zatvoru**, al' to nema veze. (S4E4, 11:09)

Substitution is arguably the most interesting translation strategy that, in the hands of a creative translator, can elevate humorous content for the target audience.

5.6. Omission

Omission is used when a translator determines that an element is unnecessary in conveying a message. When subtitled, the reason is primarily to preserve the space and keep the subtitle as short as possible, so that it is easier to read it in a short time. In example (37), which contains two cases of detention, the geographical EC (Nine Elms) is omitted:

(37)

Del Boy: But me, I'm one of them that's accepted anywhere – whether it's drinking lager with the market boys down at **Nine Elms**, or sipping Pimm's fruit cup at Hendon regatta!	Del Boy: (...) kad pijem pivo s dečkima s tržnice ili kad pijuckam "Pimms" na Hendonskoj regati! (S2E707:53)

The translator concluded that keeping this EC in this situation is not necessary, as it will not contribute anything to the Croatian audience. However, this is also

questionable for the two retained ECs in the example, in which Del Boy tries to emphasize that he is accepted in "high" society as much as he is in "ordinary" society. *Nine Elms* is an industrial region in London, and *Pimm's* is a type of fruit cocktail, the direct opposite of beer. It is possible that the term "Hendonska regata" (Hendon Regatta) is the result of confusion with the royal regatta at Henley.

In the following example, the phrase "up the wooden hill to Bedfordshire"[7] is completely omitted, the meaning of which is to go to sleep (to a room upstairs):

(38)

Del Boy: Alright then, well, why don't you go **up the wooden hill to Bedfordshire** and check it out?[7]	Del Boy: Idi ti gore i provjeri! (S3E3, 12:59)

The omission is clearly visible in the difference in the length of the original text and the translation in example (39):

(39)

Rodney: I can now leap out of the **Vauxhall Velox**, **Dukes of Hazzard** fashion, make a **chapati** and say get stuffed in Urdu!	Rodney: Sad znam sve pozdrave i psovke na urdskom! (S1E3, 06:12)

Vauxhall Velox has already been mentioned in the paper, so it does not require explanation. *Dukes of Hazzard* is an American comedy series, first aired in 1979, known for scenes in which a car flies into the air. *Chapati* is an Indian type of unleavened bread. It is evident that only one-third of the message from the original text was transmitted in the translation.

Regarding the acceptability of omission as a translation strategy, Pedersen quotes Leppihalme: "the translator may choose to use omission responsibly, having rejected all alternative strategies, or irresponsibly, in order to avoid seeking information about something unknown to him" (Leppihalme 1994: 93, cited in Pedersen 2011: 96) .[8] It would not be fair to say that in these cases the translator chose omission because of laziness or ignorance. When subtitling, sacrificing certain parts of the text is inevitable due to adaptation to the medium, especially in cases of quick speech and very complex ECs, the subtitling of which would leave a large amount of text on the screen in a very short time.

7 Literal translation: "po drvenom brežuljku (stepenice) u Bedfordshire ('grad kreveta')".

8 "A translator may choose omission responsibly, after rejecting all alternative strategies, or irresponsibly, to save him/herself the trouble of looking up something s/he does not know."

5.7. The official equivalent

This strategy involves reaching for an already existing solution, which has already been decided by some authority. This is often the case when mentioning films or literary works that have already been translated for the Croatian market.

(40)

Del Boy: Alright then, who have you seen **Hawkeye**?	Del Boy: Koga si vidio, **Oko Sokolovo?** (S3E3, 10:31)

(41)

Del Boy: As Macbeth said to Hamlet in **A Midsummer Night's Dream**, 'We've been done up like a couple of kippers.'	Del Boy: Kao što je Hamlet rekao Macbethu u **Snu Ivanjske noći**: ispali smo pravi mulci. (S1E3, 27:57)

Oko Sokolovo (40) is a hero from Marvel comics that are sold all over the world. *San Ivanjske noći* (41) is a classic comedy by one of the most famous writers in the world, William Shakespeare, so it hasbeen translated into Croatian as well. Due to the very fact that official equivalents exist, it can be said that these ECs are transcultural.

Units of measurement are also almost always transmitted by official translation. As the Imperial System of Measures is used in the United Kingdom, it is customary for the quantities expressed in this system to be converted into the metric system used in Croatia:

(42)

Lady Ridgemere: I'm trying to get to Ridgemere Hall, it's that large estate about **five miles** back up the road.	Lady Ridgemere: Idem u Ridgemere Hall. To je posjed udaljen **8 kilometara**. (S2E7, 04:41)

The official equivalent may arise using any of the strategies mentioned above (except omission).

6. Conclusion

The frequency of the elements of culture is expected, but still staggering. A closer consideration of everything that the elements of culture actually encompass makes it even clearer how inevitable they are in the text. The translator has used all the above translation strategies, but one can discern his fondness for generalizing and substitution, that is, strategies oriented towards the target text that reduce the foreignness in the source text and bring it closer to the target audience. This can also be seen in the fact that, reading only the translation, many elements of the culture of the original language remain unnoticed. This was also expected—since the genre was a comedy, the translator had to ensure that its primary goal, humour, was available to the target audience, even if it meant losing the cultural specificity of

the original. The very popularity of the series *Only Fools and Horses* in Croatia proves that he was successful in the transmission of humour. Nevertheless, it is clear that a certain percentage of humour has been lost in translation, precisely because of the impossibility of transmitting the elements of culture. It must also be noted that many elements of culture in the original may not be recognizable to members of the original culture as well, which makes their transfer to the target culture even more difficult.

From the analysis of the translations, it is clear that the very choice of translation strategies does not depend strictly on the category of the element of culture being translated, but on the context of the expression. Pedersen's taxonomy of translation strategies has proven to be a useful tool for evaluating translations, as well as for exploring ways of translating elements of culture that can help with future translations. However, it must be pointed out that the procedures used in the translation are not strictly defined by the above strategies and can arise from their combination. A translator can use previously defined strategies, approaches, aids, and norms in creating translations, but in the end their mind is their most important tool, which is why two translators will never produce exactly the same translation of the same text.

7. List of episodes

1. "Go West Young Man." *Only Fools and Horses*, created by John Sullivan, season 1, episode 2, BBC, 1981.
2. "Cash and Curry." *Only Fools and Horses*, created by John Sullivan, season 1, episode 3, BBC, 1981.
3. "The Second Time Around." *Only Fools and Horses*, created by John Sullivan, season 1, episode 4, BBC, 1981.
4. "The Russians Are Coming." *Only Fools and Horses*, created by John Sullivan, season 1, episode 6, BBC, 1981.
5. "The Long Legs of the Law." *Only Fools and Horses*, created by John Sullivan, season 2, episode 1, BBC, 1981.
6. "No Greater Love." *Only Fools and Horses*, created by John Sullivan, season 2, episode 4, BBC, 1981.
7. "A Touch of Glass." *Only Fools and Horses*, created by John Sullivan, season 2, episode 7, BBC, 1982.
8. "Friday the 14th." Only Fools and Horses, created by John Sullivan, season 3, episode 3, BBC, 1983.
9. "Yesterday Never Comes." *Only Fools and Horses*, created by John Sullivan, season 3, episode 4, BBC, 1983.

10. "Wanted." *Only Fools and Horses*, created by John Sullivan, season 3, episode 6, BBC, 1983.
11. "Who's a Pretty Boy?" *Only Fools and Horses,* created by John Sullivan, season 3, episode 7, BBC, 1983.
12. "Strained Relations." *Only Fools and Horses*, created by John Sullivan, season 4, episode 2, BBC, 1985.
13. "It's Only Rock and Roll." *Only Fools and Horses*, created by John Sullivan, season 4, episode 4, BBC, 1985.
14. "Sleeping Dogs Lie." *Only Fools and Horses*, created by John Sullivan, season 4, episode 5, BBC, 1985.
15. "As One Door Closes." *Only Fools and Horses*, created by John Sullivan, season 4, episode 7, BBC, 1985.
16. "From Prussia with Love." *Only Fools and Horses*, created by John Sullivan, season 5, episode 1, BBC, 1986.
17. "The Longest Night." *Only Fools and Horses*, created by John Sullivan, season 5, episode 3, BBC, 1986.
18. "Who Wants to Be a Millionaire?" *Only Fools and Horses*, created by John Sullivan, season 5, episode 6, BBC, 1986.
19. "Stage Fright." *Only Fools and Horses*, created by John Sullivan, season 7, episode 3, BBC, 1991.
20. "The Class of '62." *Only Fools and Horses*, created by John Sullivan, season 7, episode 4, BBC, 1991.

References

Apter, Emily. 2009. "Translation -9/11: Terrorism, Immigration, Language Politics." In Bielsa, E. & Hughes, C. (Eds.) *Globalization, Political Violence and Translation*.. New York: Palgrave. 195–206.

Baldo, Michela. 2016. *Analysing Cultural Translation of Post-Migrant Writing: Italian Narratives of Return*. London: Palgrave.

Bandia, Paul F. 2009. "Translation Matters: Linguistic and Cultural Representation." In Inggs, J. & Meintjes, L. (Eds.) *Translation Studies in Africa*. London: Continuum. 1–20.

Barker, Chris; Galasinski, Dariusz. 2001. *Cultural Studies and Discourse Analysis*. London: SAGE Publications.

Bassnett, Susan; Lefevere, Andre. 1998. *Constructing Cultures: Essays On Literary Translation*. Bristol: Multilingual Matters.

Bielsa, Esperança. 2016. *Cosmopolitanism and Translation: Investigations into the*

Experience of the Foreign. London and New York: Routledge.

Caffrey, Colm. 2009. "Relevant Abuse? Investigating the Effects of an Abusive Subtitling Procedure on the Perception of TV Anime Using Eye Tracker and Questionnaire." PhD dissertation, Dublin City University, 2009. Accessed June 8, 2023. http://doras.dcu.ie/14835/1/Colm_PhDCorrections.pdf

Carbonell, Ovidi. 1996. "The Exotic Space of Cultural Translation." In Alvarez, R. & Vidal, M. C. A. (Eds.) *Translation, Power, Subversion*. Clevedon, UK: Multilingual Matters. 79–98.

Chen, Yan; Huang, Jingjing. 2014. "The Culture Turn in Translation Studies". *Open Journal of Modern Linguistics* 4(4). 487–494.

Denotacija. Bruno Kragić (Ed.). 2021. Leksikografski zavod Miroslav Krleža. https://www.enciklopedija.hr/natuknica.aspx?ID=14586.

Franco Aixelà, Javier. 1996. "Culture-Specific Items in Translation." In Alvarez, Roman & Vidal, M. C. A. (Eds.). *Translation, Power, Subversion*. Clevedon: Multilingual Matters LTD. 52–79.

Hartley, John. 2002. *Communication, Cultural and Media Studies, The Key Concepts*. London: Routledge, Taylor & Francis Group.

Gentzler, Edwin. 1998. "Foreword." In Bassnett, S. & Lefevere, A. (Eds.) *Constructing Cultures: Essays on Literary Translation*, i–xxii. Clevedon, UK: Multilingual Matters.

Guillot, Marie-Noëlle. 2010. "Film Subtitles from a Cross-Cultural Pragmatics Perspective: Issues of Linguistic and Cultural Representation." *The Translator* 16(1). 67–92.

House, Juliane. 2009. "Moving across Languages and Cultures in Translation as Intercultural Communication." In Bührig, K., House, J. & ten Thije, J. D. (Eds.) *Translatory Action and Intercultural Communication*. Manchester: St. Jerome. 7–17.

House, Julianne. 2015. *Translation Quality Assessment: Past and present*. London: Routledge.

Hussein, Basel Al-Sheikh. 2012. "The Sapir-Whorf Hypothesis Today." *Theory and Practice in Language Studies* 2(3). 642–646.

Katan, David. 2002. "Mediating the Point of Refraction and Playing with the Perlocutionary Effect: A Translator's Choice?" In Herbrechter, S. (Ed.) *Cultural Studies, Interdisciplinarity, and Translation*, 20. Amsterdam: Rodopi. 177–195.

Kharina, Alla. 2018. *Realia in Literary Translation*. University of Oslo.

Kramsch, Claire. 1998. *Language and Culture*. Oxford: Oxford University Press.

Kroeber, A. L.; Kluckhohn, Clyde. 1952. "Culture: a critical review of concepts and definitions." *Papers. Peabody Museum of Archaeology & Ethnology* 47(1). Harvard University. viii.

Kultura. (Ed.) Bruno Kragić. 2021. Leksikografski zavod Miroslav Krleža. https://www.enciklopedija.hr/natuknica.aspx?ID=34552.

Nedergaard-Larsen, Birgit. 1993. "Culture-bound problems in subtitling." *Perspectives: Studies in Translatology* 2. 207–241.

Nida, Eugene; Taber, Charles. 1969. *The Theory and Practice of Translation*. Leiden: E. J. Brill.

Pedersen, Jan. 2011. *Subtitling Norms for Television; An exploration focussing on extra-linguistic cultural references*. John Benjamins Publishing Company: Amsterdam.

Reiss, Katharina; Vermeer, Hans J.. 2014. *Towards a General Theory of Translational Action; Skopos Theory Explained.* London: Routledge.

Schleiermacher, Friedrich. 1963. "Ueber die verschiedenen Methoden des Uebersezens." In Störig, H. J. (Ed.) *Das Problem des Übersetzens*. Stuttgart. 38–69.

Snell-Hornby, Mary. 1988. *Translation Studies: An Integrated Approach*. Amsterdam: J. Benjamins Publishing Company.

Tylor, Edward B. 1871. *Primitive Culture*. London: John Murray.

Tymoczko, Maria. 2005. "Trajectories of Research in Translation Studies". *Meta* 50(4): 1082–1097.

Venuti, Lawrence. 1995. *The Translator's Invisibility; A history of translation*. London: Routledge.

Venuti, Lawrence. 1996. "Translation as a Social Practice: Or, the Violence of Translation." *Translation Perspectives* 9. 195–213.

Venuti, Lawrence. 1998. *The Scandals of Translation: Towards an Ethics of Difference*. London: Routledge.

Venuti, Lawrence. 2013. *Translation Changes Everything*. Abingdon, UK: Routledge.

Williams, Raymond. 1976. *A Vocabulary of Culture and Society*. New York: Oxford University Press.

Znak. Bruno Kragić. (Ed.) 2021. Leksikografski zavod Miroslav Krleža. (August 30, 2023). https://www.enciklopedija.hr/natuknica.aspx?ID=67350.

Eriola Qafzezi

Fan S. Noli University Korça, Albania
eriola_bonja@yahoo.com

Inside Out: A Corpus-Driven Study of Expressions with Parts of the Body in the Albanian Language

Abstract

The aim of this paper is to illustrate expressions with parts of the body in the Albanian language, based on a corpus-driven study. Research that is based on corpora is an area of study yet to be explored in the Albanian language. This is due to the fact that corpora of the Albanian language have only recently been developed, and thus corpus-based linguistic research and/or corpus-driven research present challenging and innovative pursuits. Our study primarily adopts a qualitative approach, based on a corpus-driven analysis, enriched by a few quantitative data that show the number of occurrences of linguistic expressions for each part of the body under investigation. We have extracted examples from the Albanian National Corpus (ANC). We begin the paper by presenting the research questions and then we outline some relevant literature background. The corpus and methodology of the study are described in the third section, followed by the analysis and discussion of the data in the fourth section. The paper ends with conclusions based on the current research and suggestions on how to extend the study in the future.

Keywords: corpora, Albanian language, culture, parts, body, ANC

1. Introduction

Linguists have always been interested in discovering meaning. It is specifically due to collocation and the neoFirthian approach to word meaning that we conclude that meaning does not reside with a word in isolation, but with other words that are combined with it. Furthermore, the study of language use from a cognitive linguistic point of view endeavors to identify underlying conceptual systems that allow for meaning construction. In support of these views, we have focused our research on linguistic expressions that contain parts of the body in the Albanian language, based on the Albanian National Corpus (ANC). The research is limited to these terms: *kokë (head)*, *këmbë (foot)*, *zemër (heart)*, *gojë (mouth)*, and *hundë (nose)*. Our research questions are outlined below:

What are the most common uses of expressions with parts of the body in the Albanian language based on the ANC?

How can corpus-driven research aid our understanding of Albanian culture?

The paper is organized as follows. Section 2 gives a brief overview of related literature, specifically on the conceptualization of body parts from the perspective of cognitive linguistics. Section 3 describes the corpus and methodology of the current study. Section 4 analyses the data gathered in the current research and discusses the results of our corpus-driven research, providing taxonomies for the uses of linguistic expressions for each body part. The final section outlines conclusions based on the research and suggestions for further research inspired by the current one.

2. Review of Related Literature

The current research combines corpus studies with cognitive linguistics and cultural studies. As already mentioned, the data that support the research with examples and classifications are extracted from the ANC. One of the central themes in cognitive linguistics is the uniquely human development of some higher potential called the 'mind' and, specifically, the intertwining of body and mind, which has come to be known as 'embodiment' (e.g., Gibbs 2006; Johnson 1987; Lakoff 1987; Lakoff and Johnson 1999). 'Embodiment' refers not only to the process of cognition through bodily experience, but, more broadly, it refers to more abstract domains of cognition, such as those of thought, emotion, and language, based on human body and conceptualization of the internal body parts. As Gibbs puts it: "The key feature of this premise is that understanding the embodied nature of human cognition demands that researchers specifically look for possible mind-body and language-body connections." (Gibbs 2006: 9). Through this research, we intend to illustrate corpus-driven research by extracting from the ANC several examples of linguistic expressions that contain body parts. In the past, such research has been conducted for several languages and cultures with reference to diverse body parts as well as internal organs, as mentioned below.

Deignan and Potter (2004) provide a study of metaphors and metonyms in English and Italian. Based on large, computerized corpora of English and Italian, they examine the power of conceptual metaphor theory to explain the non-literal senses of lexis from the field of human body. Their study provides interesting insights into metonymy and its interactions with metaphor, which account for more non-literal expressions than metaphor alone, both in terms of type and token. Researchers also conclude that, despite some differences at a fairly detailed level, very similar patterns were noted in English and in Italian, so that the two languages appear to be similar both in the types of non-literal language that is used and in its grounds: both show interactions between metaphor and metonymy, and both draw on roughly the same small set of body-mind mappings (Deignan and Potter 2004: 1251).

Morrow (2009) has undertaken another such comparative study of the uses and phraseology associated with the two common nouns *hand* and *heart*. The main aims of his research are to identify, analyze, and describe phrasal patterns associated

with the lexical items *hand* and *heart*, as both are very common and frequent nouns in English. He investigated the use of these lexical items in phrases by extracting and analyzing phrases containing those words from the British National Corpus (BNC), using the interface called Phrases in English (PIE) developed by Fletcher (2003/2004). The researcher concludes that the investigation of *hand* and *heart* use in the BNC showed that their high frequency of usage was related to their extensive metaphorical use. He also found that in the case of *hand*, the high frequency could be attributed partly to the tendency of this word to be used in phrases, while *heart* did not exhibit a strong tendency to be used in phrases; however, its use in a metaphorical sense was noteworthy, which, as the researcher maintains, is hardly surprising since the heart symbol is widely associated with emotion. Furthermore, *hand* and *heart* were frequently used in locative expressions, and their usage in locative expressions contributed to their high overall frequency in the BNC. However, the two words showed different patterns of usage. *Hand* was used with very high frequency but in a restricted set of phrases, particularly in directional phrases where it collocated with *right* or *left*, whereas *heart* was used in a metaphorical sense. Morrow's research contributes to other studies that support the thesis that body part names are a very important source for metaphors to describe human experience (Morrow 2009: 19).

Iranian researchers Atef-Vahid and Zahedi provide more data for a cross-linguistic analysis of body-part metaphor conceptualizations from a cognitive semiosis perspective, aiming to contribute to research which can reveal how different languages attempt to convey certain ideas through the metaphorical mapping of body parts, citing examples from other research into languages of Brazil, Indochina, Ommura, Shona, Chinese, Mwan, South Mande, etc. (2013). They use a mixed-method approach for the analysis of the Farsi and English body metaphors, based on Lakovian cognitive linguistics. They group metaphors into relevant linguistic categories: hair, forehead, nose, lips, face, and tongue, and then compare the frequency of linguistic categories by body parts. Their findings indicate that limitations in terms of availability, semantic domain and range, and linguistic manifestation of metaphors and the accuracy and appropriateness of their application vary from one language to the other, because metaphorical expressions are profoundly embedded and intertwined in people's cognitive abilities of semiotic representations. They also suggest the existence of a universal cognitive framework for humanizing the external world through semiosis, since body parts are commonly utilized metaphorically, in various extents by different languages. Researchers conclude that the shared cognitive pool may be selected and externalized differently by different languages, influenced by culture and possibly religion (Atef-Vahid and Zahedi 2013: 138).

Kiš Žuvela and Parizoska (2023), supporting the view that in cognitive linguistics the human body is the basis of our understanding of the world, focus their research on the constructions with the noun *lice* (Eng. face) in Croatian and reveal what aspects of the face are salient in our conceptualization of certain experiences.

Their research has been performed in the Croatian web corpus hrWaC (1.2 billion words) using the Sketch Engine. The results of their research show that word combinations in which the noun *lice* appears reflect two main cultural models of the English noun *face* in Croatian. One is the communication model (*face* as the most prominent part of body in human interaction), and the other is the emotion model (*face* indicates a state that a person is in or the emotion they are experiencing) (Kiš Žuvela and Parizoska 2023: 173). The researchers have divided the lexico-grammatical constructions with the noun *lice* into four meaning groups based on the cultural model they reflect (or a combination of the two models). The results of their study in the Croatian web corpus hrWaC show that the forms and meanings of particular lexico-grammatical constructions in which *lice* occurs are closely related to the cultural model of the the English noun *face* motivating them (Kiš Žuvela and Parizoska 2023: 186). Researchers also conclude that figurative uses of *lice* are fairly restricted lexically and grammatically and a number of those relatively fixed expressions are idioms in their own right; therefore, information about form is key to interpreting the meaning of particular constructions in which *lice* occurs (Kiš Žuvela and Parizoska 2023: 188).

Another useful resource for the research into cognitive linguistics is the rich source of examples provided in the book *Culture, Body, and Language* edited by Sharifian et al. (2008), with extensive examples about conceptualizations of internal body organs across cultures and languages. There are a multitude of studies, ranging from abdomen-centering conceptualizations, focusing on *liver, heart, guts* in languages such as Indonesian, Malay, and Basque; followed by holistic heart-centering conceptualizations, materialized by research into the Chinese, Japanese, and Korean *heart*; concluded with another extensive chapter about the dualistic *heart/head* and *heart-stomach* centering conceptualizations in Persian, Northeastern Neo-Aramaic, and classical Syriac model of temperaments, *hearts*, and *minds* in Old English, *heart* in Dutch, and *heart* and cultural embodiment in Tunisian Arabic.

3. The Corpus and Methodology of Research

The data used in this research have been extracted from the Albanian National Corpus (ANC), a collection of 31.12 million words. Text collection involves collaboration with publishing houses in Kosovo and Albania (Morozova and Rusakov 2013: 86). There are two corpora available in the ANC: Corpus of modern literary Albanian (Main Corpus) and Corpus of early Albanian texts. The difference lies in the kind of texts they contain and how these texts are presented, whereas search capabilities and annotation are mostly identical. The corpus of the Albanian language did not exist until the end of 2011, when the Corpus was developed as a result of efforts of the creative community of linguists from Saint Petersburg (Institute for Linguistic Studies of the Russian Academy of Sciences) and Moscow (School of Linguistics at HSE). The current version of the ANC uses the morphological *analyzer* and the

tsakorpus platform and provides reference data for both professional linguists and anyone interested in the Albanian language and its history, Albanian lexicon and grammar, as well as language changes which happened in Albanian in the previous centuries. This study has extracted data from the Main Corpus of the ANC. Primary reasons cited for the utilization of the ANC are related to the grammar, history, and lexicon of the Albanian language, applicable to both native Albanian speakers and learners of Albanian as a foreign language (Morozova and Rusakov 2013: 95). Among other aims, we intend to demonstrate that the ANC can extend its applicability to fields such as cultural studies, cognitive studies, and sociolinguistic studies, among applications strictly related to grammar and morphology (e.g., Morozova 2012, 2013, 2015). Yet, such corpus-driven research is still scarce and the current study aims to fill this gap and contribute to such research.

In essence, corpus-driven research is more exploratory, allowing for the discovery of new linguistic patterns, while corpus-based research is more hypothesis-driven, using the corpus to test or validate existing theories or hypotheses about language. Both approaches are valuable in linguistics and language studies, providing insights into how language works and evolves. The current study adopts a corpus-driven approach, in which the researcher has been involved in a research with the aim of identifying the linguistic expressions with parts of the body in the ANC. Patterns and regularities of use will be outlined, if observed, in order to identify linguistic phenomena that guide our exploration and analysis of expressions with parts of the body. From all body parts, the researcher has limited the research into the investigation of *kokë* (Eng. head)*, këmbë (foot)*, *zemër (heart)*, *gojë (mouth)*, and *hundë (nose)*. The noun *dorë (hand)* has been excluded from the actual research because it is currently under review in another journal, in a paper which compares the use of *dorë* in Albanian and *hand* in English, in the ANC and COCA respectively. Meanwhile, the words *sy (eye)* and *vesh (ear)* have been excluded from the current study because they yielded higher occurrences and required greater space and time for analysis and could not be included within the limitations of a single paper. The figures below illustrate instances of our search in the ANC. In Figure 1, we can see the way in which the search is conducted in the ANC, when we type, for example, '*zemër'* (heart) under *lemma,* and the search results which show that '*zemër'* appears in approximately 2167 documents, with a frequency of 6571. The statistics for the usage of this lemma are shown in Figure 2, regarding its usage in press, fiction, non-fiction, poetry, etc. In the ANC, we can continue our search by clicking on *Search Sentences* and be provided with numerous sentences in which the lemma is used, as illustrated in Figure 3.

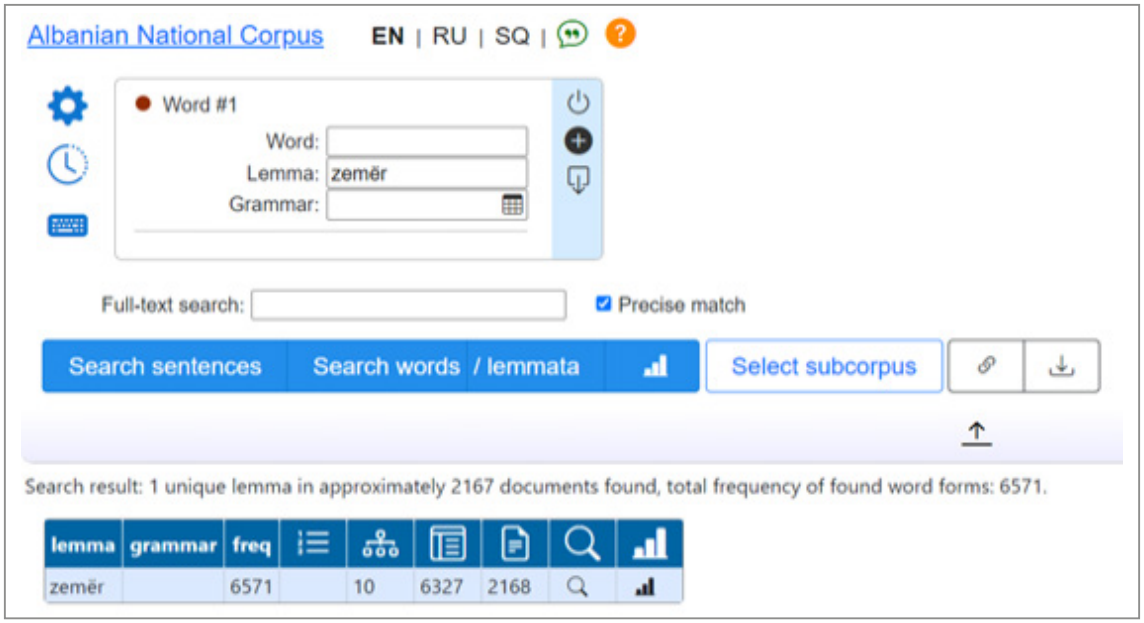

Figure 1. Interface of the ANC for the search results of zemër *under lemma*

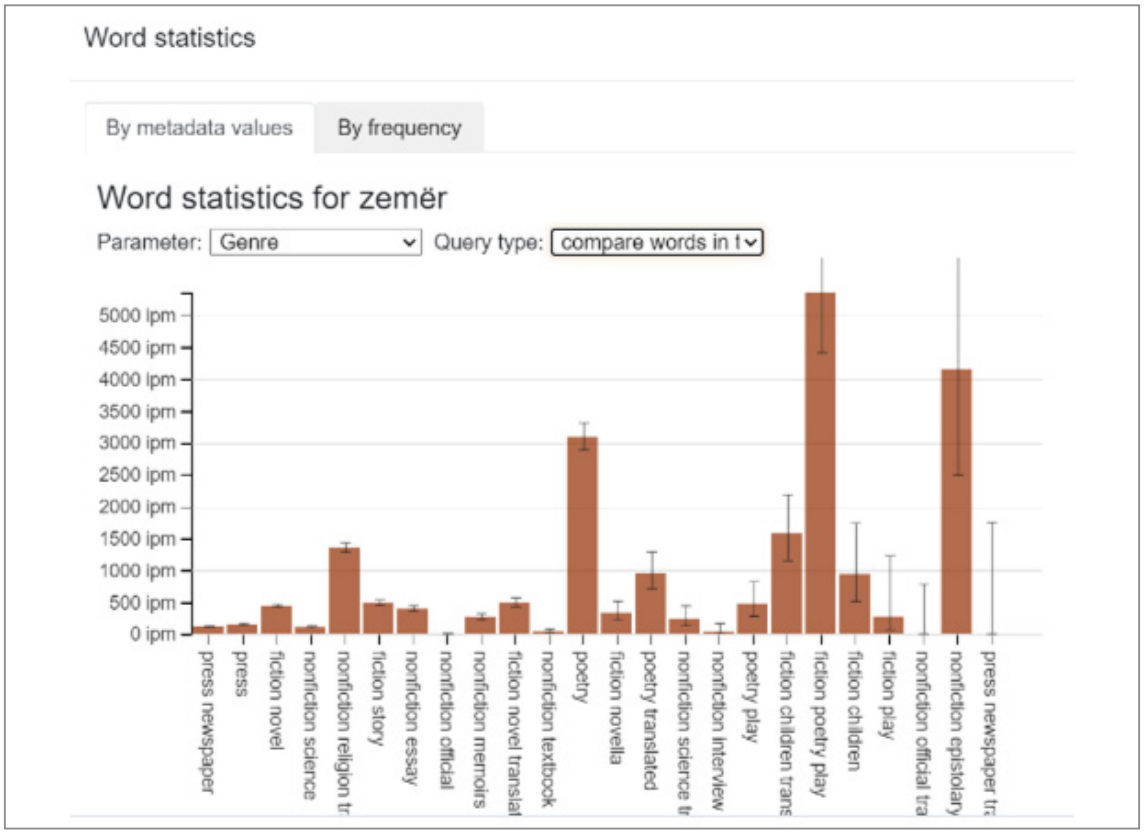

Figure 2. Interface of the ANC for statistics of the lemma zemër *across genres*

Figure 3. Interface of the ANC for the lemma zemër *under Search Sentences*

Since our study adopts mainly a qualitative approach, we only mention a few statistical data to provide some general idea about the number of occurrences of each lemma, as well as the number of sentences and the number of documents in which the lemma appears. The results yielded from the search in the ANC are outlined in Table 1, Section 4. All lines produced from the search in the ANC were automatically manually analyzed, and then, based on the total examples extracted, a classification of uses of each part of the body was created, which will be presented in Section 4. Examples will also be provided based on the extracts from the ANC. The data analysis and discussion section which follows provides both the quantitative and qualitative data that will help towards a better comprehension of the research questions posed by the current study.

4. Data Analysis and Discussion

Based on the search conducted in the ANC for each of the parts of the body, the following data were collected with reference to the number of occurrences found. Albanian is a syntactic language, with a complex system of inflections which indicate definiteness, case, gender, number, and tense. That is why we conducted our search in the ANC for lemmas rather than for words, which allows for a more comprehensive analysis to uncover patterns in word usage across different forms of a word. Thus, the data in Table 1 show the results of searches under *lemma* for each of the words under investigation that denote body parts.

Table 1. Results of searches of occurrences in the ANC

Part of body	No. of occurrences	No. of sentences	No. of documents
kokë	10434	9890	2249
këmbë	6500	6306	1892
*zem*ër	6571	6327	2167
*goj*ë	3723	3598	1341
hundë	1015	987	418

We observe that the noun *kokë (head)* is the most frequent part of the body that appears in the ANC, followed by *zemër (heart), këmbë (foot), gojë (mouth)*, and *hundë (nose)*. After manual analysis of all the examples for each part of the body in our corpus, we present below a classification of the main uses for each part of the body. Each classification provides a list of examples for the linguistic expressions for each part of the body, which answers the first question of our research: What are the most common expressions with parts of the body in the Albanian language based on the ANC? Each list is concluded with a section which illustrates culture-specific expressions with body parts in the Albanian language. These expressions often carry cultural nuances, historical references, or specific contexts, and they can be a window into the Albanian values, beliefs, and shared experiences of Albanian culture. Such examples aim to answer the second question of our research: How can corpus-driven research aid our understanding of Albanian culture?

4.1. Classification of Uses for the Linguistic Expressions with *Kokë* (Eng. Head)

1. **Part of the body:**

   Alb.: "Me xhaketën e trashë prej meshini dhe kasketën që s'e hiqte kurrë, ky më ktheu shpinën, kurse mjeku, si shfryu përpjetë një shtëllungë tymi, më vështroi nga këmbët te *koka*."

   Eng.: "With his thick leather jacket and the cap that he never took off, he turned his back to me, while the doctor, exhaling a puff of smoke, looked me over from *head* to toe."

2. **Control:**

   Alb.: "Tek e fundit, vetë ai ishte një vartës dhe ka një shef mbi *kokë*."

   Eng.: "After all, he himself was an employee and had a boss above his *head*."

3. **Subversion:**

   Alb.: "Nuk kisha asnjë nga arsyet që unë të jepja dorëheqjen dhe dhënia e dorëheqjes Kryeministrit, i cili në thelb atakonte qenien time në qeveri, ishte një

ulje *koke* që unë nuk e pranova."

Eng.: "I had none of the reasons to resign, and submitting my resignation to the Prime Minister, who fundamentally attacked my presence in the government, was a bowing of the *head* that I did not accept."

4. **Violent actions:**

Alb.: "Në orën 08.00 të mëngjesit të po kësaj date, ishte ekzekutuar me plumb pas *koke* Drejtori i Përgjithshëm i Burgjeve, Bujar Kaloshi."

Eng.: "At 8:00 in the morning on this same date, the General Director of Prisons, Bujar Kaloshi, was executed by a gunshot to the *head*."

5. **Difficult situation:**

Alb.: "Neve na vjen *koka* rrotull, ti këndon gazetën e kukurisesh."

Eng.: "Our *head* is spinning around, while you are reading the newspaper and chuckling."

6. **Proximity:**

Alb.: "Ish-kryeministri Sali Berisha në përgjigjen e pyetjes se a ka komunikuar me Lulzim Bashën mbrëmjen që ai u mbyll 3 orë në Kryesinë e Kuvendit *kokë më kokë* me Ramën dhe dolën me një marrëveshje, tha se nuk kishte mundësi ta bënte këtë."

Eng.: "Former Prime Minister Sali Berisha, in response to the question of whether he had communicated with Lulzim Basha the evening they spent three hours in the Parliament's Directorate discussing with Edi Rama *head to head* and reached an agreement, said he did not have the opportunity to do so."

7. **Rebellion in search for freedom:**

Alb.: "Ata na e kthyen në normalitet faktin që kush punon me djersë në këtë vend nuk ngre kurrë *kokë*, e kush vjedh e shet drejtësi duhet të kapardiset edhe me fodullëk."

Eng.: "They made it sound normal that whoever sweats while working in this country never raises their *head*, and whoever steals and sells justice should also be flaunting with pride."

8. **Position of authority:**

Alb.: "Ai është *koka* e fshatit, jo kryeplaku, o Sabri."

Eng.: "Sabri, he is the *head* of the village, and not the reeve."

9. **Mind:**

Alb.: "Sa më shumë që dëshiron të hysh në *kokat* e heronjve të tu, aq më shumë do të dish çka po ndodh në kokën tënde."

Eng.: "The more you get inside the *heads* of your heroes, the more you will know what is happening in your own *head*."

10. Metonymy:

Alb.: "Të tjerë duan *koka* drejtorësh dhe ministrash."

Eng.: "Others desire the *heads* of directors and ministers."

11. Figur1ative uses:

Alb.: "Mos kini turp ta vishni, sepse do të ktheni *koka*, në kuptimin e mirë të fjalës."

Eng.: "Do not be ashamed to wear it, because you will turn *heads*, in the positive sense of the word."

12. *Kokë* and *zemër* (Eng. head and heart)

Alb.: "O, Zot! bëri ai sërish dhe u përpoq të ndalte mendimet, që më shumë se nga *koka*, i vinin nga *zemra*."

Eng.: "Oh, God! he exclaimed again, trying to halt his thoughts, which were coming more from the *heart* than from the *head*."

13. Culture-specific expressions:

Alb.: "Bariu mbeti atje i vetmuar dhe i tmerruar dhe në mendje i sillej thënia e vjetër: '*Koka* e bën, *koka* e pëson.'"

Eng.: "The shepherd remained there alone and frightened, and in his mind, the old saying echoed: 'You reap what you sow' (What your *head* does, your *head* suffers)."

The expression 'What your head does, your head suffers' is a typical expression in the Albanian language which means that somebody has 'to reap what they sow' or 'lie on the bed they made'.

4.2. Classification of Uses for the Linguistic Expressions with *Këmbë* (Eng. Foot)

1. Part of the body:

Alb.: "Braziliani i ri i Manchester City, Gabriel Jesus, ka pësuar një frakturë në *këmbë*, bëri të ditur të martën klubi i tij."

Eng.: "The young Brazilian of Manchester City, Gabriel Jesus, has suffered a *leg* fracture, his club announced on Tuesday."

2. Walking on *foot:*

Alb.: "Ka bërë rreth 70 metra në *këmbë*."

Eng.: "He has walked about 70 meters on *foot*."

3. **Bad handwriting:**

   Alb.: "Mëso të shkruash shqip se të duhet për veten tende se s'i shkon burrit të shkruajë me *këmbë* pule gjuhën e nënës."

   Eng.: "Learn to write in Albanian because you need it for yourself, as it doesn't suit a man to write his mother tongue with chicken *feet*."

4. **Proximity and collaboration:**

   Alb.: "Është koha për t'u ulur *këmbë* kryq me bazën për t'i dëgjuar e zgjidhur hallet që ata kanë ndryshe një ditë shpejt apo vonë do të ndëshkoheni nga ky popull."

   Eng.: "It's time to sit down with crossed *legs* to listen and solve the issues they have, otherwise one day, sooner or later, you will be held accountable by this people."

5. **Metonymy:**

   Alb.: "Edhe këtë ditë nuk u duk *këmbë* njeriu nëpër rrugë, (për të punuar e kam fjalën), ose ndonjë makinë pune e elektrikut apo ndonjë excavator a diçka e tillë, vetëm policët e shkretë ishin në krye të detyrës, pa semafor dhe me rrezik kriminaliteti të shtuar."

   Eng.: "Even today, not a single *foot* was seen on the streets (I mean, to work), or any work vehicles, electric cars, or excavators, only the poor police were on duty without traffic lights and with the added risk of crime."

6. **Figurative uses:**

   Alb.: "Messi te Inter është vecse nje ëndërr, ndaj tekniku Stefano Pioli qëndron me *këmbë* në tokë."

   Eng.: "Messi in Inter is nothing but a dream, so coach Stefano Pioli remains with his *feet* on the ground."

7. ***Këmbë* and *kokë* (Eng. foot and head):**

   Alb.: "Qëllimi ynë kryesor dhe i përbashkët është që t'i shkurtojmë sa më shumë ditët kësaj qeverie dhe këtij kryeministri të lidhur *kokë* e *këmbë* me krimin dhe trafikun e drogës."

   Eng.: "Our main and common goal is to shorten the days of this government and this prime minister, who are *head and foot* linked to crime and drug trafficking."

8. **Culture-specific expressions:**

   Alb.: "Shumë politikanë do t'i bien kokës me grushte që kërkuan Europën' me-demek se Europa po ua fut *këmbët* në një këpucë."

Eng.: "Many politicians will be banging their heads with fists for asking to join Europe, as it seems Europe is putting both their *feet* in a shoe."

The expression 'put both feet in one shoe' in the Albanian language means 'to force somebody to do something'.

4.3. Classification of Uses for the Linguistic Expressions with *Zemër* (Eng. Heart)

1. **Part of the body:**

   Alb.: "Kjo shkakton anemi dhe në rast se nuk mjekohet, *zemra* dhe organe të tjera në trup nuk do të arrijnë të kryejnë funksionet e tyre, si rezultat i mungesës së oksigjenit."

   Eng.: "This causes anemia, and if not treated, the *heart* and other organs in the body will not be able to perform their functions due to the lack of oxygen."

2. **(Lack of) emotions:**

   Alb.: "Korovievi buzëqeshi në mënyrë domethënëse, duke përkulur trupin dhe Margaritës përsëri iu bë *zemra* akull."

   Eng.: "Koroviev smiled meaningfully, bowing his body, and Margarita's *heart* turned into ice once again."

   Alb.: Evgjitët tanë nuk e kanë humbur traditën e bukur të tyre, *zemra* atyre u këndon, ata u bien instrumenteve."

   Eng.: "Our Roma people have not lost their beautiful tradition; their *heart* sings, they play their instruments."

3. **Location:**

   Alb.: "Ky tunel është me standartet e fundit europiane, është me dy tuba në *zemër* të malit me gjatësi mesatare rreth 2.5 km secili."

   Eng.: This tunnel complies with the latest European standards; it has two tubes in the *heart* of the mountain, each with an average length of about 2.5 km."

4. **Metonymy:**

   Alb.: "Kozmai, sa më shumë largohej nga *zemra* që e kishte bërë për vehte, aq më të fuqishme e ndjente tërheqjen e saj dhe aq më të shpejtë e më të vrullshëm priste kthimin e shëmbjen..."

   Eng.: "The more Kozma distanced himself from the *heart* he had attracted, the more powerfully he felt its attraction, and the faster and more vibrant he awaited the downfall..."

5. **Figurative uses:**

   Alb.: "Kur mbështeti kokën në kraharorin tim, kur ia përkëdhela ata flokë të zez si *zemra* e natës, ajo nxori nga xhepi i këmishës së bardhë një margaritar, të cilin ma lëshoi në dorë."

   Eng.: "When she rested her head on my shoulder, when I caressed those black hair like the *heart* of the night, she pulled a gem from the pocket of her white shirt and handed it to me."

6. ***Zemër* and other parts of the body:**

   Alb.: "Më udhëheq *truri* e *zemra*, jo *barku* dhe llogaritë!"

   Eng.: "I'm guided by my *mind* and my *heart*, not by my *stomach* and my calculations!"

7. **Culture-specific expressions:**

   Alb.: "E kur u thotë 'po' atëherë shumica e tyre i shtrëngojnë dorën dhe urimi i tyre është: 'të këndoftë *zemra*'."

   Eng.: "And when she says 'yes,' then most of them shake hands, and their wish is: 'may your heart *sing*'."

The expression 'may your heart sing' is typically used in Albanian to express a wish for the others to feel happy and joyful.

4.4. Classification of Uses for the Linguistic Expressions with *Gojë* (Eng. Mouth)

1. **Part of the body:**

   Alb.: "*Goja* vjen në kontakt me shumë substanca që përmbajnë proteina, karbohidrate dhe glukozë."

   Eng.: "The *mouth* comes into contact with many substances that contain proteins, carbohydrates, and glucose."

2. **Sound:**

   Alb.: "Në fund të sheshit një ushtar italian i binte një muzike *goje*, duke vështruar vajzat që kalonin."

   Eng.: "At the end of the square, an Italian soldier was *mouthing* a song, observing the girls passing by."

3. **By word of *mouth:***

   Alb.: "Një prej këtyre manifestimeve është edhe fjala që bredh *gojë prej goje*, në shumë demagogë tanë për organizatorë të huaj."

Eng.: "One of these manifestations is also the word that spreads *from mouth to mouth* from a lot of our demagogues to foreign organizers."

4. **Suppress opinions:**

   Alb.: "Iu është qepur *goja* të gjithëve."

   Eng.: "Everybody's *mouths* have been sealed."

5. **Metonymy:**

   Alb.: "*Gojët* e liga thonë se je më afër se kurrë postit të Presidentit të Republikës?"

   Eng.: "The wicked *mouths* say that you are closer than ever to the post of the President of the Republic?"

6. **Figurative uses:**

   Alb.: "Aq më tepër që ushtria gjermane nuk ishte as dyzet milje larg, kështu që, me siguri, *goja* kishte nisur t'i lëshonte lëng, si ujkut përballë një qengji."

   Eng.: "Moreover, the German army was not even forty miles away, so surely the *mouth* had started to salivate, like a wolf in front of a lamb."

7. ***Gojë* and other parts of the body:**

   Alb.: "*Zemra* e njeriut me *mend* kërkon dijen, por *goja* e budallenjve ushqehet me marrëzi."

   Eng.: "The *heart* of a person with understanding seeks knowledge, but the *mouth* of fools feeds on foolishness."

8. **Culture-specific expressions:**

   Alb.: "Ia ktheva edhe unë buzëqeshjen duke shtuar: 'Të lumtë *goja*, zotëri, s'do ta harroj kurrsesi!'"

   Eng.: "I returned the smile, adding: 'Blessed be your *mouth*, sir, I will never forget it!'"

   Alb.: "'Tu thaftë *goja*!' i kishte thënë ajo të shoqit dhe ia kishte përkëdheluar faqet e përlotura."

   Eng.: "'Cursed be your *mouth*!' she had said to her husband and had caressed his tearful cheeks."

The expressions 'blessed/cursed be your mouth' are used in the Albanian language on occasions when the person would like to encourage or dismiss current event, respectively.

4.5. Classification of Uses for the Linguistic Expressions with *Hundë* (Eng. Nose)

1. **Part of the body:**

   Alb.: "Dhe me të vërtetë, Xha Brahua i ngjante mjaft, nga fytyra, Skënderbeut tonë: i tretur, mjekërbardhë, *hundë* me samar—tamam *hundë* shqiptari."

   Eng.: "And truly, Uncle Braho resembled him quite a bit, in appearance, to our Skanderbeg: beardless, fair-skinned, hooked *nose*—exactly the *nose* of an Albanian."

2. **Nasal sound:**

   Alb.: "Kisha vënë re se, kur jepte këshilla, zëri i bëhej më me *hundë*."

   Eng.: "I had noticed that when he gave advice, his voice became more *nasal*."

3. **Smoking:**

   Alb.: "Më shumë u lodha nga që m'u ça *hunda* për një cigare."

   Eng.: "I got more tired because my *nose* was longing for a cigarette."

4. **Use of drugs:**

   Alb.: "Në rregull, ia ktheva dhe mora një vizë me *hundë*."

   Eng.: "I agreed and then took a sniff with my *nose*."

5. **Lack of foresight:**

   Alb.: "Por mbrojtësit e sotëm të asaj Lufte, ose shohin deri te *hunda*, ose ca më keq, nuk janë prekur ata vetë apo rrethi i tyre ngushtë, dhe s'mendojnë as për viktimat e rrethit të gjerë."

   Eng.: "But the current defenders of that War, either they see only up to their *noses*, or worse, since they themselves or their narrow circle are not affected, and they don't even think about the victims of the wider circle."

6. **Metonymy:**

   Alb.: "Me të hyrë në pavion dhe sa më shumë që i afrohej dhomës së izolimit të Mark Dobjanit, aq më shumë *hunda* e tij prej hetuesi me eksperiencë nuhaste vërtet gjëra të vogla që bashkoheshin me një lloj shpërkujdesje të përgjithshme."

   Eng.: "As he entered the ward and the more he approached Mark Dobjan's isolation room, his experienced investigator's *nose* sniffed out really small things that joined in a general imprecision."

7. **Figurative uses:**

   Alb.: "Sikur nën *hundë* po flitet, sikur po belbëzohet lidhur me këtë çështje, por

ndonjë organ akoma nuk e ka shqyrtuar këtë mundësi apo ndryshimin e qëndrimit."

Eng.: "It's as if something is being whispered under the *nose*, as if it's being murmured about this issue, but some authority still hasn't considered this possibility or change in stance."

8. ***Hundë* and other parts of the body:**

Alb.: "Pushteti nuk ka as *sy*, as *veshë*, as *hundë*, vetëm *gojë*, pavarësisht ngjyresës së tij politike."

Eng.: "Power has neither *eyes*, nor *ears*, nor *nose*, only a *mouth*, regardless of its political colour."

9. **Culture-specific expressions:**

Alb.: "Përse këto punë nuk i lihen ekspertëve, por fut *hundët* pushteti?"

Eng.: "Why aren't these matters left to the experts, but the authorities get their *noses* into it?"

The expression 'get/poke their noses into something' is typically used in Albanian with the meaning 'interfere'.

5. Conclusion

The aim of this research was to illustrate expressions with parts of the body in the Albanian language, based on a corpus-driven study. The classification of linguistic expressions related to body parts *kokë, këmbë, zemër, gojë, hundë* (Eng. *head, foot, heart, mouth, nose*) provides examples which include figurative uses, metonymy, and culture-specific expressions, among the more specific nuances of meaning. The paper also fulfilled its aim to show that the ANC is a valuable resource for the potential of corpus-driven research, an area little explored for the Albanian language. We also support the thesis that corpus-driven research can aid our understanding of Albanian culture, since there were examples of expressions that are specific to the Albanian language and culture, as mentioned in the paper.

The current study could be expanded in the future by broadening the scope and comparative analysis, examining how expressions related to body parts manifest in different cultural and linguistic contexts. Another interesting perspective could be to study changes in the usage of these expressions in different genres, thus delving into subcorpora. By focusing on one or more body parts, future research can thus contribute to a deeper understanding of the intricate interplay between language, culture, and cognition, further enhancing the application of corpora in linguistic studies.

References

Atef-Vahid, Sara; Zahedi, Keivan. 2013. "Cross-linguistic Analysis of Body Part Metaphor Conceptualization from a Cognitive Semiosis Perspective". *BRAIN. Broad Research in Artificial Intelligence and Neuroscience* 4(1-4). 126-140.

Deignan, Alice; Potter, Liz. 2004. "A corpus study of metaphors and metonyms in English and Italian". *Journal of Pragmatics* 36, 1231-1252.

Gibbs, Raymond. 2006. *Embodiment and Cognitive Science.* Cambridge: Cambridge University Press.

Johnson, Mark. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason.* Chicago: University Press of Chicago.

Kiš Žuvela, Sanja; Parizoska, Jelena. 2023. "Constructions with *lice* ('Face') in Croatian: Lexico-Grammar and Cultural Models". In Pattillo, K. & Waśniewska, M. (Eds.) *Embodiment in Cross-Linguistic Studies.* Brill, Leiden, Boston. 170-190.

Lakoff, George; Johnson, Mark. 1999. *Philosophy in the Flesh. The Embodied Mind and its Challenge to Western Thought.* New York: Basic Books.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind.* Chicago: The University of Chicago Press.

Morozova, Maria; Rusakov, Alexander; Arkhangelskiy, Timofey. *Albanian National Corpus.* (Available online at: albanian.web-corpora.net).

Morozova, Maria; Rusakov, Alexander. 2013. "Korpusi elektronik i shqipes: përpunimi, përmbajtja dhe përdorimi". *The XXXII International Seminar For Albanian Language, Literature And Culture*. Prishtinë, Fakulteti i Filologjisë.

Morozova, Maria; and Rusakov, Alexander. 2015. "Albanian National Corpus: Composition, Text Processing and Corpus-oriented Grammar Development". *Albanische Forschungen. Sprache und Kultur der Albaner: Zeitlitche und räumliche Dimensionen. Akten der 5. Deutsch-albanischen kulturwissenschaftlichen Tagung*, Hubert & Co., Göttingen. 270-306.

Morozova, Maria. 2012. "Shënime për standardin morfologjik të Korpusit nacional të shqipes [Notes on the morphological standard of the Albanian National Corpus]". *Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare. Materialet e punimeve të Seminarit XXXI Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*: Prishtinë. Universiteti i Prishtinës: Fakulteti i filologjisë, 31/1. 153–156.

Morrow, Phillip R. 2009. "Hand and Heart: A Study of the Uses and Phraseology Associated with Two Common Nouns." *The Nagoya Gakuin Daigaku Ronshū: Journal of Nagoya Gakuin University; Language and Culture* 20(2). 11-20. doi: http://doi.org/10.15012/00000535

Sharifian, Farzad; Dirven, R.; Yu, Ning; Niemer, Suzanne. 2008. *Culture, Body, and Language. Conceptualizations of Internal Body Organs across Languages and Cultures.* Mouton de Gruyter, Germany.

Internet Sources

http://albanian.web-corpora.net/albanian_corpus/search. Accessed February 2023 to January 2024.

University of Zadar
Universitas Studiorum
Jadertina | 1396 | 2002 |