

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΞΕΝΩΝ ΓΛΩΣΣΩΝ, ΜΕΤΑΦΡΑΣΗΣ ΚΑΙ ΔΙΕΡΜΗΝΕΙΑΣ



Αλίνα Καρακαντά
ΞΔ2009017

Πτυχιακή εργασία
«Linguistic variation in translated and interpreted Greek discourse»

Επιτροπή
Επόπτης: Βλαχόπουλος Στέφανος
Δεύτερος επιβλέπων: Αναστάσιος Ιωαννίδης

Linguistic variation in translated and
interpreted Greek discourse

Γλωσσική μεταβλητότητα στον
μεταφρασμένο και διερμηνευμένο
ελληνικό λόγο

Table of Contents

Abstract.....	4
Περίληψη	5
1. Introduction.....	6
2. Related work.....	7
2.1. Interpreting corpora	7
2.2. Interprete vs. translationese	8
3. Methodology.....	9
3.1. Corpus compilation.....	9
Transcription.....	10
Corpus statistics	11
3.2. Analysis of interprete vs. translationese features.....	11
Simplification.....	11
Distinctive linguistic features.....	12
4. Results.....	13
4.1. Do interpreted texts from English into Greek show higher simplification than translated texts? 13	
4.2. What are the most distinctive linguistic features of translation vs. interpreting (and vice versa) into Greek?	14
5. Conclusions.....	16
References.....	17

Abstract

The proliferation of corpus-based approaches to translation and interpreting has enabled data-driven studies on the properties of oral and written mediated discourse. These studies have provided evidence of ‘translationese’ and ‘interpretese’ effects, showing that translated and interpreted texts have characteristics that differentiate them from original written and spoken texts, such as lower information density and lexical variety. When it comes to mode of discourse (translated vs interpreted), interpreting has been claimed to demonstrate larger similarities to original spoken discourse than written translations but also more salient features of translated texts. For instance, higher simplification effects have been reported in interpreting compared to translation, but the results seem to depend on the language pair. These studies have mostly included major European languages (Italian, Spanish, German, French), focusing on political discourse in the European Parliament. Even though Greek is an official language in the European Parliament (EP), to date there is no study on the properties of simultaneous interpreting (SI) into Greek, rendering it an under-represented language in corpus-based interpreting studies.

In this thesis, we perform an analysis of linguistic variation in Greek mediated discourse, investigating the properties of translated (TR) vs. interpreted (SI) texts from English into Greek. To this aim, we compile EPICEL, the first intermodal TR-SI Greek corpus, based on EP speeches. Using EPICEL, we explore two research questions: Do interpreted texts from English into Greek show higher simplification than translated texts in the same language pair? What are the linguistic features most distinctive of translation vs. interpreting into Greek? Our findings confirm previous research, which suggests shorter renderings and lower lexical variation in interpreting. In addition, we observe a tendency towards verbal structures and coordination in interpreting, while translation prefers nominalisation and subordination. We hope these findings will pave the way towards corpus-based interpreting research into Greek and inform interpreting training frameworks for Greek.

Keywords: interpreting, translation, corpora, simplification, translationese, interpretese, mediation, Greek

Περίληψη

Η πληθώρα σωμάτων κειμένων μετάφρασης και διερμηνείας καθώς και προσεγγίσεων βασισμένων σε αυτά έχει οδηγήσει σε μελέτες σχετικά με τα χαρακτηριστικά του προφορικού και γραπτού διαμεσολαβούμενου λόγου. Αυτές οι μελέτες έχουν δείξει ότι τα μεταφρασμένα και διερμηνευμένα κείμενα έχουν χαρακτηριστικά που τα διαφοροποιούν από τα πρωτότυπα γραπτά και προφορικά κείμενα, όπως χαμηλότερη πυκνότητα πληροφοριών και μικρότερη λεξική ποικιλία. Όσον αφορά τον τρόπο εκφοράς του λόγου (μεταφρασμένος έναντι διερμηνευόμενος), η διερμηνεία παρουσιάζει μεγαλύτερες ομοιότητες με τον μη-διαμεσολαβούμενο προφορικό λόγο από ότι οι γραπτές μεταφράσεις, αλλά και πιο εξέχοντα χαρακτηριστικά μεταφρασμένων κειμένων. Για παράδειγμα, έχει αναφερθεί εντονότερη απλοποίηση στη διερμηνεία σε σύγκριση με τη μετάφραση, αλλά τα αποτελέσματα φαίνεται να εξαρτώνται από το γλωσσικό ζεύγος. Αυτές οι μελέτες έχουν γίνει ως επί το πλείστον σε «κύριες» ευρωπαϊκές γλώσσες (Ιταλικά, Ισπανικά, Γερμανικά, Γαλλικά), εστιάζοντας στον πολιτικό λόγο στο Ευρωπαϊκό Κοινοβούλιο. Παρόλο που η ελληνική είναι επίσημη γλώσσα του Ευρωπαϊκού Κοινοβουλίου (ΕΚ), μέχρι σήμερα τα χαρακτηριστικά της ταυτόχρονης διερμηνείας στα ελληνικά δεν έχουν ερευνηθεί με μεθόδους βασισμένες σε σώματα κειμένων.

Αυτή η πτυχιακή εργασία παρουσιάζει μια ανάλυση γλωσσικής μεταβλητότητας στον ελληνικό διαμεσολαβούμενο λόγο, εστιάζοντας στα χαρακτηριστικά των μεταφρασμένων έναντι διερμηνευμένων κειμένων στο γλωσσικό ζεύγος Αγγλικά-Ελληνικά. Για το σκοπό αυτό, συντάχθηκε το EPICEL, το πρώτο διατροπικό ελληνικό σώμα κειμένων μετάφρασης-διερμηνείας, από ομιλίες του ΕΚ. Χρησιμοποιώντας το EPICEL, θέτονται δύο ερευνητικά ερωτήματα: Παρουσιάζει η διερμηνεία από τα αγγλικά στα ελληνικά μεγαλύτερη απλοποίηση από ότι η μετάφραση; Ποια είναι τα γλωσσικά χαρακτηριστικά που διακρίνουν τη μετάφραση από τη διερμηνεία (και αντιστρόφως) στα ελληνικά; Τα ευρήματά επιβεβαιώνουν προηγούμενες έρευνες, οι οποίες αναφέρουν συντομότερες αποδόσεις και μικρότερη λεξική ποικιλία στη διερμηνεία. Επιπλέον, παρατηρείται μια τάση προς τις ρηματικές φράσεις και παρατακτική σύνδεση στη διερμηνεία, έναντι ονοματικών φράσεων και υποτακτικής σύνδεσης στη μετάφραση. Ελπίζουμε ότι αυτά τα ευρήματα θα αποτελέσουν έναυσμα για περαιτέρω μελέτες των χαρακτηριστικών της διερμηνείας στα ελληνικά και θα χρησιμοποιηθούν στην εκπαίδευση διερμηνέων για τα ελληνικά.

Λέξεις-κλειδιά: διερμηνεία, μετάφραση, σώματα κειμένων, απλοποίηση, διαμεσολάβηση, Ελληνικά

1. Introduction

Corpus-based Translation & Interpreting studies has been established as a blooming research field of translation, enriching the understanding of mediated discourse. For translation, the usefulness of corpus-based approaches has been recognised and widely adopted since the seminal work of Baker (1993). Interpreting studies have been slower to adopt corpus-based approaches (Schlesinger, 1998), mainly due to the difficulty of collecting, transcribing and adequately representing interpreting data. In the recent decades, collaborative research attempts have led to the creation of small to medium-sized interpreting corpora, the most notable being the family of European Parliament Interpreting corpora (Russo et al., 2005; Kajzer-Wietrzny, 2012; Defrancq et al., 2015; Przybyl et al., 2022a; Korzinek and Chmiel, 2021).

The availability and readiness of interpreting corpora has allowed for comparative analyses of interpreted discourse vs. original spoken discourse and interpreted vs. translated texts (Schlesinger & Ordan, 2012; Kajzer-Wietrzny, 2012; Russo et al., 2012; Defrancq et al., 2015; Bernardini et al., 2016; Ferraresi and Miličević, 2017; Lapshinova-Koltunski et al., 2021; Przybyl et al., 2022b). Mediated discourse (translated/interpreted) can be distinguished from non-mediated (original production) based on specific characteristics. These characteristics, such as simplification, normalisation, explicitation, and seem to be independent of language, to the extent that scholars have referred to them as ‘translation universals’ (Baker, 2003; Chesterman, 2012). In addition to ontology (mediated vs. non-mediated), modality (oral vs. written) is another factor of variation, with translated texts and interpreted texts showing different characteristics. Interpreting seems to be an extreme case of translation, where the features related to mediated discourse are more salient, but also shows a strong tendency towards orality (Schlesinger & Ordan, 2012; Lapshinova-Koltunski et al., 2021). For example, a higher degree of simplification for interpreting has been reported in several studies (Kajzer-Wietrzny, 2012; Przybyl et al., 2022b). However, differences between translations and interpretations seem to be dependent on language direction (Dayter, 2018; Ferraresi et al., 2018).

Corpus-based research in simultaneous interpreting (SI) has mainly focused on major languages such as English, Spanish, Italian, French, and German. Even though Greek is an official language in the European Parliament (EP), to date there is no simultaneous interpreting corpus for the Greek language. Studies in simultaneous interpreting are limited (Petrocheilou, 2020), which constitutes a bottleneck for research in Greek language interpreting and interpreter training. In linguistic terms, Greek displays interesting properties for interpreting research, such as a rich inflectional system, relatively flexible word order, a large number of clitics, the ‘pro-drop’ property, and the co-existence of weak and strong pronouns (Joseph & Tserdanelis, 2003). With these characteristics, the study of Greek interpreted discourse can contribute to the discussion on the language-specificity or universality of ‘interpretese’ features.

In this thesis, we compile EPICEL, the first intermodal SI-TR corpus for Greek, from speeches of the EP. In this way, we contribute to the line of research in interpreted and, more generally, mediated discourse in the EP by focusing on Greek, an under-represented language in corpus-based interpreting studies which is less closely related to English in terms of vocabulary and structure. We focus on the differences between modality (translation vs. interpreting). Using EPIC, we pursue two research questions:

- Do interpreted texts from English into Greek show higher simplification than translated texts?

Specifically, we explore whether the simplification hypothesis is confirmed in the case of mediated discourse into Greek. We expect that interpreted texts will demonstrate higher simplification compared to written translations into Greek, which can be decomposed in lower lexical variety, shorter sentences and choice of more high-frequency words.

- What are the most distinctive linguistic features of translation vs. interpreting (and vice versa) into Greek?

We investigate features that distinguish interpreted texts from translated texts at the lexical level. We expect a strong presence of features of spoken discourse, a tendency towards verbalisation and coordination.

This thesis is structured as follows: Section 2 presents previous studies on interpreting corpora and on the features of interpreting vs. translated discourse. The methodology for compiling EPICEL and for the analysis of the features is presented in Section 3. The results of the analysis are shown in Section 4, and the conclusions in Section 5.

2. Related work

2.1. Interpreting corpora

Simultaneous interpreting corpora have served as sources of evidence for validating pre-existing theories and models in interpreting. In general, corpora can be classified as parallel and comparable. Parallel corpora include source texts and their translations/interpretations, normally aligned at the level of sentences or sentence-like units. Comparable corpora are collections of texts that are similar to each other based on specific criteria, for example interpretations and oral non-mediated production (original speeches) in the same language. Corpora can also contain different modalities of translation, for example translation and interpreting, as per Shlesinger's (2009) proposal. Such corpora are called intermodal.

The majority of simultaneous interpreting corpora comes from institutional settings, mainly the European Parliament (EP). Freely available interpreting data in sufficient quantities for conducting reliable statistical analyses are sparse. The EP is an ideal source of interpreting data, since it offers a large amount of original speeches, translations and interpretations in a variety of topics, produced in a formal and institutionalised setting. Rich metadata is available, which allow to control variables such as speaker, party, date, etc. Moreover, data quality is ensured, since interpreters and translators are professionals selected based on strict selection procedures.

The European Parliament Interpreting Corpus (EPIC) (Russo et al., 2005) is among the first collections of interpreting data. EPIC is a parallel corpus of European Parliament source speeches and simultaneously-interpreted target speeches in English, Italian and Spanish. Inspired by EPIC, EP interpreting data were collected for other languages, with EPICG covering English, French, Dutch and Spanish (Defrancq et al., 2015), PINC for Polish and English (Chmiel et al., 2021). Later, comparative approaches between translation and interpreting led to the need for building intermodal corpora, such as the European Parliament Translation & Interpreting Corpus EPTIC (Ferraresi & Bernardini, 2019), an extension of EPIC with the translation mode. The “EPIC suite” has complemented with other intermodal corpora; the Translation and Interpreting corpus (TIC) described in (Kajzer-Wietrzny, 2012) for English and Polish, EPIC-UdS for English, German and Spanish (Przybyl et al., 2022a) and EPTIC-SI for English and Slovenian (Južnič & Peterlin, 2023).

Besides the EP corpora, interpreting corpora have been collected in other institutional contexts. SIREN (Dayter, 2018) is an English-Russian corpus of speeches and interpretations held at the United Nations and CEPIC (Pan, 2019) is a large-scale Chinese-English translation and interpreting corpus of political speeches. There is currently no corpus of interpreting for Greek, the language in focus of this thesis. By building and releasing EPICEL, we contribute to the line of EPTIC corpora by adding a new language and to corpus-based interpreting studies for Greek.

2.2. Interpretese vs. translationese

It has been a long claim that translations have characteristics that differentiate them from original texts written or spoken in the same language. These characteristics are ‘marks’ that the translation process leaves on the product, and they are so particular to translated texts that scholars refer to them as ‘translation universals’. Such universals are simplification (Toury, 1995), explicitation (Blum-Kulka, 1986), normalisation, shining-through (Teich, 2003) and convergence (Laviosa, 2002). Chesterman (2004) makes the distinction between S-universals, which are related to the dependence of the translation on the source text, and T-universals, which relate to the way translators use the target language.

Simplification is a widely studied universal in translation. Translators seem to “produce texts that are less informationally dense and less lexically varied than those produced by people writing or uttering their own texts or speeches in the same language, in similar circumstances.” (Baker 1993, p.4). This hypothesis has been confirmed in a line of studies, where translations were found to have lower lexical density and use more frequent words than originals (Laviosa, 1998a; 1998b; Bernardini et al., 2018).

For interpreting, simplification effects have presented a mixed picture. Comparing original speeches to interpreted discourse, Bernardini et al. (2018) found interpreted texts lexically simpler than originals in EPTIC, but a higher lexical density for interpreted texts was found for Spanish (IT,EN>ES) (Russo et al., 2006), and for English-Polish (Kajzer-Wietrzny, 2012), which was contrary to expectations (and Laviosa’s findings for translation). Different trends depending on the language direction were reported for English-Italian (Sandrelli and

Bendazzoli, 2005) and Russian-English (Dayter, 2018). A possible explanation for this discordance is that simplification may be affected by the mode of delivery of the source speeches (impromptu or semi-scripted versus read out), since spontaneous speech tends to have different characteristics than prepared. The high lexical density of interpreted texts may be a result of condensation techniques which interpreters use to save time (Kajzer- Wietrzny, 2012). For example, interpreters may tend to use lexical items not appearing in the original text to achieve lexical cohesion, for example through repetitions or synonyms.

Comparing interpreting and translation, research has shown that interpreted texts are lexically less varied than translations (Shlesinger, 2009; Bernardini et al., 2018). However, parameters of simplification seem to apply differently at different linguistic levels. For example, English is simpler at the lexical level, while Italian at the lexico-syntactic level. Looking at phraseological patterns, Ferraresi, A. and M. Miličević (2017) found that translations are more phraseologically conventional than interpretations, and that phraseological patterns are more dissimilar in mediated discourse (between translations and interpretations) than they are in non-mediated discourse (between original spoken and written texts). In terms of syntactic simplicity, interpreted texts tend to have shorter sentences (Russo et al., 2012) and shorter syntactic dependencies (Liang et al., 2017), which may be an effect of attempts to reduce cognitive load.

These findings have led to investigations of the specific characteristics of interpreting. Interpreting lies in the space between spoken and translated and this raises the question whether the specific characteristics of interpreting are related to mediation or to orality. Shlesinger (1989) claimed that “interpreting exerts a levelling effect: oral texts become more literate, literate texts become more oral”. Interpreting is seen as an extreme case of translation where the discriminative features of mediation become even more salient (Shlesinger and Ordan, 2012). In fact, interpreting was found to be the most distinctive kind of production compared to translated and written/spoken originals in computational studies using both hand-crafted features and information-theoretic measures (Lapshinova-Koltunski et al., 2021). Additionally, interpreting has been found to be more verbal than nominal (Przybyl et al., 2022b; Gast & Borges, 2023), and the unique features for interpreting are signals of spoken, online production. In this study, we explore variation in Greek mediated discourse, focusing on simplification and distinctiveness of interpreting and translation into Greek.

3. Methodology

3.1. Corpus compilation

EPICEL, as an intermodal corpus, contains i) original English transcribed speeches (ORG_SP_EN), ii) their transcribed interpretations into Greek (SI_EN_EL) and iii) their written translations into Greek (TR_EN_EL). For comparability with previous studies, the original English speeches selected are the English talks from TIC (Kajzer-Wietrzny, 2012), which are paired with their Greek interpretations. The paragraphs are manually aligned to create a parallel EN→EL SI corpus. The comparable sub-corpus containing the written

translations of the respective speeches is also aligned at the paragraph level and allows for comparative research into the properties of translation vs. interpreting. Since this study focuses on properties of translation vs. interpreting, we leave the study of ontology differences (original vs. translated/interpreting) to future work.¹

Transcription

The Greek interpreting videos were first automatically transcribed using the transcription function in Matesub². The transcription was then edited manually, following the transcription guidelines of EPTIC (Bernardini et al., 2018). As already mentioned, transcribing spoken data poses several challenges, related to the representation of spoken elements, such as pauses, mispronunciations, etc. Preserving spoken elements may allow for studying oral or cognitive factors but hampers with studying grammatic and syntactic factors. As in EPTIC, we prioritise the normalised version compared to an ‘as-is’ transcription (Defrancq et al., 2015) to facilitate the comparison with the translated texts. The intended form of the word is kept in the running text, while mispronunciations, filled pauses and interruptions are given in brackets. The choice of brackets for the spoken elements facilitates processing the text for turning these elements ‘on/off’ to obtain a clean version for further processing and analysis. Table 1 gives an overview of the transcription guidelines.

Feature	EPTIC	EPICEL
Silent pause	...	/
Filled pause	ehm	[ε], [α], [ο]
Mid-word pause	proposal /pro_posal/	πρόταση [πρό τηση]
Rising intonation	?	;
Non-verbalised noises	[applause]	[χειροκρότημα]
Non-standard pronunciation	proposal /prooposal/	πρόταση [πρό:ταση]
Mispronunciation	Parlamento / Parlomento	πρόταση [πράταση]
Truncated words	propo-	πρό-
Sentence-like segments	.	\n (new line)
Numerals	2004	2004
Foreign words/names	-	Latin spelling, e.g. Reding instead of Ρέντινγ

Table 1. Transcription guidelines in EPTIC and EPICEL.

The discrepancy for the symbol for silent pauses between EPTIC and EPICEL is to match the English (source) sub-corpus for which different transcription guidelines were applied. Still, a simple substitution would suffice to make the SI Greek sub-corpus fully compatible with EPTIC. Two examples of the transcribed text are shown in Table 2.

Example 1

¹ Studying the differences between interpreted and original spoken Greek discourse would require the compilation of a comparable corpus of original Greek speeches, including identification of speeches, collection of metadata, file management and transcription, which would not be justifiable given the time and space constraints of this thesis.

² matesub.com/

SI	ποιο είναι το θέμα [;] / ποιος είναι ο πυρήνας του ζητήματος [;] είναι Πρόεδρε ότι / στην Ευρωπαϊκή Ένωση / οι αρχές / δαπανούν για [για:] προϊόντα και υπηρεσίες πάρα πολλά χρήματα
TR	Οπότε, περί τίνος πρόκειται; Οπότε, η ουσία του ζητήματος είναι ότι οι δημόσιες αρχές στην Ευρωπαϊκή Ένωση ξοδεύουν τεράστια ποσά από το δημόσιο κορβανά για την αγορά προϊόντων και υπηρεσιών.
EN	so what am I talking about / well the core of the issue president is that public authorities in the european union <ehm> spend huge amounts of public money on buying products and services /
Example 2	
SI	δεν θέλουμε [ε:] / va- μάλλον αυτή τη στιγμή βρισκόμαστε εν τω μέσω τεράστιων επενδυτικών σχεδίων όπως οι [ε:] Ολυμπιακοί Αγώνες [ε:] του δύο χιλιάδες δώδεκα και λοιπά χωρίς όμως να δίνουμε δουλειά στους Βρετανούς εργαζόμενους /
TR	Φοβάμαι ότι τα πράγματα θα χειροτερέψουν πολύ περισσότερο. Καθώς ξεκινάμε ένα γύρο μαζικών δημόσιων δαπανών σε έργα, όπως οι Ολυμπιακοί αγώνες και η κατασκευή εργατικών κατοικιών, χιλιάδες ευρωπαίοι εργάτες είναι αποφασισμένοι να επωφεληθούν.
EN	[breath] I'm afraid it's set to get quite a lot worse because as we embark upon / a round of massive public spending projects such as the Olympics / [breath] and council house building / thousands of European workers are set to benefit / British jobs for British workers /

Table 2: Examples of the transcription in EPICEL, along with the official translated version and the English original transcription taken from TIC (Kajzer-Wietrzny, 2012). Elements of spoken discourse are in bold.

Finally, the transcription was split into sentence-like segments by the annotator based on syntactic and prosodic characteristics.

Corpus statistics

EPICEL contains the first 61 speeches of the TIC corpus, amounting to ~4 hours of original speech. The paragraphs are aligned across the three sub-corpora, allowing for cross-lingual analyses. However, achieving perfect parallel alignments at sentence level was not possible due to omissions and reordering in the interpreted texts. The translated texts were split at full stop marks. The corpus statistics are shown in Table 3. EPICEL corpus statisticsThe corpus can be accessed at <http://u.pc.cd/5WArtalK>.

	Speeches	Paragraphs	Sentences	Words
ORG_SP_EN	61	335	892	26,739
SI_EL_EN	61	335	864	19,451
TR_EL_EN	61	335	930	22,361

Table 3. EPICEL corpus statistics.

3.2. Analysis of interpretese vs. translationese features

Simplification

To answer the first research question, “Does interpreted discourse show higher simplification than translated discourse for Greek?”, we employ traditional corpus analysis measures of

lexical diversity, lexical density and mean length. For the results to be comparable across different corpora sizes and topics of the speeches, we use metrics of lexical diversity which have been shown to be insensitive to text length and lack of textual homogeneity.

- **Standardised Type-Token Ratio (STTR)**: Running average of the ratio of types (unique words) to tokens (total words) for every $n=1000$ words.
- **Hypergeometric distribution D (HDD)** McCarthy and Jarvis (2010): the probability of drawing (without replacement) a certain number of tokens of a particular type from a sample of size $n=42$.
- **Measure of lexical textual diversity (MTLD)** McCarthy and Jarvis (2010): mean length of sequential word strings in a text that maintain a given TTR value (here, TTR= 0.720).³

Lexical density measures the linguistic complexity based on the portion of function and content words in a text. The higher the number of content words, the higher the lexical density. We define lexical density Ld as follows:

$$Ld = \left(\frac{N_{lex}}{N} \right) \times 100$$

Where:

N_{lex} : Number of content words (nouns, verbs, adjectives, adverbs, numerals)

N : Total number of words

Lastly, simplification in terms of sequence length is measured with **Mean sentence length**, defined as the ratio of total words divided by the number of sentences in each sub-corpus.

The corpora were lower-cased and spoken elements (in brackets) as well as punctuation were removed. For obtaining the part-of-speech tags to compute lexical density, the corpora were POS-tagged with Universal Dependencies⁴ using the Stanza toolkit (Qi et al., 2020).

Distinctive linguistic features

For the second research question “What are the most distinctive linguistic features of translation vs. interpreting (and vice versa) into Greek?” we use a data-driven approach based on n -gram language models and Kullback-Leibler Divergence (KLD), following Karakanta et al. (2021). KLD, also called relative entropy, is a measure which shows how different one probability distribution is to a second one. In plain terms, KLD shows how much more surprising it would be to find a word typical for interpreting in a distribution of translated words. To give a concrete example, given a probability distribution of all single words (unigrams) in the interpreting corpus I and a probability distribution of all single words in the translation corpus T , KLD for encoding I with a model based on T is defined as follows:

$$D_{KL}(I||T) = \sum_{x \in X} I(x) \log \left(\frac{I(x)}{T(x)} \right)$$

³ HDD and MTLD were computed with <https://pypi.org/project/lexical-diversity/>

⁴ <https://universaldependencies.org/u/pos/>

Where:

I is the actual distribution

x is each individual unigram (single word)

X is the set of all the unigrams in the corpus

T is the reference distribution

KLD shows how much extra information (in bits) is required to encode the actual ‘interpreting distribution’ using a model optimised for the reference ‘translation distribution’. Since the total KLD value is the sum of the KLD of all individual words x , we know the contribution of each word x to the total KLD (pointwise KLD). In this way, we can detect the most typical words for each distribution, i.e., the words that contribute most to the distinction. KLD is asymmetric, which means that modelling translation based on interpreting will give a different value than modelling interpreting based on translation for the same unigram.

Contrary to frequency which was used in previous studies, KLD has the benefit of unravelling the distinctive words regardless of how often they appear in the corpus. For example, the word ‘and’ will be very frequent in both translation and interpreting. KLD will show ‘and’ as a typical word for one distribution only if it is distinctive of the actual distribution.

To summarise, KLD allows us to explore the following questions:

- Which lexical items are typical for translation but not interpreting?
- Which lexical items are typical for interpreting but not translation?

KLD is computed over the same sampling space, which means that the translation and interpreting distribution should have the same lexical items, which understandably is not the case. For this reason, add- e smoothing was applied. The missing words were added to the distribution with a very small probability ($e = 10^{-5}$) and e was also added to the probabilities of all existing words.⁵

To allow for an easy interpretation of the results, we visualise the 20 most distinctive words with word clouds⁶, where the size of the word is weighted based on their KLD value (the larger the more distinctive).

4. Results

4.1. Do interpreted texts from English into Greek show higher simplification than translated texts?

For the first research question, we compare the results of simplification metrics for translation and interpreting into Greek. We additionally report simplification results for the original English speeches. Table 4 shows the results for lexical diversity, lexical density and mean sentence length. For **lexical diversity**, all three measures show that translations are more lexically diverse than interpretations into Greek. This is in line with previous studies reporting that interpretations are less lexically varied than translations. English shows a much lower lexical diversity than both Greek sub-corpora. This is to be expected due to the

⁵ <http://hanj.cs.illinois.edu/cs412/bk3/KL-divergence.pdf>

⁶ <https://www.wordclouds.com/>

differences between languages and is possibly due to English being morphologically poorer than Greek. A similar tendency is observed for **lexical density**, with translations being the most lexically dense. The ratio of content to all words for interpreting is below 0.5, meaning that function words are the majority in interpreting. This ratio may be higher, for example, because pronouns are preferred instead of nouns in interpreting, or due to the use of particles. Still, these measures allow only for speculations, for which the exploration of the distinctive features of interpreting using KLD can shed light to this hypothesis. **Mean sentence length** is lower for interpreting confirming that shorter renditions are preferred or even unavoidable. Interpreting has on average 4 tokens less per sentence than translation. It is worth noting that English originals have the longest mean sentence length, possibly due to repetitions, figures of speech and other elements of institutional discourse, or simply because of language differences. All in all, the results confirm that interpreting shows higher simplification compared to translation for the language pair English into Greek.

	ORG_SP_EN	TR_EN_EL	SI_EN_EL
Hypergeometric distribution D (HDD)	0.86	0.89	0.88
Measure of lexical textual diversity (MLTD)	123.59	190.29	175.64
Standardised Type-Token Ratio (STTR)	0.46	0.54	0.51
Lexical density	0.493	0.501	0.490
Mean sent length	26.49	24.19	20.62
No of tokens	23627	22498	17817

Table 4: Simplification results for lexical diversity (HDD, MLTD, STTR), lexical density and mean sentence length for the three subcopora: English original speeches (ORG_SP_EN), translations into GREEK (TR_EN_EL) and interpretations into Greek (SI_EN_EL).

4.2. What are the most distinctive linguistic features of translation vs. interpreting (and vice versa) into Greek?

For the second research question, we inspect the lexical items that received the highest KLD scores for the two distributions, as shown in the word clouds. Figure 1: Lexical items typical of translation from interpreting and Figure 2: Lexical items typical of interpreting from translation.



Figure 1: Lexical items typical of translation



Figure 2: Lexical items typical of interpreting

For translation (Figure 1), inspection of the word cloud reveals that determiners (definite articles) and third-person weak pronouns such as *την*, *τη*, *της*, *του*, *των* and the indefinite feminine article *μια* (one) are among the most distinctive features.⁷ Determiners accompany nouns and especially the genitive form of the determiners suggests the presence of noun phrases in translation. The second category of distinctive features is prepositions, such as *σε*, *στην*, *με* (with), *από* (from/by), *σχετικά* (regarding), *λόγω* (due to), which form prepositional phrases combined with nouns. Finally, translations are distinguished from interpreting by subordination, shown by the items *καθώς* (since), *ότι* (that), *ωστόσο* (however), *εάν* (if) and the use of more formal conjunctions than interpreting (*γιατί* - because/why). All these features suggest more complex grammatic and syntactic structures, with subordinated clauses, prepositional phrases and nominalisation.

Interpreting, on the other hand, is distinguished by features of spoken language. The feature with the highest KLD value is the hesitation *ε*, which is to be expected since it does not appear in translations. Other items related to spoken discourse are *ευχαριστώ* (thank you) and *λοιπόν* (so). *Λοιπόν* is a coordinating conjunction which is very common in spoken discourse to conclude a topic or transition to a new one. Orality is also indicated by the distinctiveness of two alternative forms between translation and interpreting; interpreting prefers contraction *απ'* instead of *από*, and the stressed variety of the feminine indefinite article *μία* instead of *μια*. An oral, direct style is suggested also by the first-person pronoun in nominative *εγώ* (I). Strong pronouns *Αυτά*, *αυτές* (they) are distinctive of interpreting compared to weak ones in translation. Interpreters use pronouns to substitute or avoid repetition of nouns, which can be attributed to time constraints and management of cognitive effort. Common adverbs *πάρα* (very), *εδώ* (here) were also found to be distinctive. Adverbs possibly substitute prepositional

⁷ These words can be determiners or weak pronouns, depending on their function in the sentence. Since our approach is based on unigrams (single words), it is not possible to determine the class in this word cloud. However, annotating these words with their POS tag (e.g. *τον_PRO* or *τον_DET*) and re-computing KLD would allow to differentiate between pronouns and determiners.

phrases and create emphasis without the interpreter having to resort to longer renditions. Particles *να* (to-subjunctive), *θα* (will-future) as well as the frequent verbs *πρέπει* (must), *έχουμε* (we have) show a tendency towards verbalisation in interpreting, in line with Lapshinova-Koltunski et al. (2021) and Przybyl et al., (2022b). Lastly, coordination is preferred in interpreting, with the paratactic conjunction *και* (and) standing out as one of the most distinctive elements. This also confirms the success of our approach in disentangling distinctiveness from frequency effects.

5. Conclusions

In this thesis we explored linguistic variation in translation vs. interpreting in mediated discourse from English into Greek. Interpreting into Greek was found to be less lexically varied and dense than translation, and shorter renderings were observed, confirming the simplification hypothesis, at least for the current scenario. The analysis of typical words based on KLD confirmed previous studies suggesting a more nominal style for translation as suggested by genitive and prepositions, more formal language and subordination. Interpreting, on the contrary, demonstrates features of spoken language, more verbal style, as suggested by common verbs and particles, and a preference for coordinating structures. Other features of interpreting are the use of pronouns instead of nouns, and adverbs instead of prepositions. These features can be used in more traditional corpus-based studies to investigate specific interpreting strategies and inform interpreter training curricula for Greek. The thesis has opened directions for future work. During the transcription process, we noticed that there are large individual differences in choice of terms, stylistics and rate of condensation among interpreters. Upon completion of the compilation process, sound-to-text alignment (Chmiel et al., 2022) and interpreter identification (Korzinek and Chmiel, 2021) will be added to facilitate further analyses into individual interpreter differences and prosodic features of simultaneous interpreting into Greek. Moreover, it has been suggested that the type of source speech (prepared, semi-prepared, impromptu) influences the results, especially when it comes to simplification (Kajzer-Wietrzny, 2012). The actual size of the corpus did not allow for reliable results after separating the speeches, but an extension of the corpus would allow to replicate this study for the different types of speeches. This could be achieved with the inclusion of Greek in the EPTIC suite. Finally, the compilation of a comparable sub-corpus of Greek original speeches to study variation in original oral and interpreted discourse, to complete the picture of the features of mediated vs non-mediated discourse for Greek.

References

- Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.
- Bernardini, S., A. Ferraresi and M. Miličević (2016). From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28: 61-86.
- Bernardini, S., A. Ferraresi, M. Russo, C. Collard and B. Defrancq (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In Russo, M., C. Bendazzoli and B. Defrancq (eds.) *Making Way in Corpus-based Interpreting Studies* (pp. 21-42). Singapore: Springer Nature.
- Blum-Kulka, Shoshana (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* (pp. 17–35). Tübingen: Narr.
- Chmiel, A., Janikowski, P., Danijel, K., Lijewska, A., Kajzer-Wietrzny, M., and Jakubowski, D. (2021). Lexical frequency modulates current cognitive load, but triggers no spillover effect in interpreting. In *Proceedings of the 3rd International Conference on Translation, Interpreting and Cognition*, Forli, November.
- Defrancq, B., Plevoets, K., & Magnifico, C. (2015). Connective items in interpreting and translation: Where do they come from? In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics* (pp. 195–222). Springer International Publishing.
- Ferraresi, A. (2016). Intermodal corpora and the translation classroom: What can translation trainers and trainees learn from interpreting?. *Linguaculture* 2: 27-51.
- Ferraresi, A., & Miličević, M. (2017). Phraseological patterns in interpreting and translation. Similar or different? In G. De Sutter, M.-A. Lefer, and I. Delaere (eds), *Empirical Translation Studies. New Methodological and Theoretical Traditions*, volume 300 of Trends in Linguistics (pp. 157–182). Studies and Monographs [TiLSM]. Mouton de Gruyter.
- Ferraresi, A., Bernardini, S., Petrović, M. M. & Lefer, M.-A. (2018). Simplified or not Simplified? The Different Guises of Mediated English at the European Parliament. *Meta*, 63(3), 717–738. <https://doi.org/10.7202/1060170ar>
- Gast, V., & Borges, R. (2023). Nouns, Verbs and Other Parts of Speech in Translation and Interpreting: Evidence from English Speeches Made in the European Parliament and Their German Translations and Interpretations. *Languages*, 8(1), 39.
- Joseph, B. D., & Tserdanelis, G. (2003). Modern Greek. In T. Roelcke (Ed.), *Variationstypologie / Variation Typology: Ein sprachtypologisches Handbuch der europäischen Sprachen in Geschichte und Gegenwart / A Typological Handbook of European Languages* (pp. 823–836). De Gruyter. <https://doi.org/10.1515/9783110202021.5.823>
- Kajzer-Wietrzny, M. (2012). *Interpreting universals and interpreting style*. PhD thesis, Adam Mickiewicz University.
- Karakanta, A., Przybyl H. & Teich, E. (2021). Exploring variation in translation with probabilistic language models. In Julia Lavid-López, Carmen Maíz-Arévalo & Juan Rafael Zamorano-Mansilla (eds.), *Corpora in translation and contrastive research in the digital age. Recent advances and explorations* (pp. 308–323). Amsterdam/Philadelphia: John Benjamins.

- Lapshinova-Koltunski, E., Bizzoni, Y., Przybyl, H., & Teich, E. (2021). Found in translation/interpreting: Combining data-driven and supervised methods to analyze cross-linguistically mediated communication. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age* (pp. 82–90). Association for Computational Linguistics.
- Laviosa, S. (1998a). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43(4), 557–570.
- Laviosa, S. (1998b). The English Comparable Corpus: A Resource and a Methodology. In L. Bowker, M. Cronin, D. Kenny, & J. Pearson (Eds.), *In Unity in Diversity? Current Trends in Translation Studies* (pp. 101–112). St. Jerome.
- McCarthy, P. M., & Jarvis, S. (2010). MTL-D, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2), 381–392. <https://link.springer.com/content/pdf/10.3758/BRM.42.2.381.pdf>
- Pan, J. (2019). The Chinese/English Political Interpreting Corpus (CEPIC): A New Electronic Resource for Translators and Interpreters. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)* (pp 82–88). Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.
- Przybyl, H., Lapshinova-Koltunski, E., Menzel, K., Fischer, S., & Teich, E. (2022a). EPIC-UdS: Creation and applications of a simultaneous interpreting corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 1193–1200). European Language Resources Association (ELRA).
- Przybyl, H., Karakanta, A., Menzel, K. & Teich, E. (2022b). Exploring linguistic variation in mediated discourse: Translation vs. interpreting. In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska, & S. Bernardini (Eds.), *Mediated discourse at the European Parliament: Empirical investigations* (pp. 191–218). Language Science Press. DOI: 10.5281/zenodo.697705
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*.
- Russo, M., Bendazzoli, C., & Sandrelli, A. (2006). Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: Extended analysis of EPIC. *Forum*, 4(1), 221–254.
- Sandrelli, A., & Bendazzoli, C. (2005). Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). In *Proceedings from the Corpus Linguistics Conference Series 1*. University of Birmingham, Birmingham.
- Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43 (4): 486–493.
- Shlesinger, M., & Ordan, N. (2012). More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target*, 24(1), 43–60.
- Shlesinger, M. (2009). Towards a definition of interpretese: An intermodal, corpus-based study. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile* (pp. 237–253). John Benjamins.