Research article
Submitted: 1 May 2024
Accepted: 15 Sept 2024

TERMINOLOGICAL CONSISTENCY IN CROATIAN TRANSLATIONS OF EU LEGISLATION: A CORPUS-BASED STUDY

Laura Marić, University of Zagreb

Abstract

The EU produces thousands of translations, with particular emphasis on terminological consistency as it can impact clarity and legal certainty. However, the fast-paced, "multitranslator" environment raises the question of how consistently terminology is used. The aim of this study is to check for inconsistencies in the use of trade-related terminology in Croatian translations of EU legislation, and to examine the correlation between the frequency and structure of terms and their consistency. Terminological consistency is measured using the HHI, following the example of Itagaki et al. (2007). Trade-related terms are taken from IATE and analysed on a corpus compiling DGT's English-Croatian translation memories from 2020. The results confirm the presence of terminological inconsistency and show a weak positive correlation between consistency and frequency, and no correlation between consistency and structure. The study also highlights the usefulness of EU materials and the HHI for linguistic and terminological research.

Key words: terminology, terminological consistency, European Union, EU translation, Herfindahl–Hirschman Index

1. Introduction

The European Union has become the most multilingual body of institutions in the world. According to data provided by European Commission (2023), more than 2.5 million pages get translated every year, with legislation making up 54% of that amount in 2023. This includes translations into all 24 official languages of

the EU owing to its language policy[1]. To ensure terminological consistency across all these languages and documents, significant effort has been put into terminological work, with one of the biggest milestones being the introduction of the EU's multilingual terminology database, *Interactive Terminology for Europe* (IATE) in 2004. However, research suggests that there might still be terminological inconsistencies, especially in Croatian translations, as the *acquis*, i.e. the complete body of EU law, was infamously translated "under pressure" and "by numerous translators of various degrees of expertise and experience" (Bratanić and Lončar, 2016, 210). That, together with the fast-paced, "multitranslator" environment of EU translation raises questions about how consistently terminology is used and translated. The aim of this study is, therefore, to check for inconsistencies in the use of terminology in Croatian translations of EU legislation and, further, to determine whether there is a correlation between the frequency and structure of a term and its terminological consistency where inconsistencies are detected. The scope was reduced to trade-related terminology only. The study was conducted on a corpus compiling translation memories comprising texts from 2020, made available by the European Commission's Directorate-General for Translation. The analysed terms were extracted from IATE. Finally, terminological consistency was measured using the Herfindahl–Hirschman Index (HHI) as proposed by Itagaki et al. (2007).

Section 2 provides an overview of previous research related to the topic and Section 3 presents definitions of key concepts. Section 4 outlines the aims and hypotheses, while Section 5 describes the methodology and resources. Finally, in Sections 6 and 7 the study's findings are reported and further discussed.

2. Previous research

In Croatia, terminology and terminological consistency in Croatian legal translations gained attention in the scholarly sphere when Croatia was preparing to accede to the EU. One of the works discussing the pre-accession state of Croatian legal terminology was *Hrvatski jezik na putu u EU* edited by Maja Bratanić (2011). It comprises articles about various terminological issues, from

[1] https://www.europarl.europa.eu/factsheets/en/sheet/142/language-policy

term formation to term standardisation. In one contribution, Bajčić and Stepanić (2011) highlighted the inconsistent use of competition law terminology. Their study, conducted on a number of English-Croatian legal translations and terminological resources, such as *Eurovoc* and *Euroterm*, found that several competition law terms were inconsistently translated. The authors underlined the importance of terminological consistency in legal translation, advocating for more cooperation between domain experts, terminologists and translators in the then upcoming translation of EU legislation. Although the resources used in the study are less relevant today, because Croatia has since joined the EU and its terminology and translation framework, its results illustrate that the problem of terminological inconsistency was present in Croatian legal translation even before the accession, which probably influenced the consistency in later translations.

Bratanić and Lončar (2016) also examined the myth of terminological consistency in the EU on the example of the Croatian translation of the *acquis*, documents that constitute the body of EU law. The study exemplified instances of terminological inconsistency, aiming to provide an overview of both linguistic and extra-linguistic reasons behind it. The linguistic reasons mostly included unclear definitions of terms, while the extra-linguistic ones included relatively adverse circumstances in which the *acquis* was translated.

A more international perspective on the EU's terminological work and consistency is given in the works of Stefaniak (2017) and Pozzo (2020). Stefaniak (2017) gave a detailed overview of its terminological process and described most common terminological problems in translation. She pointed out the importance of terminological consistency and the consequences inconsistency can have in the legal context of the EU. Pozzo (2020) analysed how multilingualism has impacted the harmonisation process of European private law. She found that despite the EU's efforts to harmonise private law terminology, inconsistency was still present, both at a monolingual level and in translations. Although there is more research concerning the terminology process in the EU, very few studies observe consistency using an example-based, empirical approach. Moreover, all aforementioned studies observed terminological inconsistency qualitatively.

However, Gašpar (2013) and later Gašpar et al. (2022) proved that, by using the HHI method first introduced by Itagaki et al. (2007), terminological

consistency can also be quantitatively assessed. The method, which utilizes the Herfindahl–Hirschman Index, commonly used to measure market concentration is more thoroughly explained in the following sections. The first study (Gašpar, 2013) assessed terminological consistency in a Croatian-English legal parallel corpus which showed an expected low index of consistency in Croatian-English legal translations and supported the implementation of the HHI for the measurement of terminological consistency. In the second study, Gašpar et al. (2022) employed the same method, but on a bigger corpus. The corpus consisted of three types of legal subcorpora, Croatian-English parallel corpus (1991–2009), Latin-English and Latin-Croatian versions of the Code of Canon Law (1983), and the English and Croatian versions of the EU legislation (2013–). The results confirmed the presence of inconsistency in all corpora, with the consistency index being higher for the English-Croatian language pairs. They also showed a diachronic increase in consistency and supported the implementation of HHI on the Croatian-English language pair. However, the study had certain limitations due to the small terminology dataset. Only the most frequent terms from each corpus were analysed, with the numbers of terms per corpus being the following: Croatian-English parallel corpus, 100; Canon Law corpus, 25; EU legislation corpus, 15. The limitations leave the question of how representative the consistency index was as a whole, since there was not a great distribution in terms of frequency. This study tries to bring more insight into this issue. It should also provide a more recent assessment of the state of terminological consistency in Croatian translations of EU legislation.

3. Key concepts

3.1 Terminology *and* term

Both terminology as a discipline and the term *term* have been defined and redefined by many scholars since the publication of Wüster's seminal work *General Theory of Terminology* in 1979, which set the groundwork for the development of the discipline. Traditionally, terminology has been defined as "the study of and the field of activity concerned with the collection, description, processing and presentation of terms" (Sager, 1990, 2). It is also often used to refer to "internally consistent and coherent set of terms belonging to a single subject field" (Sager, 1990, 3). It plays a crucial role in specialized fields,

including law, as it provides a framework that establishes connections between specific *concepts,* or ideas, and their lexical counterparts, or *terms,* as well as maps relationships between those concepts. That connection is formed primarily through a definition which gives a precise description and reference of the concept. It is further strengthened by its relationship to other concepts in the same domain. The definition can also be complemented by other morphological, syntactic and pragmatic specifications, such as information about context and usage (Sager, 1990, 21-40). As follows, *terms* can be defined as "items which are characterised by special reference within a discipline" (Sager, 1990, 19), as opposed to *words*, which are linguistic units with a general reference in a language. Since the meaning and function of terms is defined and confined by the domain they exist in, they can also be described as "a functional class of lexical units" (Sager, 1998, as cited in Kockaert and Steurs, 2015, 48).

In translation, as explained by Fischer (2022), terms can also be evaluated in a broader sense, encompassing any lexical unit which is expected to be translated in a specific way, i.e. any word, phrase or sentence that restricts the translator's freedom. She also insists that such approach should be followed in all discussions surrounding EU translation, since EU's terminology work is largely influenced by translation (Fischer, 2022). Moreover, the application of these principles can be noticed in the criteria for inclusion of terms in IATE, as it, in addition to terms, comprises special expressions that are useful in translation, but that would not be considered terms in the traditional sense (Fischer, 2010). Furthermore, one of the main characteristics of terms in traditional terminology is univocity, designating both monosemy (one concept per term) and mononymy (one term per concept) (Temmerman, 2000, 10). That means that each concept can be signified by only one term, and in turn that term cannot be used to refer to another concept. This implies the elimination of term synonymy and polysemy as well as terminological variation.

This view has, however, been challenged by a more recent sociocognitive terminology theory developed by Rita Temmerman. Temmerman (2000) based her approach on frameworks proposed by sociolinguistics and cognitive linguistics. She describes terms as prototypical in nature, with their categorization being based on similarity, not defining characteristics, and their structure being less delineated and more malleable (Temmerman, 2020, 63-66).

Likewise, concepts are replaced by less restrictive and mostly prototypical *units of understanding*, emphasizing the cognitive dimension they function in. In this view, synonymy, polysemy and terminological variation are consequently accepted as functional, as they express changes in meaning and show different perspectives; "[c]ategories evolve, terms change in meaning, understanding develops" (Temmerman, 2020, 16). In the context of EU translation, the sociocognitive approach can nevertheless be hard to accept because of its implications for law interpretation and legal certainty[2], even though it might provide a better insight into the ways in which we process and understand terminology. As Bratanić and Lončar (2016) point out, due to the multilingual, multicultural and "multilegal" context that EU translation is situated in translators already have to face many challenges to ensure the uniformity of law interpretation across all languages, as misinterpretations can have serious legal consequences. That uniformity hinges on unambiguity and clarity which could be even harder to achieve in the presence of terminological variation, which is why in such cases "a more traditional approach to term harmonisation and standardisation should still be at the forefront of the discussion" (Bratanić and Lončar, 2016, 217). Following these arguments, this study draws on the traditional definitions of terminology and terms and gives precedence to terminological consistency over terminological variation.

3.2 Terminological consistency *and* variation

Although it is a characteristic of both monolingual and translated specialized texts, *terminological consistency* is more commonly observed in translations and, in this study, it will only be observed from that perspective. *Terminological consistency* can, therefore, be defined as the use of one and the same translation equivalent, or terminological variant, for a given source term (Gašpar, 2013, 1). Additionally, consistency does not only refer to the use of "the same term for the same referent throughout a particular communication" (Rogers, 2008, 107), but also "throughout all communications within a particular organisation if a terminology policy is in place" (Rogers, 2008, 107). That is certainly the case in the EU.

[2] Legal certainty is a principle that "rules should be clear and precise, so that individuals may be able to ascertain unequivocally what their rights and obligations are and may take steps accordingly" (Craig, 2012, 549).

On the other hand, *terminological variation* refers to a broader phenomenon, resulting from term polysemy and synonymy, "where the same concepts can be expressed differently" (L'Homme, 2023, 153). ISO 1087 standard defines synonymy as a "relation in which a designation represents two or more concepts" and polysemy as "a relation between designations in a given natural language representing the same concept" (2019, 11). As follows, terminological variation in translation can be understood as one source term having two or more *terminological variants* in the target language. Terminological variants are most frequently defined as variants of a term which are "semantically and conceptually related to an original term" (Daille et al. 1996, 201, as cited in Biel, 2023, 92). Therefore, they cover instances of both conceptual sameness (*denominative variants*) and similarity (*conceptual variants*). They can be further categorized according to formal and conceptual distance to the base term, time, acceptability and distribution (Biel, 2023). In the context in which an ideal of consistency is present, terminological variation is replaced by *terminological inconsistency*. Inconsistency highlights the enforcement of a standard of univocity, as any instance of variation is considered to be an unwanted divergence from that standard, i.e. the consistent use of one chosen term.

Finally, terminological consistency can be analysed quantitatively and qualitatively, and in this study the former approach was employed. The method was first proposed by Itagaki et al. (2007), who utilized the Herfindahl–Hirschman Index (HHI), commonly used to measure market concentration to measure terminological consistency in a given text. The HHI is usually calculated using the following formula:

$$HHI = \sum_{i=1}^{n} {s_i}^2$$

where *S* indicates the market share of a firm in the market, and *n* is the number of firms. If the index is 10000, or 100^2 that means one firm dominates the market. When applied to translation, Itagaki et al. (2007) explain that "*S* becomes the ratio of each translation (*i*) to the total number of translations (*n*) within a product" (5). The index was later applied by Gašpar (2013) and Gašpar et al. (2022) in studies on Croatian-English and English-Croatian corpora, using the following adapted formula:

$$C_t = \sum_{n=1}^{n} \left(\frac{f}{k} \times 100\right)_i^2$$

The calculation and application of the HHI index in this study is further explained in Section 5.

4. Aims and hypotheses

The main aim of this study was to check for inconsistencies in the use of terminology in Croatian translations of EU legislation. Due to its small scope, the observed terminology was limited to trade-related terms only. Furthermore, where inconsistencies were detected, the second aim was to determine whether there is a correlation between a term's frequency and its structure (i.e., the number of words it consists of) on the one hand and how consistently it is translated on the other. The following hypotheses were tested:

H1 There are inconsistencies in the use of trade terminology in Croatian translations of EU legislation.

H2 Terms that have a higher frequency, i.e. those that are used more often, are translated more consistently.

H3 Longer terms, i.e. those that consist of more words are translated less consistently.

The basis for the first hypothesis was the previously discussed expectation that terminology might not be consistently translated based on the context of EU translation and the suggestions of previous studies (e.g. Gašpar, 2022). That hypothesis also lays the foundation for the other two hypotheses: they can be tested only if it is accepted, i.e. if inconsistencies are found.

Furthermore, it is expected that more frequent terms will be translated more consistently because they have more established terminological equivalents since they are used more often. Another aspect that also plays a role in this correlation is the use of, and reliance on translation memories (TMs). All EU translators work with *Euramis*, the EU's central translation memory, which automatically retrieves similar segments that it recognizes as useful for the translation of the new document. This not only allows for the translations to be produced at a faster rate, but also helps ensure consistency across all documents. In 2016, Euramis contained over 1 billion segments across all official EU languages (European

Commision, 2016). Owning to this fact, it is presumed that terms that are more frequent will be present more often in the TMs and automatically be translated more consistently, as opposed to less frequent terms which might not be present in the TMs and for which the translator will have to do extra research to find or create an appropriate equivalent. More frequent terms are usually also better known, so the translator might know their equivalents even without double-checking the term base or the TM.

In the third hypothesis it is expected that a term's number of words might pose a challenge to translators in some contexts, especially when it comes to longer terms. If there are inconsistencies, maybe they arose because the intended equivalent made the sentence less readable and understandable, or the translator did not recognize the whole phrase as a term and translated it only partially consistently.

Initially, the fourth hypothesis was supposed to examine if terminological consistency correlated with a term's part of speech. Yet, analysis of a random sample revealed only one verb, one adjective, and the remaining terms were nouns or noun phrases. Due to insufficient data, this hypothesis was then dismissed.

5. Methodology[3]

5.1 Corpus compilation

Since this is a corpus-based study, the first step regarding data collection was finding or compiling a relevant corpus. There are several pre-existing EU legislative corpora that were considered first. The most recent public corpus is *EUR-Lex 2/2016 parallel,* a parallel corpus with multilingual subcorpora in all official languages of the European Union. It can be accessed through Sketch Engine, and it compiles European Union law and other public documents up until 2016 that are available in EUR-Lex, the online database of EU legal documents.[4]*The Digital Corpus of the European Parliament*[5] or *DCEP* is another

[3] I would like to thank Prof. Mateusz-Milan Stanojević for his invaluable guidance on the methodological approach, particularly corpus tools and analysis, as well as Prof. Mirjana Tonković for advice regarding the statistical analysis.
[4] https://eur-lex.europa.eu/homepage.html
[5] https://joint-research-centre.ec.europa.eu/language-technology-resources/dcep-digital-corpus-european-parliament_en

publicly available corpus comprising documents, including legislative documents, published on the European Parliament's official website between 2001 and 2012. Although these corpora are the most recent publicly accessible corpora of EU legislation, their data, i.e. texts published up until 2012 and 2016, were deemed too dated for this study, especially since Croatia joined the EU in 2013.

The final corpus that was considered is *DGT Translation Memory parallel corpus* in 24 official EU languages, including Croatian. It is also available on Sketch Engine; however, there is no clear indication of when it was created, which texts it compiles and from which time period. Therefore, a new, more relevant corpus had to be created. This was done using TMs made public by the European Commission's Directorate-General for Translation on their website[6]. These TMs consist of parallel texts from the *acquis communautaire* and some other texts, in all 24 official EU languages. The *acquis* comprises all treaties, regulations and directives adopted by the European Union, or in other words the EU legislation. According to the European Commission (n.d.), the texts were aligned in accordance with the DGT's segmentation rules and were pre-processed "to reduce the number of entries of low value for the translators (short sentences, long sentences, obvious mismatches, etc.)". There are several versions of the TMs depending on the year they were released. Since this study has a smaller scope, the corpus compiles only the most recently released TMs with parallel texts from 2020. As mentioned, they are available in all 24 official EU languages, so to acquire the English-Croatian TMs, a bilingual extraction was performed using the extraction tool TMXtract made available by the DGT.

Finally, TMs consisting of only English-Croatian parallel texts from 2020 were uploaded to Sketch Engine, a web-based corpus tool which offers features like corpus creation, automatic lemmatization and tagging for parts of speech, term extraction etc. The resulting corpus consists of two parallel corpora, with the English one compiling 7,170,658 words and the Croatian one 6,445,680 words.

5.2 Term list

The second step consisted of compiling a term list with terms to be examined in the study. EU legislation covers a wide range of domains and areas of human

[6] https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en#dgt-memory

activity, which has led to the development of very diverse terminology in the EU. To improve terminology management, the EU has created IATE, its own term base which is available as an online tool and can also be downloaded in multiple formats and all official EU languages. On IATE there is a filtering option that lists all 22 domains the EU terminology can be categorized into, ranging from law and economics to industry and energy. Since this is a small-scale study, the scope was reduced to analysing only trade-related EU terminology. Trade was chosen since is one of the main domains of EU legislation, so it is the subject of many legislative documents and therefore has a well-developed and comprehensive terminology. It should be noted though that the domain variable could also be an interesting point in further research regarding terminological consistency, since the nature of the domain and the way its terminology is dealt with could influence its consistency in translation. However, this correlation will not be explored in this study.

As mentioned, terms in IATE are categorized by domain, which made it easier to compile a list of only trade-related terms for the analysis. In general, IATE comprises the majority of English terms used in EU documents and provides a number of translation equivalents in official languages. It is concept-based, meaning that each entry should correspond to a single concept. Furthermore, since its main purpose is to "facilitate multilingual drafting and translating of EU legal texts" (Stefaniak, 2017, 111), it does not only contain terminology, but also proper names, titles of documents and agreements, abbreviations and a number of phrases that could not necessarily be considered terms, but that occur often in EU texts and should be uniformly translated. However, IATE still has a number of downfalls, as pointed out by Stefaniak (2017) and Bratanić and Lončar (2016), such as the varying quality of entries, with some containing little to no information or many entries having low reliability as evaluated by IATE's own reliability system. The Croatian terminological network on IATE can especially be lacking at times, since a significant number of English terms still do not have a listed Croatian equivalent. Nonetheless, this does not affect the results of this study since it only observes the translation equivalents of terms, or its variants, in the TMs. This is also why the term list is monolingual, i.e. contains only terms in English.

The base of the term list was created by downloading the English IATE term

base for the trade domain in .xsl format. That list was then manually edited so that it is accessible and ready for further processing, which included separating entries that were initially in the same cell in Excel, so that each respective entry can be properly recognized by Excel, as well as deleting terms that are abbreviations and Latin terms. Those entries are not of interest for this study since they are always translated in the same way, so there is little to no room for possible inconsistencies apart from translator's lack of attention. Since the list comprised thousands of terms which cannot all be guaranteed to be found in the study's corpus, the next step was to do an automatic term extraction from the English part of the parallel corpus in Sketch Engine. The extraction resulted in a list of over 150,000 single- and multi-word units that Sketch Engine recognized as terms and as expected there were "instances of both noise (non-pertinent items identified) and silence (relevant terms missed)" (Bowker 2015, 310). For this reason, Bowker (2015) emphasizes the importance of manually editing the list, usually done by a domain expert. In this study, the IATE term list was used as reference for the validity of the recognized terms instead. The automatically generated list of terms was compared to the IATE term list in Excel to determine which of the IATE terms are present in the corpus. The final term list consisted of 909 English terms (see Appendix A), and then a random sample of one hundred terms was taken using Excel. The random sample was supposed to provide a more even distribution of terms with different frequencies and structures, and possible inconsistencies.

Due to IATE's aforementioned varying reliability and quality of listed entries, there are limitations to this study, as the validity of certain terms can be questioned. The reliability values on IATE are reflected through a star rating system, with entries having 3 or 4 stars being manually verified and considered reliable and very reliable, respectively, and 1 or 2 stars indicating unverified and low reliability. Some terms also have no listed definition or other pragmatic information. That might weaken their reliability because, as explained in Section 3, a term's meaning, or its connection to the concept it describes is established through a definition (Sager, 1990, 21). The lack of definition can also lead to different interpretations of the term's reference, especially if observed in the context of different national legal systems (Ferrari, 2010, as cited in Pozzo, 2020). It is also interesting to note that many terms are listed in IATE as

separate entries under different domains, and some of them have a listed definition under only one of the entries. For example, replacement certificate has five separate entries, one under finance, one under international relations and three under trade, with two of them having low reliability value, and none of them having a definition. Moreover, some terms cannot be found on IATE's online search tool and only exist in the downloadable term base. Out of 100 terms analysed in this study, 22 terms have a satisfactory reliability value, but no listed definition, six terms have an unsatisfactory reliability value, and one is not listed on the online database. The remaining 71 terms fit all the reliability criteria, having both a listed definition and a high reliability score.

5.3 Analysis

The process of gathering data for the analysis was corpus-based. In other terms, it involved searching for the occurrences of the terms in the corpus and then recording the relevant data which included the frequency of a term in the corpus, the number of words a term consists of, and its translation equivalents present in the corpus, i.e. its terminological variants in the target language. When it comes to frequency, two types of frequency could have been used: a term's relative frequency in the corpus, or its absolute frequency. Relative frequency shows the relation between the number of occurrences of a term and the total number of tokens, or words, in the corpus. It is usually used to compare frequencies between corpora of different sizes. Absolute frequency is just the number of individual occurrences, or hits, in the corpus, which is why it is also referred to as raw frequency. For example, the relative frequency of *trade committee* is 0.001878%, or 18.78 per million tokens, while its absolute frequency is 165. This study observes and uses only the absolute frequency for the following reasons. Firstly, since the study is conducted on only one corpus, its size is not an important factor that would have had to be taken into consideration if this were a multi-corpus study, in which case the relative frequency would have been a better representation of a term's frequency. Secondly, not all hits in the corpus contained only the relevant term. The absolute frequency of certain terms was lower than indicated in the corpus, which meant the relative frequency would also have to be recalculated. This calculation was essentially much harder to do in comparison to the simple manual adjustment of the absolute frequency. Finally, absolute frequency was also needed to calculate the consistency index,

or the HHI, so it was decided that only the absolute frequency and its correlation to terminological consistency would be analysed in this study.

As mentioned, the corpus search doesn't always yield results containing only relevant term occurrences. Therefore, a set of criteria for the exclusion of certain results was laid down. New terms are often created by principle of recursion, i.e. by taking established terms and combining them into new phrases with different meanings, as exemplified by terms *trade → trade policy → Trade Policy Committee.* Because of this phenomenon, when searching for certain terms in the corpus, especially single-word ones, it generates results that essentially contain a different, expanded term. In the context of EU legislation, that also frequently happens with official documents or agreements that contain terms in their names and are considered separate terms themselves. All of these instances had to be taken note of regarding the absolute frequency of terms, since any instances where the resulting corpus example contained what could be considered a separate term had to be eliminated from the analysis. For example, 43 instances of the term *international trade* were eliminated due to the term being a part of other terms like *international trade rules,* or *Convention on International Trade in Endangered Species of Wild Fauna and Flora*. The other exclusion criteria concerned the translation of the term, namely the instances where the term was transposed, i.e. replaced with a different word class, or even excluded from the translation. The exclusion criterion was especially important because of the calculation of the HHI. The HHI for each variant is a ratio of its absolute frequency and the absolute frequency of all terminological variants found in the translation. If the examples where the term was excluded in translation were counted towards a term's absolute frequency, the ratio would be skewed, and the HHI would consequently be inaccurate. In addition to data about overall frequency of the term, the number of its terminological variants was recorded, as well as the variants' form and frequency. To test the third hypothesis, each term was analysed on the level of structure, i.e. how many words it is made up of, and whether it is a single- or a multi-word unit. As mentioned, there was an intention to analyse the correlation between a term's type of phrase or part of speech and its consistency, but the data to test it was insufficient. Therefore, to control that variable, those terms were replaced by the next two randomly sampled terms that were nouns or noun phrases.

The last step of the analysis was to calculate the consistency index for every term. As explained in Section 2, Herfindahl Hirschman Index is a measure of market concentration in economics. However, it was introduced into terminological research by Itagaki et al. (2007) as a way to measure and automatically validate terminological consistency which was up until then evaluated only qualitatively. Their study was focused on terminology in localized materials, like manuals, and in training of example-based and statistical MT systems. The authors adapted the HHI formula to fit the context of translation and terminology which resulted in the following formula:

$$C_t = \frac{\sum^{p} \sum_{n=1}^{n} (\frac{f}{k} \text{x} 100)_i^2}{p}$$

C is the consistency index for a specific term (*t*), *p* is the number of texts that contain the term, *f* is the absolute frequency of a particular translation variant, and *k* is the total number of occurrences of the term, or a sum of absolute frequencies of all variants, within a text or corpus (Itagaki et al. 2007, 5). This formula, especially the *p* variable, was pertinent to their methodology and aims, because the study analysed terminological consistency across multiple groups of texts belonging to different products. When Gašpar (2013) applied this method to assess the terminological consistency of translated terms in a Croatian-English parallel corpus of legislative texts, she further adapted the formula by removing the *p* variable to calculate the HHI score for individual terms. The adapted formula was as follows:

$$C_t = \sum_{n=1}^{n} \left(\frac{f}{k} \times 100\right)_i^2$$

It can be said that the consistency index of a particular term is the sum of the consistency indexes of all its respective variants found in a text or corpus. The "frequency share" for each variant is calculated as a ratio of its absolute frequency and the total occurrence of all variants of that term. Gašpar et al. (2022) applied this formula again on an expanded range of corpora, this time including also Latin-English and Latin-Croatian versions of the Code of Canon Law (1983), and the English and Croatian versions of the EU legislation (2013-). Both studies confirmed that HHI as a measure of terminological consistency can be successfully applied to Croatian-English and English-Croatian legal

translations. Therefore, the aforementioned formula to calculate the HHI will be used in this study as well. Table 1. shows an example of an HHI calculation. The final values of the HHI, seen in the far right column of Table 1., were normalized to a range of 0-100, with 0 marking complete terminological inconsistency, i.e. the term being translated differently in every instance, and 100 marking complete terminological consistency, i.e. the term being translated using the same terminological variant in every instance.

Table 1. HHI calculation for the term *price suppression*

SOURCE TERM	VARIANTS	FREQUENCY	$\left(\frac{f}{k}\times 100\right)$	$\sum_{n=1}^{n}\left(\frac{f}{k}\times 100\right)_i^2$
price suppression	sprečavanje rasta cijena	15	4253.31	4593.56
	smanjenje cijena	2	7.61	
	pritisak na cijene	3	170.13	
	pad cijena	1	18.90	
	sniženje cijena	2	7.61	

Although the number of terminological variants indicates the presence of terminological inconsistency, it is actually the ratio of these variants' frequency that affects the index the most. For example, *subtotal* had two terminological variants, *međuzbroj* and *ukupno*, and its HHI was 71.18, while *import price* had four terminological variants *uvozna cijena, cijena uvoza*, *izvozna cijena* and *obujam uvoza* and its HHI was higher, at 87.16. This is because the two variants for *subtotal* were relatively evenly distributed, while *uvozna cijena* was the clearly dominant variant for *import price.* It follows that the consistency index evaluates terms with multiple variants out of which one is dominant as "more consistent" than terms with fewer variants that are equally distributed.

Finally, due to the skewness of the HHI distribution, another value was added to each term for statistical analysis purposes. Their HHI scores were ranked on an ordinal scale as shown in Table 2., and each term got assigned an ordinal value. For the final spreadsheet with all data for all 100 terms, see Appendix B.

Table 2. HHI scores with assigned ordinal values

HHI SCORE	ORDINAL VALUE
0.00-19.99	1
20.00-39.99	2
40.00-59.99	3

60.00-79.99	4
80.00-100.00	5

Statistical analysis and correlation tests were performed using JASP (version 0.17.2.0), an open-source program for statistical analysis. Since the distributions of all three variables (frequency, number of words, HHI score) were skewed, as seen in Figures 1., 2. and 3., Spearman's rank correlation, a non-parametric correlation test, was employed for the analysis.

6. Results

The following sections present the results of the statistical analysis. The first subsection gives an overview of descriptive results for each of the variables used in the correlation tests. The second and third subsections report the results of correlation tests regarding terminological consistency and frequency, and term structure, respectively.

6.1 Descriptive results

6.1.1 Frequency

First variable that was analysed was frequency. As previously discussed, only absolute frequencies of terms were recorded. Frequency distribution was right-skewed (see Figure 1.). The median frequency was 11.5 (IQR = 5-29). The term with the highest frequency was *consignment* with 1645 occurrences, followed by *trade* (uncountable noun) with 1304 occurrences. *Middle-value contract* had the lowest frequency, occurring only twice in the corpus.

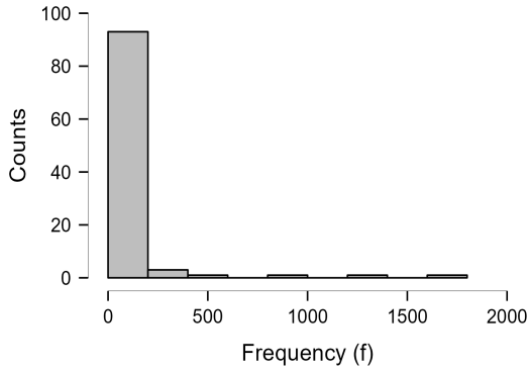

Figure 1. Frequency distribution

6.1.2 Number of words

The distribution of a term's number of words was slightly right-skewed as well, as shown in Figure 2. Out of 100 terms, four terms were single-word units and the remaining 96 were multi-word units; sixteen terms consisted of three words; eight of four words and two of five words. The most terms, 70 of them, consisted of two words. The median was 2 (IQR = 2-3).

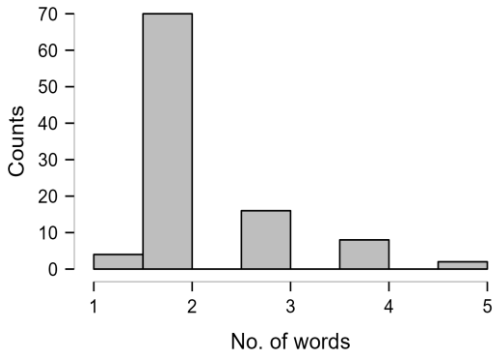

Figure 2. Number of words distribution

6.1.3 Herfindahl Hirschman Index score

The distribution of the HHI scores was left-skewed (see Figure 3.), with 51 terms having optimal scores of 100. This means that the overall consistency was fairly high, with the median score being 100.00 (IQR = 63.31-100.00). The term with the lowest consistency of 33.33 was *specific contract*, followed by *corporate entity* (C_t = 33.56) and *supply contract* (C_t = 42.15). However, the terms with the most terminological variants (5) were *price suppression* (C_t = 45.94) and *trade* (C_t = 96.23).

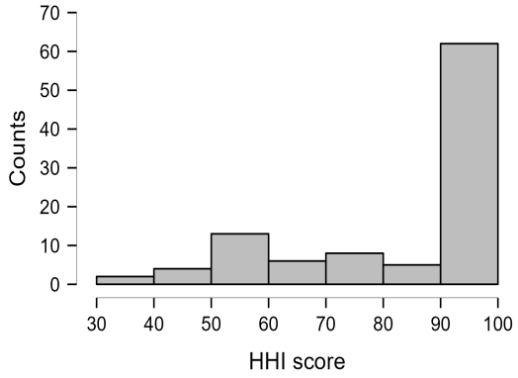

Figure 3. HHI score distribution

6.2 Correlation between frequency and terminological consistency

As explained in Section 5, Spearman's rank correlation coefficient, or Spearman's rho (ρ) was computed to assess the correlations. Additionally, the aforementioned ordinal HHI variable, as opposed to the HHI score, was used as a measure of consistency for both correlations. The results of the correlation test, as seen in Figure 4., report a weak positive correlation between frequency and ordinal HHI score, $\rho(98) = .32$, $p < .001$. This means that more frequent terms tend to have a higher HHI score, or rather are translated more consistently. The effect size, as measured by Spearman's rho, indicates a medium effect (Goss-Sampson 2022, 41).

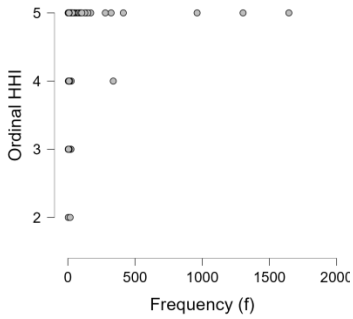

Figure 4. Scatterplot of correlation between frequency and ordinal HHI score

As the reliability of analysed terms was discussed, the correlation test was furthermore performed on two other sets of data. The first one excluded the seven terms with low reliability scores in IATE, and the second one excluded those seven, as well as the 22 terms that had no listed definition on IATE (see Appendix C). These tests were performed to check if there was a probability that the reliability of terms on IATE could affect the correlation in any way. After the elimination of the seven unreliable terms, the rank correlation coefficient only slightly changed, $\rho(98) = .33$, $p < .001$. However, the correlation coefficient for the second set of data including only the 71 reliable terms, changed significantly, $\rho(98) = .47$, $p < .001$. That result would indicate a moderate positive correlation between frequency and ordinal HHI score, or terminological consistency. The effect size still denotes medium effect.

6.3 Correlation between structure and terminological consistency

Spearman's rank correlation coefficient was employed to test the correlation between the number of words a term consists of and its terminological consistency. The results show that there is no correlation between the number of words and ordinal HHI score, ρ(98) = -0.002, p = .982 (see Figure 5.). As follows, both single-unit and multi-unit terms have the same probability to be (in)consistently translated as there is no correlation between terminological consistency and the number of words constituting a term.

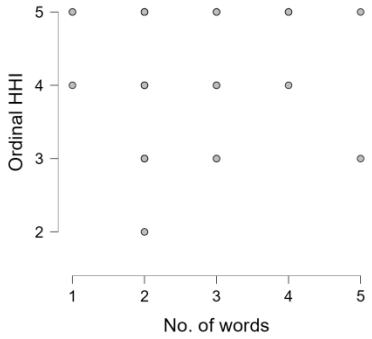

Figure 5. Scatterplot of correlation between no. of words and ordinal HHI score

7. Discussion

7.1 Terminological consistency in Croatian translations

The main aim of this study was to determine whether there are terminological inconsistencies in the Croatian translations of the EU legislation. Terminological consistency is considered one of the most essential features of specialized texts, and in turn translations. It enhances readability and information transfer, as well as reduces the possibility of misunderstanding and ambiguity (Gašpar et al., 2022, 2). In the context of the EU, consistency is even more insisted upon, which can be observed in its drafting rules. One of the general principles in the drafting of EU legislation is that "the terminology used in a given act shall be consistent both internally and with acts already in force, especially in the same field. Identical concepts shall be expressed in the same terms, as far as possible without departing from their meaning in ordinary, legal or technical language" (European Commission, 2015, 20). Since EU legislation is implemented into the law of every Member State, terminological consistency ensures there are no

ambiguities or difficulties in its interpretation, and consequently enhances the harmonisation of laws between Member States. Legal uncertainty can have serious consequences at both national and EU level, resulting in misinterpretation of rights and obligations, or even legal disputes (Stefaniak, 2017). Due to those circumstances, terminology work in drafting and translation processes has been brought to the forefront of the EU's language service tasks. Nonetheless, the results of the analysis confirm that there are terminological inconsistencies in Croatian translations, i.e. that translators rendered certain terms using more than one translation equivalent, or terminological variant. However, although inconsistency is present, it can be said that the overall consistency of the analysed terms in this corpus was relatively high, since 51 of the 100 examined terms had the optimal HHI score of 100.

7.2 Correlation between frequency and terminological consistency

Since inconsistencies were detected, the next aim was to examine the correlation between (in)consistency and a term's frequency, or more precisely, to determine whether higher frequency positively correlated with higher consistency. The sample consisted of many more infrequent terms than frequent ones. This, however, was expected due to the skewness of frequency being "a design feature of language" (Taylor, 2012, 180). As Taylor (2012) explains, this means that normally "a small number of very common words make up the bulk of a text, a fair number of moderately frequent words constitute somewhat smaller proportion, while a very large number of infrequent words account for only a tiny amount of a text" (156). In view of this, the skewed distribution was accounted for in the analysis by using Spearman's non-parametric correlation test. The results found a weak positive correlation between frequency and terminological consistency, confirming the second hypothesis. This means that the more frequent the term is, the more consistently it tends to be translated. Nevertheless, it is important to remember that consistency, as measured by HHI, does not account as much for the number of variants, as for their ratio across all occurrences. As a result, a term like *Union producer* with four terminological variants, still has a high HHI, *Ct* = 95.71, while a term with two variants like *special fiscal territory* has a low HHI, *Ct* = 50.78, because the variants are relatively equally distributed. These findings could be partially explained by the

use of TMs in EU translation. As explained, the EU has a central translation memory called *Euramis*, which is automatically integrated into the translation process. It offers the translator already translated segments stored in the memory that are similar to the ones being translated. The translator can choose to copy them, retain them with alterations or ignore them, depending on how similar the retrieved and new segments are. The TMs can also be used to look up terms or phrases in older documents to get an overview of the context they occur in. As follows, the highly frequent terms are more likely to occur in previous documents, i.e. be present and found by the translator in the TMs. On the other hand, less frequent terms might not be as present. In such cases the translator might have to use other resources, like IATE which can, as mentioned, be lacking, to find a term's right translation equivalent. If none can be found, they might have to create a new one. However, because of their dependence on older translations, TMs can both enhance and reduce terminological consistency. If they consist of translations with multiple terminological variants for one term, the TM system might suggest two segments containing different variants to two different translators, depending on the context the term occurs in. If the translator does not look up the established term translation in their language, but rather automatically copies the unsuitable term, its frequency in the TM might increase, perpetuating the process. It is possible that such line of events is the reason behind certain terms like *subtotal* having a rather equal distribution of variants. For this reason, all language departments in the DGT have so-called sentence managers, whose main task is to update translation memories, as well as update IATE in cooperation with the terminologists (European Commission 2012, 25).

Another aspect of this correlation is that translators are more likely to learn, and recognize more frequent terms, and consequently, know how they should be translated. Even if they do not know the preferred translation equivalent, they would be aware that it is a term they should look up to retain consistency. Conversely, less frequent terms might go unrecognized and therefore be translated differently depending on the context. For instance, *specific contract* had a low absolute frequency in this corpus, $f = 3$, and was translated differently each time. According to IATE, the term denotes a "contract specifying details of a particular task based on the previously signed framework contract or agreement,

dynamic purchasing system or qualification system", and its Croatian equivalent is *posebni ugovor*. However, two translations contained variants *pojedini* and *pojedinačni ugovor*, pointing to the fact that the translators probably failed to recognize the phrase as a term, thus did not look it up, and translated it using a phrase with a more general meaning which seemingly fit the context, but did not retain the term's true meaning.

As previously discussed, one of the study's limitations regarding term extraction was possible unreliability of IATE's entries. For this reason, the correlation test was also performed on data which excluded possible unreliable terms from the study's term list. The criteria of unreliability included the term not existing in IATE's online base, its reliability score on IATE being low, or the lack of a listed definition for the term on IATE. The correlation coefficient for data excluding the seven terms with a low reliability score didn't change significantly. However, when the test was performed on data excluding additional 22 terms with no listed definition, the correlation coefficient increased significantly, indicating a moderate correlation between frequency and terminological consistency. It is hard to say if the coefficient's value changed due to the smaller sample size, or if IATE's unreliability and its possible effect on the translation process presented itself as a confounding variable. Nonetheless, the results of the study indicate that a correlation between terminological consistency and frequency does exist, but they should be taken as preliminary due to certain methodological limitations and the lack of other analogous research in the area.

7.3 Correlation between structure and terminological consistency

Conversely, the same cannot be said for the correlation between terminological consistency and the number of words a term consists of. The results of the correlation test showed no correlation between those two variables, rejecting the third hypothesis. The basis of the hypothesis was the expectation that longer terms might be translated differently in contexts where their intended translation equivalent would decrease readability. It was also expected that, similarly to infrequent terms, translators might not recognize the whole phrase as a term in case of longer terms which would result in partially consistent translation. No such cases were recorded, which could indicate that consistency is indeed given precedence over decreased readability. Moreover, a term's length might not be a particularly relevant factor in the translation process, especially since the

distribution of terms indicates that most terms consist of two words (see Figure 2.).

7.4 Limitations and relevance

The main limitations of this study, as pointed out in Section 5, come from the use of IATE as a reference for the validity of the extracted terms. However, the applied methodological approach does give an insight into the current state and practical use of IATE, which presents itself as an additional contribution. Other limitations stem from the use of nonparametric correlation tests which are less precise than parametric tests and need a larger sample size to show sufficient results (Eddington 2015, 37). The results should consequently be taken as preliminary, as the scope of the study was rather small, focusing only on trade-related terminology. It is also hard to account for the confounding variable of the translator's lack of attention or skill which might have influenced the translation in some cases. Furthermore, despite the general discourse about terminological (in)consistency in the EU, there have not been many studies exploring the issue using a quantitative approach, so the findings cannot be compared or interpreted in a broader context of the research area.

The results, however, do provide an insight into the state of terminological consistency in recent Croatian translations of EU legislation. From the perspective of translation, the study illustrates the undesirable inconsistency, while also presenting potential causes and aspects that should be observed in practice in order to improve consistency, and consequently quality, of EU translations. From the perspective of terminology, it adds to the recently opened discussion about terminological variation by providing material and reference for further research, whether theoretical or empirical. The EU is today a cornerstone of multilingual, multidomain terminological work, so an empirical analysis of terminological variants in its translations can help provide a better understanding of the current use and dynamics of terminology. Finally, the study expands on the implementation of the Herfindahl Hirschman Index for measuring terminological consistency and supports its use on the English-Croatian language pair.

8. Conclusion

Due to its impact on clarity and interpretation, terminology is a crucial aspect of EU translation. Its harmonisation presents itself as a challenge both because of the interplay between the EU legal system and Member States' diverse legal systems, and because of the fast-paced environment in which terminological work and translation are done. Integration of IATE, the EU's multilingual terminology database, was supposed to facilitate this process and help ensure terminological consistency across all documents and languages. The aim of this study was, therefore, to check whether there are still inconsistencies in the use of terminology in Croatian translations of EU legislation, and to determine whether there is a correlation between the frequency and structure of a term and its terminological consistency where inconsistencies were found. The consistency of 100 trade-related terms taken from IATE was analysed in a corpus compiling English-Croatian EU translation memories from 2020. It was measured using the Herfindahl Hirschman Index, a method innovated by Itagaki et al. (2007). The results have found terminological inconsistency to be present, with the overall consistency still being relatively high. Statistical analysis reported a weak positive correlation between consistency and frequency, and no correlation between consistency and structure. However, due to the limitations, notably the small dataset and lack of analogous research, the results are to be taken as preliminary.

The study's main contribution is thus to provide an insight into the current state of terminological consistency in Croatian translations of EU legislation and add to further discussions and research on terminological consistency in EU translation.

Additionally, it provides an overview of several EU language resources, such as IATE and DGT's TMs, and illustrates how useful they can be for linguistic and terminological research. Future studies could focus on the domain-specificity of terms in the context of consistency. The correlation between frequency and terminological consistency within the context of TMs and relevance could also be further analysed, as that might have practical outcomes.

Lastly, this study continued the work of Itagaki et al. (2007), Gašpar (2013) and Gašpar et al. (2022) by implementing the Herfindahl Hirschman Index to measure terminological consistency. The method was proved to be a suitable tool

for quantitative terminological research, with the results supporting its use on the English-Croatian language pair and in EU legislation.

References

Bajčić, Martina and Stepanić, Martina. 2011. "Nedosljednost u prevođenju pojmova iz prava tržišnoga natjecanja Europske unije". In Hrvatski na putu u EU, edited by Maja Bratanić, 133-151. Zagreb: Institut za hrvatski jezik i jezikoslovlje, Hrvatska sveučilišna naklada.

Biel, Łucja. 2023. "Variation of legal terms in monolingual and multilingual contexts: Types, distribution, attitudes and causes". In Handbook of Terminology.: Volume 3, edited by Łucja Biel and Hendrik J. Kockaert. 90-124. Amsterdam, Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hot.3

Bratanić, Maja and Lončar Maja. 2016. "The Myth of EU Terminology Harmonization on National and EU Level". In Language and Culture in EU Law, edited by Susan Šarčević, 207-218. London, New York: Routledge.

Bowker, Lynne. 2015. "Terminology and translation". In Handbook of Terminology.: Volume 1, edited by Hendrik J. Kockaert and Frida Steurs. 304-323. Amsterdam, Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hot.1

Craig, Paul. 2012. *EU Administrative Law 2nd ed*. New York: Oxford Univ. Press. https://doi.org/10.1093/acprof:oso/9780199568628.003.0018

Eddington, David. 2015. *Statistics for Linguists: A Step-by-Step Guide for Novices.* Newcastle upon Tyne: Cambridge Scholars Publishing.

European Commission, Directorate-General for Translation. 2012. *Quantifying quality costs and the cost of poor quality in translation : quality efforts and the consequences of poor quality in the European Commission's Directorate-General for Translation.* Publications Office of the European Union. https://data.europa.eu/doi/10.2782/44381

European Commission: Legal service. 2015. *Joint practical guide of the European Parliament, the Council and the Commission for persons involved in the drafting of European Union legislation*. Publications Office of the European Union. https://data.europa.eu/doi/10.2880/5575

European Commission, Directorate-General for Translation. 2016. *Translation tools and workflow*. Publications Office of the European Union. https://data.europa.eu/doi/10.2782/703257

European Commission, Directorate-General for Translation. 2023. *Translation in figures 2023.* Publications Office of the European Union. https://data.europa.eu/doi/10.2782/438164

European Commission. n.d. "DGT-Translation Memory*".* Accessed January 26, 2024. https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en#dgt-memory

Fischer, Márta. 2010. "Language policy, translation and terminology in the European Union". In Terminology on everyday life: Terminology and Lexicography Research and Practice 13, edited by Marcel Thelen and Frieda Steurs. 21-34. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Fischer, Márta. 2022. "Horizontal and vertical terminology work in the context of EU translations". Porta Lingua (1): 7–15. https://doi.org/10.48040/pl.2022.1.1

Gašpar, Angelina. 2013. "Računalno potpomognuta provjera terminološke dosljednosti prijevoda hrvatskog zakonodavstva na engleski jezik". PhD diss. University of Zagreb.

Gašpar, Angelina, Seljan, Sanja and Kučiš Vlasta. 2022. "Measuring Terminology Consistency in Translated Corpora: Implementation of the Herfindahl Hirshman Index". Information 13 (2): 43. https://doi.org/10.3390/info13020043

Goss-Sampson, Mark A. 2022. *Statistical analysis in JASP: a guide for students*. https://jasp-stats.org/jasp-materials/statistical-analysis-in-jasp-a-students-guide-v16/

Itagaki, Masaki, Aikawa, Takako and He, Xiaodong. 2007. Automatic validation of terminology translation consistency with statistical method. Paper presented at *Machine Translation Summit XI, Copenhagen, September 10.* https://aclanthology.org/2007.mtsummit-papers.36.pdf

ISO 1087. 2019. Terminology work and terminology science – Vocabulary. Geneva: International Standardization Organization.

Kockaert, Hendrik J., and Steurs, Frida Eds. 2015. *Handbook of terminology.: Volume 1*. Amsterdam, Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hot.1

L'Homme, Marie-Claude. 2020. *Lexical Semantics for Terminology: An introduction*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.20

Pozzo, Barbara. 2020. "Looking for a consistent terminology in European contract law". *Lingue* Culture Mediazioni – Languages Cultures Mediation LCM Journal (71): 103–126. https://doi.org/10.7358/lcm-2020-001-pozz

Rogers, Margaret. 2008. "Consistency in Terminological Choice: Holy Grailor False Prophet?" In SYNAPS 21: Festskrift for Magnar Brekke, edited by Ingrid Simmons, 107-113. Bergen: NNH.

Sager, Juan C. 1990. *Practical Course in Terminology Processing*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Stefaniak, Karolina. 2017. "Terminology work in the European Commission: Ensuring high-quality translation in a multilingual environment". In Quality aspects in institutional translation, edited by Tomáš Svoboda, Łucja Biel and Krzysztof Łoboda, 109–121. Berlin: Language Science Press. DOI:10.5281/zenodo.1048192

Taylor, John R. 2012. *The Mental Corpus: How Language Is Represented in the Mind*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199290802.001.0001

Temmerman, Rita. 2000. *Towards New Ways of Terminology Description: The sociocognitive approach*. Amsterdam, Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/tlr3

TERMINOLOŠKA DOSLJEDNOST U HRVATSKIM PRIJEVODIMA ZAKONODAVSTVA EU-A: KORPUSNO ISTRAŽIVANJE

Sažetak

Svake godine u EU-u prevedu se tisuće dokumenata. Pritom se posebna važnost pridaje terminološkoj dosljednosti jer utječe na jasnoću i pravnu sigurnost. Međutim, uzimajući u obzir tijek rada prevoditelja, točnije kratke rokove i veliki obujam posla raspodijeljen na više prevoditelja, nameće se pitanje koliko se dosljedno terminologija koristi. Cilj je ovog istraživanja provjeriti postoje li nedosljednosti u uporabi trgovinske terminologije u prijevodima zakonodavstva EU-a na hrvatski jezik te ispitati povezanost frekvencije i strukture termina te njihove dosljednosti. Terminološka dosljednost mjeri se pomoću HHI-ja, slijedeći primjer Itagakija i sur. (2007). Termini su preuzeti iz baze IATE i analizirani u korpusu sastavljenom od englesko-hrvatskih prijevodnih memorija Glavne uprave za prevođenje Europske komisije iz 2020. Rezultati su potvrdili prisutnost terminološke nedosljednosti, kao i slabu pozitivnu korelaciju između dosljednosti i frekvencije te izostanak korelacije između dosljednosti i strukture. Istraživanjem se također namjerava istaknuti korisnost jezičnih materijala EU-a, i

HHI-ja za lingvistička i terminološka istraživanja.

Ključne riječi: terminološka dosljednost, Europska unija, prevođenje, HHI