Exploring Word Formation Trends in Written, Spoken, Translated and Interpreted European Parliament Data – A Case Study on Initialisms in English and German

Katrin Menzel
Saarland University
Campus A2.2, 66123 Saarbrücken, Germany
k.menzel@mx.uni-saarland.de

Abstract

This paper demonstrates the research potential of a unique European Parliament dataset for register studies, contrastive linguistics, translation and interpreting studies. The dataset consists of parallel data for several European languages, including written source texts and their translations as well as spoken source texts and the transcripts of their simultaneously interpreted versions. The paper presents a cross-linguistic, corpus-based case study on a word formation phenomenon in these European Parliament data that are enriched with various linguistic annotations and metadata as well as with information-theoretic surprisal scores. The paper specifically addresses the questions of how initialisms are used across languages and production modes in the English and German corpus sections of these European Parliament data and whether there is a correlation between the use of initialisms and the use of their corresponding multiword full forms in the analysed corpus sections. The correlation analysis particularly addresses the question of whether initialisms in the analysed discourse types function as synonymous alternatives used in alternation with their full forms or primarily as replacements increasing compactness and lexical economy, but not necessarily transparency. Additionally, the paper explores what insights might be gained from an analysis of information-theoretic surprisal values with regard to the informativity and possible processing difficulties of initialisms. The results show that English written originals and German translations are the corpus sections with the highest frequencies of initialisms. The majority of cross-language transfer situations lead to fewer initialisms in the target texts than in the source texts, which means that they are either entirely omitted or other means are used to replace them in mediated discourse, e.g. hypernyms as less specific terms or multiword terms as semantically more explicit variants. In the English data, there is a positive correlation between the frequency of initialisms and the frequency of the respective full forms. There is a similar correlation in the German data, apart from the interpreted data. Additionally, the results show that initialisms represent peaks of information with regard to their surprisal values within their segments. Particularly the German data show higher surprisal values of initialisms in mediated language than in non-mediated discourse types, which indicates that in German mediated discourse, initialisms tend to be used in less conventionalised textual contexts than in English.

Keywords: European Parliament data, translation and interpreting data, corpus analysis

1. Introduction

1.1 Background and Motivation

This paper presents an example of the research potential of a unique European Parliament dataset consisting of parallel data for several European languages, including written source texts and their translations as well as spoken source texts and the transcripts of their simultaneously interpreted versions. The paper presents a cross-linguistic, corpus-based case study on a word formation phenomenon in these data that are enriched with various linguistic annotations and metadata as well as with information-theoretic surprisal scores calculated from the probabilities output of a 4-gram model trained for each language on an external domain-comparable resource. It thus applies language modelling to the recently published multilingual resource for the cross-lingual retrieval and analysis of selected word formation types within their contexts in the respective discourse types.

The case study presented in this paper compares English and German in EU parliamentary debate speeches. Furthermore, written and spoken mode are compared as the dataset includes edited, published records from the parliamentary debates and verbatim transcripts of the debates reflecting spoken language features such as repetitions, unfinished sentences, reformulations etc. Additionally, the non-mediated language of source texts is compared to mediated, i.e., translated and interpreted language. Here the focus is on initialisms as a particular type of word formation choices in European Parliament texts that are characterised by informative and persuasive messages and that need to be transferred to other languages.

Only a few case studies have discussed morphology and word formation in the context of contrastive research and translation studies (e.g. Cartoni and Lefer, 2011; Lefer, 2012; Defrancq and Rawoens, 2016; Berg, 2017). Ström Herold et al. (2021) specifically looked at initialisms in parallel data. Nevertheless, word formation remains an understudied area in corpus work on specialised registers and on translated and interpreted discourse. Moreover, there is still a research gap on initialisms in corpus linguistics, contrastive linguistics and translation and interpreting studies that this paper aims to address. The theoretical morphological literature has often treated initialisms as peripheral, marginal or extra-grammatical word formation patterns (cf. Menzel, forthcoming, for a literature summary). However, initialisms are a very interesting and unique strategy for shortening multiword terms to word-like units in a one-token

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 57–65

format which gives them higher syntactic flexibility than their underlying full expressions. Initialisms have more complex functions and features than mere abbreviations, and they therefore deserve a much more prominent role in theoretical morphology and in corpus-based work.

The purpose of the analysis is to show how initialisms as a specific type of word formation and shortening strategy for multiword expressions (MWE) are used across languages and production modes in the English and German corpus sections of the selected datasets and whether there is a correlation between the use of initialisms and the use of their corresponding full forms in the analysed corpus sections. The correlation analysis particularly addresses the question of whether initialisms in the analysed discourse types function as synonymous alternatives used in alternation with their full forms or primarily as replacements increasing compactness and lexical economy, but not necessarily transparency. Additionally, the analysis presents insights on the informativity and on possible processing difficulties of initialisms gained from information-theoretic surprisal values. The data used for the surprisal calculations and for the corpus-linguistic analysis are the EuroParl_UdS[1] (Karakanta et al., 2018) and the EPIC-UdS corpora (Przybyl et al., 2022a/b, Menzel et al., forthcoming).

1.2 Initialisms

Initialisms can be defined as combinations of initial letters of multiword sequences of words functioning as shortened, more word-like forms of their spelt-out forms. Examples are the letter-by-letter initialism *EPA* in which each letter corresponds to a part of the full multiword expression ***E**conomic **P**artnership **A**greement* or the acronymic initialism *CITES* with a word-like pronunciation as a short form of ***C**onvention on **I**nternational **T**rade in **E**ndangered **S**pecies*. In a broader sense, initialisms also include shortenings of multimorphemic individual orthographic words that contain more than one meaningful part. By using this broader definition, we may include initialisms of multiword expressions that contain closed compounds (e.g. *EFSF* for ***E**uropäische **F**inanz**s**tabilisierungs**f**azilität*) that are often found in German where English typically prefers open compounds although shortening processes may lead to similar reduced forms in both languages (e.g. *EFSF* for ***E**uropean **F**inancial **S**tability **F**acility*). On the basis of this broader definition, we also include shortenings of expressions that contain individual words with combining forms whose initial letters are used in abbreviated forms as is often the case in technical and scientific concepts (e.g. *PCB* for ***p**oly**c**hlorinated **b**iphenyl* or *AIDS* for ***a**cquired **i**mmuno**d**eficiency **s**yndrome*).

Initialisms are productive in specialised registers such as political, administrative, military and business language. They function as insiders' code words giving shorter labels and an intended flavour of familiarity to concepts that already have multiword designations (Mattiello, 2013: 66). The shortened form is the result of the compression of a semantically equivalent multi-word denomination that refers to the same referent. Both the full and the short form continue to coexist as absolute synonyms, but their formal and stylistic features may make them suitable for different contexts.

Many initialisms in the register of EU parliamentary debates replace multiword proper nouns referring to institutions, groups, projects and policies that are important for the internal structure and the networks of the EU as the organisation in which the discourse takes place. The texts also contain initialisms for geographical entities and for technical and scientific concepts that play a role in the parliamentary debates.

2. Data

The written dataset EuroParl_UdS consists of parallel, sentence-aligned corpora for English, German and Spanish, and the source side contains texts only by native speakers of the respective languages. The corpus has been enriched with various metadata that were not available in previous European Parliament corpora. The EuroParl_UdS data are based on speeches adapted to the requirements of written language. They contain edited and published records of debates that took place in the European Parliament and they also contain their officially published translations. Like data from other parliamentary records such as the British Hansard (SAMUELS Consortium, 2015), they also include some written statements to the Parliament from parliamentary sessions.

The spoken dataset EPIC-UdS is also a multilingual parallel corpus of political debates from the European parliament for English, German and Spanish. Here, the release version V3 (Przybyl et al., 2022b) is used. Like in EuroParl_UdS, various metadata have been added to the EPIC-UdS texts (for instance, the speed of the speeches in words per minute and the topics of the texts). The EPIC-UdS data are unedited verbatim transcripts of what was said in parliamentary debates, and they also include simultaneous interpreting transcripts. For various written corpus texts, there are also the corresponding spoken ones in EPIC-UdS, but of course not for all of them as the spoken sections are smaller than the written ones. This paper focusses on the data from the German-English language pair and the respective corpus sections in the analysis (cf. Table 1).

[1] UdS stands for 'Universität des Saarlandes' (Saarland University)

	Corpus section	Tokens
English	EPIC-UdS EN orig. (spoken)	68.548
	EPIC-UdS EN interpr.	59.100
	EuroParl_UdS EN orig. (written)	8.693.135
	EuroParl_UdS EN transl.	6.260.869
German	EPIC-UdS DE orig. (spoken)	57.049
	EPIC-UdS DE interpr.	58.218
	EuroParl_UdS DE orig. (written)	7.869.289
	EuroParl_UdS DE transl.	3.100.647

Table 1: Corpus size of EuroParl_UdS and EPIC-UdS V3[2]

EuroParl_UdS and EPIC-UdS complement each other. Additionally, they complement other European Parliament datasets that contain translated or interpreted texts such as the EuroParl Simultaneous Interpreting Corpus (ESIC, Macháček et al., 2021), the Hungarian European Parliamentary Intermodal Corpus (HEPIC, Götz, 2020), the Polish Interpreting Corpus (PINC, Chmiel et al., 2022) and the EP-Poland Interpreting Corpus (Bartłomiejczyk et al., 2022). EPIC-UdS in particular builds on the experience of existing EPIC[3] parallel corpora developed at the University of Bologna (cf. Bendazzoli and Sandrelli, 2005; Russo et al., 2012, Bernardini et al., 2018) and EPICG at Ghent University (Defrancq et al., 2015) by using similar standards and transcription guidelines, and it extends them with the German-English language pair. There is also EPTIC (the European Parliament Translation and Interpreting Corpus), a bidirectional English-Italian corpus of interpreted and translated EU Parliament proceedings aligned to each other and to their corresponding source texts, i.e. the transcripts of the speeches and their edited and published written versions (Bernardini et al., 2016). The range of these corpora can be used to test hypotheses from translation studies in translated and / or simultaneous interpreted language. Some of these datasets have been used, for instance, to look at lexical and syntactic simplification processes, but the role of word formation patterns in parliamentary discourse and in translated or interpreted speech has not yet been a major research focus despite its potential significance in this context.

Table 2 contains example extracts from the spoken and written versions of a speech that illustrate the use of initialisms in the different corpus section types used for the analysis in this paper. In this table, we see that there are not many differences from the transcript of the live speech to the written and published version in the German example, only a grammatically correct form of the definitive article "der" replaces "des" before *EFSF* and "Einsatz" is used instead of "Nutzen" in this nominal group to use a more conventionalised context in front of the initialism. The examples in Table 2 illustrate what we might generally expect: nominal groups with initialisms sometimes become longer in translations via explicitation. Here, the full term for *EFSF* is added in the English translation before the initialism is introduced. In interpreted texts, nominal groups with initialisms remain short. Explicitation of initialisms is rare in interpreting, and these forms are used in contexts of more general vocabulary than in the other corpus sections (e.g. *"help from the EU"* in the English interpreted version vs. *"remedial measures from the EU"* in the translated version).

EPIC-UdS DE orig. (spoken)	**EPIC-UdS EN interpr.**
[…] erscheinen konzertierte Hilfsmaß-nahmen von EU und IWF das Nutzen des EFSF unausweichlich zu werden	*[…] agreed help from the EU and the IMF the use of the EFSF seem to be unavoidable*
EuroParl_UdS DE orig. (written)	**EuroParl_UdS EN transl.**
[…] erscheinen konzertierte Hilfsmaß-nahmen von EU und IWF und der Einsatz der EFSF unausweichlich zu werden.	*[…] concentrated remedial measures from the EU and the IMF and the use of the European Financial Stability Facility (EFSF) appear inescapable.*

Table 2: Example extracts from EPIC-UdS and EuroParl_UdS with initialisms

Tables 3 and 4 with longer extracts from the different versions of a parliamentary speech illustrate other examples of initialisms in the dataset that show that these forms are part of lexical chains and contribute to the network of cohesive ties in the texts.

EPIC-UdS DE orig. (spoken)	**EPIC-UdS EN interpr.**
*[…] und uns Gedanken machen wie dieses in dem Zusammenspiel mit dem **Europäischen Sozial-fonds** möglicherweise noch effizienter gestaltet werden kann*	*we want to see how we can make this even more efficient together with the **ESF** as well*
*was die Finanzierungs-quellen angeht haben Sie natürlich Recht was die Zahlungsermächtigung aus dem **ESF** angeht*	*you're quite right when it comes to payment appro-priations from the **ESF***
*aber am Ende möchte ich schon dass das Gesamt-spiel der Verpflichtung und der Zahlung sowohl für die **Strukturfonds** als auch für den **ESF** dann so aus-geht wie wir es in den Gesamtzahlen vereinbart haben*	*however what I would like to see is that the commit-ment appropriations and the payment appropriations should actually happen with the **European Struc-tural Funds** as we've set out in the interinstitutional agreement*

Table 3: Example extracts from EPIC_UdS with an initialism (*ESF*) in lexical chains establishing cohesive links between textual elements

[2] EN = English, DE = German, orig. = original (source) texts, transl. = translations, interpr. = interpreted texts

[3] European Parliament Interpreting Corpus

EuroParl_UdS DE orig. (written)	EuroParl_UdS EN transl.
*Wir müssen uns Gedanken machen, wie dies im Zusammenspiel mit dem **Europäischen Sozialfonds (ESF)** möglicherweise noch effizienter gestaltet werden kann.*	*We need to contemplate how this interaction with the **European Social Fund (ESF)** could possibly be better shaped.*
*Was die Finanzierungsquellen angeht, haben Sie, was die Zahlungsermächtigungen aus dem **ESF** angeht, natürlich Recht.*	*As far as the sources of funding are concerned you were, of course, absolutely correct in what you said about the payment appropriations from the **ESF**.*
*Aber am Ende möchte ich schon, dass das Gesamtspiel der Verpflichtungen und der Zahlungen sowohl für die **Strukturfonds** als auch für den **ESF** dann so ausgeht, wie wir es in den Gesamtzahlen vereinbart haben.*	*Ultimately, however, I would like the overall picture for the obligations and the payments, both for the **structural funds** and for the **ESF**, to be as we agreed in the overall figures.*

Table 4: Example extracts from EuroParl_UdS with an initialism (*ESF*) in lexical chains establishing cohesive links between textual elements

The extracts in Tables 3 and 4 illustrate general differences between the written and spoken versions of the parliamentary debate speeches. In these extracts, the initialism *ESF* is used several times in lexical chains to create lexical cohesion between different segments via repetition, the use of the synonymous full form and other semantic relations such as hyponym-hypernym relations. At the beginning of the written and the translated versions of the speech in both languages in the EuroParl_UdS data, we have the first use of the initialism after the use of its full MWE, similarly to what we would find in many other formal written registers in both languages. In the transcribed spoken text in EPIC-UdS, only the full form is used by the speaker at the beginning as it is less common to give a pair of a short and long form of the same concept in spoken language. The listeners have to make the implicit connection between the full and the short form in this spoken text on their own. In the interpreted version in EPIC-UdS, only the initialism is used in the first segment as it is faster to pronounce than *European Social Funds,* and we may assume that the interpreter is familiar with the term to make the connection between the full term and its short form during the interpreting process. The interpreter also seems to expect the audience to understand what *ESF* stands for. However, the word formation choices of the interpreter lead to a different cohesive structure of the target text – the full term is not mentioned before the interpreter starts using the short form. Later, the hyponym *European Structural Funds* is used in the interpreted version (in fact, the *ESF* is one of the *European Structural and Investment Funds*). Understanding the network and chains of lexical relations in the interpreted version is more demanding for the audience than in the other text versions.

3. Analysis and Results

3.1 Query Design

The retrieval process for initialisms can be compared to developing annotation guidelines for a pattern that might seem rather fuzzy in the existing literature. It involved linguistic work of decisions with regard to relevant categories and subcategories. There are various phenomena that look similar on the surface, but decisions need to be taken to determine which ones have to be excluded due to their irrelevance. Some decisions need to be made in order to optimise precision and recall (e.g. to exclude forms which are theoretically possible and exist in various text types but are marginal for our dataset). As a first step, rather broad CQP queries (Corpus Query Processor, cf. Evert 2005) were used for words containing capital letters. Irrelevant forms were excluded via refined queries, e.g. abbreviations such as *EUR*, actual words spelt with capital letters such as *CARS* (an EU Action Plan on the car industry), forms that contain splinters from source words that fall under blends (e.g. *ALTENER* for *Alternative Energy Programme*) and mixed forms with only some letters as in initialisms (e.g. *REACH* that contains more than one letter from a source word [CH = Chemicals]). As they are marginal for this dataset due to spelling rules in EU style guides,[4] initialisms with small letters or periods (e.g. *aids* or *G.M.T.*) were excluded from the queries. They would occur more frequently in other text types or in older data. Hyphenated and open compounds that start with an initialism (e.g. *HACCP-based, EIB operations*) represent an interesting case from a cross-linguistic perspective due to different compounding strategies in English and German, but they will not be a particular focus of the analysis in this paper.

3.2 The Usage of Initialisms across Languages and Production Modes

One expectation for the analysis of initialisms across languages and production modes is to find that interpreters use many initialisms instead of multiword terms to save time. However, the spoken original data usually have more initialisms than the interpreted speeches. The German interpreted data have the lowest number of initialisms of all corpus sections (cf. Fig. 1). If we look at translated and interpreted data, we also have to take the influence of the respective source texts and the frequencies of initialisms in them into account. For instance, the differences between the English original spoken data and interpreted German are more pronounced than between the German original spoken data and the

[4] cf. for instance, *English style guide – A handbook for authors and translators in the European Commission*, [Latest PDF version: https://commission.europa.eu/document/download/c45f5b70-2d0e-4da7-b181-b5fe3a16c4bb_en]

interpreted English data. Thus, initialisms as word formation patterns are not just copied one to one as anglicisms in spoken language transfer from English to German.

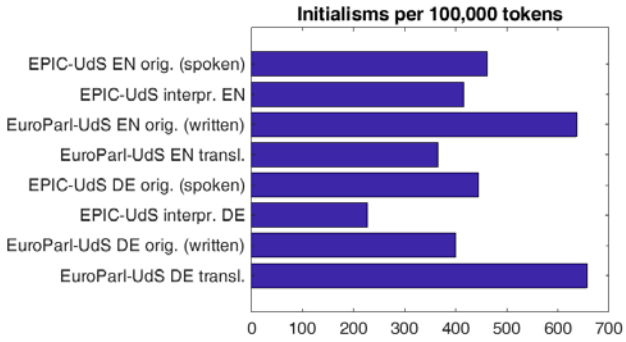

Figure 1: Usage of initialisms across languages and production modes

English written originals and German translations are the corpus sections with the highest frequencies among all. The German translations even have slightly more initialisms than the original English written texts, and German originals have a considerably lower frequency. In summary, we see the following trends in language transfer: English to German leads to fewer initialisms in spoken language transfer, but to more initialisms in written language transfer. In both spoken and written language transfer from German to English, fewer initialisms are found in the target texts than in their originals. Generally, that means that in most language transfer situations from the English-German language pair, and particularly in spoken language transfer, some initialisms from the source texts are either entirely omitted or other means are used to replace them in the target texts.

3.3 Frequency of Usage of Initialisms and their Corresponding Full Forms

This section addresses the question of whether there is a correlation between the use of initialisms and the use of their corresponding full forms in the analysed corpus sections. Measuring this correlation will reveal whether initialisms mainly function as synonyms to their MWE in this register, namely if the MWE have similar or higher frequencies than the initialisms themselves. From time to time, especially in debates on a variety of specialised topics, speakers may want to remind the audience of what an initialism as a potentially ambiguous form consisting of letters as submorphemic elements stands for. Additionally, short and long forms referring to the same concepts may be used in alternation in the texts to function like synonyms, as other types of synonyms for specialised multi-word terms or multi-word named entities are not necessarily available. However, if many initialisms are very conventionalised in this register, their full expressions may occur rarely or not at all in the texts. If this is mainly the case in the data, the most important function of the initialisms would be to give more efficient labels and an intended flavour of familiarity to specialised concepts that have lengthy multiword designations. Therefore, the more often conventionalised initialisms are used in the EU parliament discourse community, the less often the community might need to express their underlying full forms. A slightly positive correlation between frequently used initialisms and their MWE in both languages and all production modes can be expected.

Table 5 shows the correlation coefficients and the p-values in order to investigate the relationship between the normalised frequencies of initialisms and their corresponding MWE in the data. The 30 most frequent types of initialisms and their corresponding full forms in each corpus section were taken into account for this analysis. In contrast to all other forms found in the data, the initialism *"EU"* represents an extreme outlier. It is always used much more frequently than the second most frequent initialism in the respective datasets. It would have such a strong influence on the calculations and subsequent interpretation that it is excluded here in order to obtain more fine-grained insights on the other initialisms that are not characterised by such extreme values.[5]

	Corpus	**Correlation coefficient *r***	***p*-value**
English	EPIC-UdS EN orig. (spoken)	0.40	0.03
	EPIC-UdS EN interpr.	0.17	0.39
	EuroParl_UdS EN orig. (written)	0.88	4.83e-10
	EuroParl_UdS EN transl.	0.51	0.005
German	EPIC-UdS DE orig. (spoken)	0.05	0.76
	EPIC-UdS DE interpr.	0.40	0.03
	EuroParl_UdS DE orig. (written)	-0.0007	0.99
	EuroParl_UdS DE transl.	0.09	0.063

Table 5: Pearson correlation coefficients between normalised frequencies of most frequent initialisms and their corresponding MWE and significance level

[5] In all corpus sections, both the form "*EU*" and "*European Union*" were used with similar frequencies like synonyms (between 150 and 200 times per 100.000 tokens). Including these exceptions here would give us a correlation coefficient of almost 1 in all sections due to their high frequencies.

Table 5 shows that the English spoken and interpreted data have a slightly positive correlation for the frequency of initialisms and the frequency of the respective MWE, and the English written and translated data have a stronger positive correlation. There is not really any correlation to see in the German data, apart from a slightly positive one in the interpreted data. German uses some frequent multiword expressions in the original written and spoken data whose shortened forms are also among the top abbreviated forms in these data, but the usage of the full form is considerably more important than the usage of the initialism in some cases in German compared to English, while in other cases, the full form of a frequent initialism is not used at all or rarely in the German data. An obvious difference to English is that more initialisms in the German data originally represent foreign multiword expressions, but native equivalents for the full forms may exist as well. For instance, *"UN"* is used in German, but it is unusual to use the full English term in the German text. Additionally, the initialism has no visible link to the semantically equivalent German multiword term *"Vereinte Nationen"*. This may explain why in some cases neither the original full form nor an equivalent MWE is used frequently when a borrowed initialism has become conventionalised in the target language. Overall, the full expressions for frequently used initialisms seem to have become more unusual alternatives in German than in English.

3.4 Analysis of Surprisal Values

The data have been annotated with surprisal scores. Surprisal (S) has been calculated as the negative log (base 2) probability of each token (t) given its preceding context of three tokens measured in bits of information as in the following equation: $S(t_i) = -\log_2 p(t_i|(t_{i-1}\ t_{i-2}\ t_{i-3}))$. The values were calculated from the probabilities output by a KenLM 4-gram model, i.e. the model considers the three preceding words of each word to predict its surprisal. It was trained for each language on a domain-comparable resource. The data was balanced with regard to the size of the different corpus sections by discarding a number of random document pairs from the larger, written ones.

From an information-theoretic perspective, processing effort is related to surprisal that can be measured in bits (Hale, 2001; Degaetano-Ortlieb and Teich, 2022). For instance, the initialism *"CAP"* (Common Agricultural Policy) after the 3-token-sequence *"context of the"* is rather predicable with lower surprisal values in our data than *"CAP"* after a sequence such as "*be driven by"*. The assumption here was that initialisms, apart from the extremely frequent example *"EU",* would represent peaks of information with regard to their surprisal values within their segments.

Surprisal scores for all initialisms regardless of their frequencies were identified in the data and the average surprisal scores of the respective text segments were extracted together with the text of the segments (Fig. 2).

item	item_srp	raw	AvS	seg_id	doc_id	seg_num
ACTA	16.512183	Und wenn das Thema nicht so ernst w	8.4971327	ORG_SP_D	ORG_SP_DE_1	3
ACTA	16.747736	Es liegt also auf der Hand dass es in un	7.6255080	ORG_SP_D	ORG_SP_DE_1	12
ACTA	13.235625	Wenn die Kommission vermeiden möc	8.8465601	ORG_SP_D	ORG_SP_DE_1	17
ACTA	15.523390	Und drittens kann die Kommission ver	6.9500245	ORG_SP_D	ORG_SP_DE_1	22
AKP	16.478000	Dennoch im in dem Mittelpunkt steht	9.4555870	ORG_SP_D	ORG_SP_DE_0	18
AKP	17.936628	Während die regierende AKP dagegen	11.351781	ORG_SP_D	ORG_SP_DE_0	10
AKP	17.309927	Leider sehe ich auch bei der hochgelob	7.2819816	ORG_SP_D	ORG_SP_DE_0	14
AKP	17.936628	Die regierende AKP muss endlich eine	9.3342966	ORG_SP_D	ORG_SP_DE_0	16
ALDE	15.559537	Dies ist auch der Grund warum wir die	6.4293142	ORG_SP_D	ORG_SP_DE_1	12

Figure 2: Extract from table with extracted surprisal scores for initialisms (item_srp), the text segments (raw) and their average surprisal (AvS)

Figure 3 shows the range of the surprisal values of initialisms in the English data.[6] In the English translated data, surprisal is significantly higher than in the English written originals. We do not see the same difference between original spoken and interpreted discourse. Surprisal here is also generally higher in the spoken than in the written data.

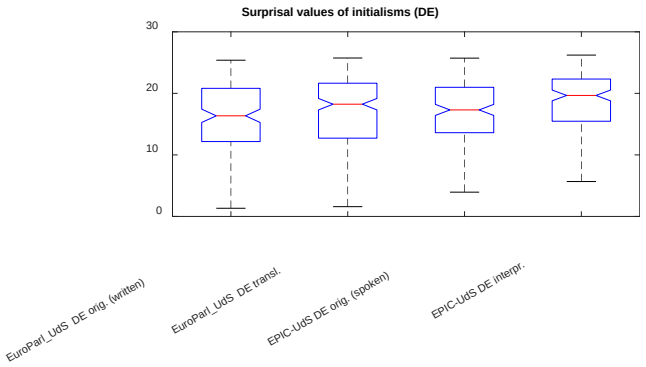

Figure 3: Surprisal values of initialisms in English

The German data in Figure 4 look slightly similar, but both types of mediated language production have significantly higher mean values than the respective non-mediated forms, which we can conclude from the plotted notches that represent the confidence interval around the median. This indicates that in German mediated discourse, be it written or spoken, initialisms are generally used in less conventionalised contexts than in original texts. Overall, the written and translated sections here in the German data turn out to be closer to the spoken and interpreted ones from the same language than in English.

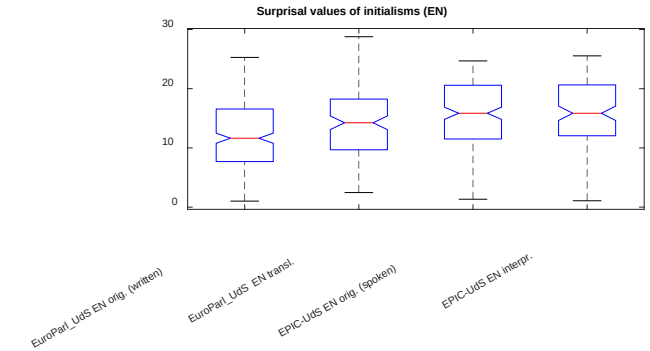

Figure 4: Surprisal values of initialisms in German

[6] Due to its exceptional frequency in all corpus sections, *"EU"* has again been excluded in this step.

The average surprisal of the entire segments in which initialisms occur is typically between 6 and 9 in all corpus sections (not plotted here). In most cases in both languages and all production modes, initialisms as condensed word-like forms of multiword terms indeed represent elements with high or very high surprisal compared to the average of their segments.[7] Interestingly, many examples do not have fixed sequences of part-of-speech patterns such as preposition + determiner before the initialism. There is generally a great variety of part-of-speech patterns in the preceding contexts, including content words such as verbs, nouns and adjectives. Generally, surprisal values for fixed elements in the full forms of the corresponding multiword expressions tend to be lower.

Initialisms achieve higher syntactic flexibility than MWE due to their one-word format. A qualitative analysis of initialisms in their contexts shows that untypical local context occur, for instance, if an initialism represents a MWE from a different language. Table 6 shows two examples of initialisms with very high surprisal values.

EPIC-UdS EN orig. (spoken):
And are we really happy that somebody who will be in charge of our overseas security policy was an activist a few years ago in an outfit like ***CND***
EPIC-UdS DE interpr.:
Und sind wir wirklich glücklich darüber, dass jemand, der für unsere außenpolitische Sicherheit zuständig ist, vor ein paar Jahren aktiv war in ***CND****.*

Table 6: Examples of initialisms with very high surprisal values (>20).

In the examples in Table 6, "*Campaign for Nuclear Disarmament"* is shortened, and already in the English original text, the value for "*CND"* was very high (20.45) due to an untypical context of the three preceding words, but in the German interpreted data, its value was one of the highest (25.64) as the form is not used in a great variety of contexts. From a cognitive perspective, reproducing a similar sequence of letters to produce a fluent target text might be less capacity-demanding in mediated discourse for the interpreters than replacing it with another structure. Nevertheless, an initialism like this might not be so common in the target language, and a different expression might normally be preferred by the target audience.

4. Conclusion and Outlook

To sum up, the case study presented in this paper has demonstrated the utility and a research context of the EuroParl_UdS and EPIC-UdS data that consist of written, spoken, translated and interpreted European Parliament texts for different languages. The case study on initialisms in English and German as a particular type of word formation and shortening strategy for MWE has shown differences and similarities between the languages and production modes in the data and provides valuable insights for the fields of register studies, contrastive linguistics, translation and interpreting studies. Some differences between the spoken and interpreted versions and the written and translated versions of parliamentary debate speeches may be due to the fact that the two former production modes directly address experts taking part in debates on specialised topics, while the two latter ones function as written documentation like reports. They address a larger, more heterogeneous audience of people including all those who did not take place in the actual debate. This explains some of the choices in the written texts, e.g. to restructure the elements and types of semantic relations in lexical chains in a different way than in the spoken texts or not to start right away with an initialism without mentioning the full form. Other strategies with regard to fixed multiword expressions and less explicit initialisms consisting of submorphemic elements reflect general mediated language effects and some are specific to interpreting due to high time pressure and cognitive effort in this language transfer task. In the annotated data, all segments have been extracted that contain no initialism, but the aligned source or target segment does contain one. Therefore, in a future analysis, it would also be useful to focus on specific contexts where initialisms were omitted or added in the translated or interpreted speeches and to analyse the types of translation/interpreting procedures in more detail. Generally, we can expect to see an overall trend towards explicitation in written translations (e.g. EuroParl_UdS DE orig. [written]: *das SIS* -> EuroParl_UdS EN transl.: *the Schengen Information System*) and the usage of less specific vocabulary, i.e. fewer initialisms and fewer multiword terms, in interpreting (e.g. EPIC-UdS DE orig. [spoken]: *das SWIFT-Abkommen* -> EPIC-UdS EN interpr.: *the agreement*).

One could further look into the subtypes of the initialisms, considering, for instance, their length, whether they have to be pronounced as one-word acronyms or letter by letter, what type of MWE they stand for (e.g. technical term or named-entity, foreign or native origin) and whether they are used as the head of a nominal group or as a premodifier of another noun as in that case they often cannot easily be replaced by the full form. A larger size of the spoken original and interpreted data would be useful for this type of analysis. Additionally, one might control for specific metadata when comparing word formation choices in the different production modes. What makes this challenging is that some types of metadata are not available although they would be relevant for particular questions (e.g. specific background information on the translators and interpreters). Other metadata types are not available for all types of production modes or difficult to use for specific studies in their current form. EPIC-UdS, for instance, contains information on the general topics and the titles of the debate as indicated by the European Parliament. However, the

[7] The outlier *"EU"* has low to medium surprisal values.

debates represent a huge variety of topics that are rather difficult to assign to overarching clear-cut categories (e.g. a debate on the beekeeping sector has been assigned the topic of “Economy”, the situation in the Middle East/Gaza Strip falls under “International affairs”, the democratic process in Turkey under “Politics” and “Food distribution to the most deprived persons in the Community (amendment of the Single CMO Regulation)” has been labelled with “Health”. Enlarging the spoken part and further enriching and enhancing the metadata would therefore be an opportunity to facilitate follow-up studies.

5. Acknowledgments

This work is based on research funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) – SFB 1102 / Project-ID 232722074, project B7 (Translation as Rational Communication). Additionally, the author is greatly indebted to Maria Kunilovskaya who calculated the surprisal scores and extracted them with the respective corpus segments.

6. Bibliographical References

Berg, T. (2017). Compounding in German and English – A quantitative translation study. In *Languages in contrast*, 17(1): 43–68.

Cartoni, B. and M. Lefer (2011). Negation and lexical morphology across languages: Insights from a trilingual translation corpus. In *Poznan studies in contemporary linguistics*, 47(4): 795–843.

Defrancq, B., K. Plevoets and C. Magnifico (2015). Connective items in interpreting and translation: Where do they come from? In J. Romero-Trillo (ed.) : *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to discourse and translation studies.* Cham: Springer International Publishing, 195–222.

Defrancq, B. and G. Rawoens (2016). Assessing morphologically motivated transfer in parallel corpora. In *Target*, 28(3): 372–398.

Degaetano-Ortlieb, S. and Teich, E. (2022). Toward an optimal code for communication: The case of scientific English. In *Corpus Linguistics and Linguistic Theory*, 18(1): 175–207.

Evert, S. (2005). *The CQP query language tutorial.* IMS: Stuttgart University.

Götz, A. (2020). Discourse markers and connectives in interpreted Hungarian discourse: A corpus-based investigation of discourse properties and their interdependence. In *Speech Science* 2020(1): 259–284.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8, Pittsburgh, Pennsylvania, June 2001, ACL.

Lefer, M. (2012). Word formation in translated language – The impact of language-pair specific features and genre variation. In *Across languages and cultures,* 13(2): 145–172.

Mattiello, E. (2013). *Extra-grammatical morphology in English. Abbreviations, blends, reduplicatives and related phenomena.* Berlin: Mouton de Gruyter.

Menzel, K. (forthcoming). Initialisms in scientific writing in the 19th and early 20th centuries. *Zeitschrift für Wortbildung,* 2/24.

Przybyl, H., A. Karakanta, K. Menzel and E. Teich, (2022a). Exploring linguistic variation in mediated discourse: Translation vs. interpreting. In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska and S. Bernardini (eds.): *Mediated discourse at the European Parliament: Empirical investigations.* Berlin: Language Science Press, 191–218.

Shannon, C. E. (1948): A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379–423.

Ström Herold, J., M. Levin and J. Tyrkkö (2021). RAF, DNA and CAPTCHA: English acronyms in German and Swedish translation. *Bergen Language and Linguistics Studies,* 11(1): 163–184.

7. Language Resource References

Bartłomiejczyk, M., E. Gumul and D. Koržinek (2022). EP-Poland: Building a bilingual parallel corpus for interpreting research. In *GEMA, Online Journal of Language Studies,* 22(1): 110–126.

Bendazzoli, C. and A. Sandrelli (2005). An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In H. Gerzymisch-Arbogast and S. Nauert (eds.): *Mutra2005 – Challenges of Multidimensional Translation. Proceedings of the Marie Curie Euroconferences.* Saarbrücken. May 2005. https://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html

Bernardini, S., A. Ferraresi and M. Miličević (2016). From EPIC to EPTIC – Exploring simplification in interpreting and translation from an Intermodal perspective. In *Target,* 28: 61–86.

Bernardini, Silvia, A. Ferraresi, M. Russo, C. Collard and B. Defrancq (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In M. Russo, C. Bendazzoli and B. Defrancq (eds.): *Making way in corpus-based interpreting studies.* Singapore: Springer Nature, 21–42.

Chmiel, Agnieszka, D. Koržinek, M. Kajzer-Wietrzny, P. Janikowski, D. Jakubowski and D. Polakowska (2022): Fluency parameters in the Polish Interpreting Corpus (PINC). In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska and S. Bernardini (eds.): *Mediated discourse at the European Parliament: Empirical Investigations.* Berlin: Language Science Press, 63–91.

Heafiel, K. (n.d.). KenLM toolkit https://kheafield.com/code/kenlm/.

Karakanta, A., M. Vela and E. Teich (2018). Europarl-UdS: Preserving metadata from parliamentary debates. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, May 2018.

Resource available at: https://fedora.clarin-d.uni-saarland.de/europarl-uds/

Macháček, D., M. Žilinec and O. Bojar (2021). *ESIC 1.0 - Europarl Simultaneous Interpreting Corpus,* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3719.

Menzel, K., H. Przybyl and E. Lapshinova-Koltunski (forthcoming). EPIC-UdS – ein mehrsprachiges Korpus als Grundlage für die korpusbasierte Dolmetsch- und Übersetzungswissenschaft. In *Proceedings of the 4th TRANSLATA Conference, 2021*, Innsbruck.

Przybyl, H., E. Lapshinova-Koltunski, K. Menzel, S. Fischer and E. Teich (2022b). EPIC UdS – Creation and applications of a simultaneous interpreting corpus. In *Proceedings of 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 1193–1200, Marseille, June 2022. Resource available at: https://fedora.clarin-d.uni-saarland.de/epic-uds/index.html

Russo, M., C. Bendazzoli, A. Sandrelli and N. Spinolo (2012). The European Parliament Interpreting Corpus (EPIC): Implementation and developments. In F. Straniero Sergio and C. Falbo (eds.): *Breaking ground in corpus-based interpreting studies.* Bern: Peter Lang, 53–90.

SAMUELS Consortium (2015). Hansard corpus. SAMUELS Project, available via https://www.english-corpora.org/hansard/