Recibido / Received: 24/05/2020
Aceptado / Accepted: 10/08/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.05

Para citar este artículo / To cite this article:
Kajzer-Wietrzny, Marta & Łukasz Grabowski. (2021) "Formulaicity in constrained communication: An intermodal approach" In: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection. MonTI* 13, pp. 148-183.

FORMULAICITY IN CONSTRAINED COMMUNICATION: AN INTERMODAL APPROACH

MARTA KAJZER-WIETRZNY
kajzer@amu.edu.pl
Adam Mickiewicz University, Poland

ŁUKASZ GRABOWSKI
lukasz@uni.opole.pl
University of Opole, Poland

Abstract

In this exploratory study bordering on corpus linguistics, formulaic language and studies on constrained communication (focusing on translation, interpreting, and L2) we aim to verify whether constrained texts found in the Polish-English component of an intermodal EPTIC corpus differ from native texts in terms of use of adjacent word combinations commonly known as bigrams and whether similar patterns can be found across spoken and written registers. To that end, we fit a Poisson regression model with fixed and random effects. The results show that the translated language variety contributes to the higher number of the most frequent bigram types in both spoken and written registers, and that the number of frequent bigrams in texts generally increases when the speech/source speech is delivered impromptu, but the effect is significant only for the written register. The findings reveal the considerable impact of individual variation on formulaicity as most of the bigram variation within both models is explained by text-specific random variables rather than fixed variables.

Keywords: Formulaic language; Interpreting; Translation; Constrained language; Corpus linguistics.

Zusammenfassung

In dieser an Korpuslinguistik, Formelsprache und Studien über eingeschränkte Kommunikation grenzende Forschungsstudie, die sich hier auf Übersetzung, Dolmetschen und L2 konzentriert, wollen wir überprüfen, ob die in der polnisch-englischen Komponente eines intermodalen EPTIC-Korpus gefundenen eingeschränkten Texte sich von den einheimischen Texten unterscheiden in Bezug auf die Verwendung benachbarter Wortkombinationen, oft Bigrams genannt, und ob ähnliche Muster in gesprochenen und geschriebenen Registern gefunden werden können. Dazu erarbeiteten wir das Poisson-Regressionsmodell mit festen und zufälligen Effekten. Die Ergebnisse zeigen, dass die übersetzte Sprache zur höheren Anzahl der häufigsten Bigram-Typen sowohl in gesprochenen als auch in geschriebenen Registern beiträgt und dass die Anzahl der häufigen Bigrams in Texten generell zunimmt, wenn die Sprache / Quellsprache spontan geliefert ist, aber der Effekt ist lediglich für das schriftliche Register statistisch signifikant. Abschließend zeigen die Ergebnisse einen signifikanten Einfluss der individuellen Variation auf die Formelsprache, da man den größten Teil der Bigram-Variation in beiden Modellen eher durch die textspezifischen zufälligen Variablen als durch die festen Variablen erklärt.

Schlagwörter: Formelsprache; Dolmetschen; Übersetzung; Eingeschränkte Sprache; Korpuslinguistik.

1. Introduction

In the last two decades research on translation and interpreting has provided ample support for the assertion that there is no unified way or method of approaching translational and non-translational texts. Consequently, contemporary Translation/Interpreting Studies resemble a cluster of overlapping perspectives, e.g. formal, pragmatic, psycholinguistic, neurolinguistic or corpus linguistic, etc. Following interest in corpus-based and corpus-driven research on translation and interpreting universals (Baker 1993; Laviosa 1998, 2002; Mauranen 2000; Olohan 2004; Chesterman 2004; Ulrych & Murphy 2008; Kajzer-Wietrzny 2012; De Sutter et al. 2013; Grabowski 2013; Biel 2014; Szymor 2018), i.e. repeatedly observed characteristics of translations, more attention has been paid recently to the concept of 'constrained communication' (Kruger 2012, Kruger & Van Rooy 2016a, Kotze 2019), where language use is constrained by mediation (translation/interpreting), foreign language use or both. For example, Lanstyák and Heltai (2012)

hypothesize that both translation and non-native production share the main constraint, i.e. the need to manage two languages and the ensuing "linguistic uncertainty resulting from the parallel activation of two languages". At the same time, they point out that constrained varieties differ in that non-native language/text production involves descriptive language use (i.e. it does not depend on any other text), translation being additionally constrained by interpretive language use (i.e. it is dependent on the source text).

Current research on translated English and non-native English appears to validate the view that there are similar linguistic tendencies with respect to "features resulting from processing strain" (Kruger & Van Rooy 2016a: 26). Among the constrained varieties, translation is usually viewed as the extreme case of bilingual activation and perceived as particularly constrained at the psycholinguistic level due to rapid bi-directional switching between languages and activation both at the level of language in general as well as the specific linguistic variants of the source text (Kruger & van Rooy 2016b: 121). On these grounds, we can argue that simultaneous interpreting is an even more extreme case due to the time constraint, which makes the entire process more rapid than written translation. Thus, it is imperative that the analysis be expanded to include interpreting, as in many respects it shows different linguistic patterns compared to written translation (cf. Sandrelli & Bendazzoli 2005; Shlesinger & Ordan 2012; Defrancq et al. 2015; Kajzer-Wietrzny 2015; Bernardini et al. 2016; Ferraresi et al. 2019). For the same reason, spoken non-native texts should also be included in this paradigm because, like interpreting, such texts are not subject to intermediate intervention (e.g. editing). That is why they may also reveal peculiar linguistic patterns.

The rapidly growing literature on constrained communication also points to shared cognitive limitations in the production of non-native and translated texts and, as pointed by Aston (2018: 84-85, after Forster 2001), cognitive resources seem to be liberated by the use of formulae which are also believed to be used in greater proportions in settings requiring more processing effort. In an exploratory study of interpreter discourse in the European Parliament, Aston (2018: 83) looks at the frequency of n-grams with 5 words or longer found in transcripts of simultaneous interpretations and argues that "the language of fluent interpreters relies heavily on recurrent formulaic phraseologies." As the formulaic repertoire of second language speakers is

supposed to be smaller than that of native speakers, Aston (2018: 83) points "to the need for interpreters working into their second language to enlarge this repertoire as far as possible", especially that linguistic preferences of translators and interpreters do not always reflect native speakers' preferences manifested, among others, in the use of the so-called formulaic language.

Although the concept of 'formulaic language' (or 'formulaicity') has been explored by linguists of various schools and research traditions as well as with various purposes in mind (descriptive, applied or otherwise), the debate about its theoretical status has been rather inconclusive and there has been little agreement as to its precise definition and operationalization (Wray 2002, 2007; Schmitt & Carter 2004; Wood 2015; Forsyth & Grabowski 2015; Buerki 2016, 2020; Myles & Cordier 2017; Pęzik 2018; Nelson 2018; Siyanova-Chanturia & Omidian 2019; Szerszunowicz 2020). Consequently, a wide variety of criteria is used in the identification and classification of various manifestations of formulaic language in texts, e.g. distributional (frequency, distribution range, collocational strength measures), syntactic (fixed versus flexible word order, substitutability), semantic (non-/compositionality of meaning), pragmatic (genres, registers etc.), to name but a few. That is why 'formulaic language' acts as an umbrella term for the many different types of linguistic items or operationalizations of recurrent patterns of language use, such as collocations, bigrams, binomials, multi-word verbs, speech formulae, routine formulae, pragmatic routines, pragmatemes, lexical bundles, idioms, winged words, proverbs, sayings, clichés etc.

As in this paper we adopt a textual, quantitative corpus linguistic perspective on constrained communication, frequency and repetition become the main criteria for us to identify formulaic language. As such, frequency constitutes a statistical property of multi-word combinations because language users, be it in translation, interpreting or native language use, generally give priority to the linguistic items that are frequently used in their discourse communities. Moreover, since formulaic phrasings are inherently repetitive, we believe that focusing on frequent bigram types will provide a cursory insight into the amount of formulaic language in the study corpus, similar to Altenberg's (2018) research on recurrent n-grams. Furthermore, the frequency-driven approach to study formulaic language is particularly attractive for the analyses of routinized or clichéd texts because such texts

rely more on restricted sets of prefabricated text chunks, notably when compared with more creative texts (literary or otherwise) (Forsyth & Grabowski 2015). Hence, the frequency-driven approach focusing on the use of contiguous sequences of words (e.g. bigrams, trigrams) seems to be well justified when exploring the properties of somewhat restricted and clichéd European Parliament discourse (Kajzer-Wietrzny 2012).

Thus, in this exploratory study, which interfaces corpus linguistics, formulaic language and studies on translation, interpreting as well as L2, we aim to verify whether constrained spoken texts (read out and delivered impromptu) differ from native spoken texts in terms of use of adjacent word combinations (bigrams). We look at the formulaicity of texts produced in English by native English speakers and native speakers of Polish as well as that from interpreters at the European Parliament working into their B (L2) language and Polish-English translations of the European Parliament debates. The study aims to verify whether such constrained texts differ from native texts regarding the number of most frequent bigram types and whether similar patterns can be found across spoken and written registers. More precisely, we put forward the claim that, due to increased processing constraints, interpreters, translators and non-native speakers rely more on the use of formulaic language (operationalized as the number of bigram types among the most frequent bigrams in the registers under study) than native speakers, and that the mode of delivery of the text and delivery rate, particularly in the case of spoken production, might impact the number of distinct bigram types, which is our working hypothesis. In other words, the discussion presented in this paper focuses on the factors that impact the use of formulaic language in constrained communication with the European Parliament discourse as a case in point. In what follows, we describe the research material and methodology of our study in greater detail.

2. Translation, interpreting and non-native language as forms of constrained communication

Constrained language is an umbrella concept marrying two independent research directions focusing on translation and foreign language. It recognizes the shared cognitive constraints in those two communicative situations

involving bilingual activation, which may help identify their shared linguistic characteristics. As already mentioned, while translation is source text dependent, non-native production is not at the same level. Both translation and non-native language use are constrained by parallel bilingual activation and the ensuing linguistic uncertainty (Lanstyák & Heltai 2012). Kruger and van Rooy (2016a) suggest also that the common denominator of the constrained language varieties is the "transfer or cross-linguistic influence (CLI)". It can therefore be expected that patterns observed in one form of language contact may be reflected in other language contact conditions.

Even though the existence of such links was suggested over a decade ago (e.g. Halverson 2003; Chesterman 2004), it has only recently been addressed in empirical investigations of different instances of constrained communication together with factors such as "processing complexity and cognitive effort, (communicative) risk avoidance, and cross-linguistic influence (CLI)" (Kruger & De Sutter 2018: 252). Kruger (2018: 10) argues that

> "constrained varieties may be seen as probabilistically conditioned by five overarching and interacting constraint dimensions (conceived as continua rather than binaries), enabling us to model the similarities and differences between varieties:
> (1) Language activation (monolingual—bilingual)
> (2) Modality and register (spoken—written—multimodal) (3) Text production (independent/unmediated dependent/mediated)
> (4) Proficiency (native/proficient—non-native/learner)
> (5) Task expertise (expert—non-expert)".

As this research direction is relatively new, the studies addressing those five constraint dimensions are still relatively scarce and mostly limited to written register. Also, the very few studies conducted so far focus on the comparisons of texts written in the English language. For example, Kruger and Van Rooy (2016a: 26) showed that translated English and non-native written English show similar tendencies with respect to "increased formality, explicitation of information through elaboration and specification, and features resulting from processing strain". Expertise and proficiency also play a role as "less advanced non-native varieties and translated texts avoid informality features in written registers to a much larger degree than more advanced

non-native varieties and native varieties", and that this tendency, which is likely to be caused by a risk-avoidance strategy, diminishes with greater proficiency (Kruger & Van Rooy 2018: 237). A similar conclusion seems to transpire from the study by De Sutter and Lefer (2020), who examined the use of explicit variant (*that* vs. zero complementizer) and observed that it is most often chosen by learners with little writing experience, followed by less experienced native writers, only then by translators and non-translators. What is more, the "the two groups of professionals hardly differ, although in some very specific contexts translators use explicit that somewhat more often than non-translators" (De Sutter & Lefer 2020). Not only is explicitation more frequent in constrained communication, but also certain structures indicating implicit syntactic relationships are underused when compared to original native texts (Ivaska et al. submitted). In other words, it seems that non-native authors use less implicit relationships than translations and the pattern is consistent across different registers. In a similar vein, Rabinovich et al. (2016: 1871) show that lexical richness of constrained varieties is lower, idiomatic expressions and pronouns are differently distributed and the proportion of more frequent words is much higher as well as that of cohesive devices.

Studies showing multimodal approaches to constrained communication are still few and far between, but they seem to confirm that non-native and translated texts share a common ground also in the spoken register. A small-scale study of non-native and interpreted texts (Kajzer-Wietrzny 2018: 111) shows that a tendency to an increased frequency of optional connective *that* can be observed in both spoken varieties of constrained communication. Kajzer-Wietrzny et al. (2019) observe that mediation has an equalizing effect on formality differences causing the mediated written and spoken varieties to be closer to each other on the formality spectrum than the native non-mediated written and spoken varieties.

Another study on lexical diversity in constrained language examined through the lens of lexical density, variability, evenness, dispersion, rarity and semantic disparity (Kajzer-Wietrzny & Ivaska 2020) confirms that both spoken and written constrained texts show a tendency similar to the "equalizing effect". First observed by Shlesinger (1989) with reference to

interpreting, it is supposed to diminish "the orality of markedly oral texts and the literateness of markedly literate ones". Moreover, constrained texts in general tend to shift towards the middle of the involved vs. informational speech production continuum. On the other hand, interpreted and translated texts show a greater uniformity, i.e. are "more like each other" (as observed by Baker 1996 and Laviosa 1998 with reference to translations), which hints at the possibility of translation-specific levelling-out effect.

It is also important to note that, especially in the context of the European Parliament, the mode of speech delivery also affects the patterns of language use in mediated discourse, both written and spoken. While orthographic transcripts[1] are considered source texts of spoken mediated texts (English interpretations), verbatim reports drafted in Polish and available at the EP website constitute the sources of written mediated texts (English translations). Moreover, it is hypothesized here that the mode of delivery of a source event (i.e. the original speaker delivering a speech at the European Parliament) impacts characteristics of both spoken (i.e. orthographic transcripts of the speech) and written texts (i.e. verbatim reports in all language versions). The impact of mode of delivery on interpreting seems to be more direct, but it is plausible that at least a selection of typically oral or typically written features that can be attributed to the mode of delivery of source events are transferred also to the target texts of translations of the verbatim reports of these. This is reflected, for example, in the lexical diversity of both simultaneous interpretations of speeches delivered at the EP as well as translations of the verbatim reports of these speeches (Kajzer-Wietrzny & Ivaska, 2020) and in cohesion patterns in these texts (Kajzer-Wietrzny, accepted). The impact of this factor seems worth investigating also in the context of formulaicity.

1. While orthographic transcripts of the source and target speeches were manually produced in the compilation process of the EPTIC corpus (Section 3), verbatim reports of the source speeches and their translations into the EU official languages have been published at the EP website (translations into EU official languages are available for all the speeches given at the EP until mid-2011).

In another study (Kajzer-Wietrzny, accepted) we also show the tendency towards increased cohesion in constrained varieties, which is, however, realized in different ways. The overall frequency of cohesive devices (excluding phrase-level coordinators) points to a significantly higher number of cohesive devices in translations when compared to native and non-native texts. A similar significant effect, albeit slightly weaker, is visible in interpretations. Non-natives do increase the overall level of cohesion of their utterances in the spoken register with an overuse of phrase-level coordinators. All those findings encouraged us to undertake a further study, this time focusing on formulaicity in spoken and written unconstrained and constrained language varieties. We believe that the patterns of use of bigrams (as we operationalize formulaic language) will cast more light on the specificity of constrained communication.

3. Methodology

3.1. Research material

The research material includes the Polish-English components of the European Parliament Translation and Interpreting Corpus[2] (henceforth EPTIC), which is an intermodal corpus rich in contextual information (e.g. speaker, delivery rate, mode of delivery of the text/source text). The texts compiled in EPTIC include speeches first delivered at the plenary sittings of the European Parliament by MEPs or Commissioners and simultaneously interpreted into official EU languages. Subsequently, verbatim reports were drawn and until 2011 they were also translated and published on the EU Parliament website (Figure 1).

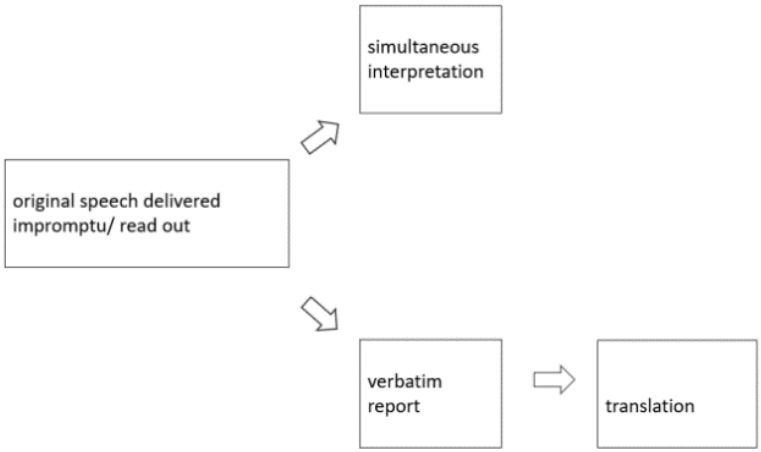

Figure 1. Text cycle at the European Parliament (adapted from Defrancq et al. 2015: 202)

EPTIC sub-corpora, which enable one to conduct a number of different comparisons (e.g. interpretations vs translations, interpreted vs non-interpreted language, native English vs non-native English etc.), include the following:

- sources – spoken: orthographic transcripts of the original speeches;
- sources – written: official verbatim reports of the source speeches;
- targets – interpreted: orthographic transcripts of the interpretations;
- targets – translated: translations of the verbatim reports.

The present study was carried out on a dataset comprising English speeches as well as English interpretations and translations from Polish selected from EPTIC and augmented with two corpora of non-native speeches delivered by Polish representatives (MEPs and Commissioners) delivering speeches in English at the European Parliament[3]. In total, the study corpus comprises 250 texts with 59,540 words (tokens), which are divided into two sub-corpora representing spoken and written registers. These, in turn, are further

3. While the core EPTIC files are based on speeches delivered at the European Parliament in 2011, the two additional corpora contain speeches delivered in 2010 and 2011.

subdivided into native English-originals, non-native English originals and interpretations[4]/translations from Polish into English (Table 1).

Spoken	Written				
Native English Originals*	Non-native English Originals**	Interpretations from Polish into English*	Native English Originals*	Non-native English Originals**	Translations from Polish into English*
9,487 w 34 texts	9,869 w 33 texts	9,567 w 58 texts	9,200 w 34 texts	9,703 w 33 texts	11,714 w 58 texts

** Components of EPTIC*
*** Corpora compiled according to EPTIC guidelines*

Table 1. Analysed dataset

3.2. Unit of analysis, methods and procedures

Apart from the fact that n-gram models, i.e. models based on contiguous sequences of *n* words, have been effective in general in modelling language data in various statistical natural language processing applications, we used bigrams (2-word sequences) as the unit of analysis because they have also been used as indicators of formulaic language in texts (Altenberg 1998). Although not all bigrams represent neat form-and-meaning pairings, they nevertheless tap into the most important aspects of formulaic language (from the corpus linguistic perspective seen primarily as recurrent use of fixed or semi-fixed multi-word units in texts), such as frequency and fixedness (Schmitt & Carter 2004; Wood 2015; Pęzik 2018; Siyanova-Chanturia & Omidian 2019). Also, the frequency-driven approach to study formulaic language is particularly useful for the analyses of clichéd texts because such texts rely more on limited stocks of prefabricated text chunks or boilerplate conventional formulas (Forsyth & Grabowski 2015). Furthermore, Nesi (2012: 422) claims that "n-grams in spoken and written texts tend to be constituted differently [...], and some genres are more formulaic than others". Another rationale behind focusing on bigrams rather than longer sequences of *n* words (e.g. trigrams, fourgrams) is the limited size of the study corpus and, consequently, the problem of data scarcity. Our preliminary inspection

4. Interpretations were carried out from Polish into English by native Polish interpreters.

of the lists of trigrams and fourgrams showed that their number was not sufficient for a large-scale statistical analysis, which would be feasible only with a larger study corpus.

Although in recent years corpus-based research on formulaic language in translation has been flourishing, most studies have been primarily descriptive rather than explanatory and pertained to native versus non-native distinction (e.g. Hu et al. 2016, Ebeling & Ebeling 2018). In this study, we aim to also address the distinction between constrained and unconstrained language as well as attempt to identify those text-related factors that condition the degree of formulaicity (operationalized as the number of bigram types) in spoken and written constrained texts under scrutiny. As mentioned earlier, in this study we investigate formulaicity only within the most frequent bigram types used in the registers under scrutiny. The tools used in the study include Formulib software package (Forsyth 2015), R (2013) and ad hoc scripts written in Python.

We explored formulaicity by identifying the 400 most frequent bigrams in spoken and written sub-corpus, which – given the small size of the sub-corpora – provides a sufficient number for an analysis. In order to avoid a topic bias, we decided to remove from the list all the bigrams that perform referential functions, such as proper names (e.g. *Lady Ashton, of Congo, in Poland, Mr Lukashenko*) or bigrams related to topics of particular speeches (e.g. *construction products, cohesion policy, foreign policy, Christians in*) as these were bound to be more dictated by the discussed problem than the hypothesized cognitive processes that might constrain the investigated forms of bilingual communication. The manual filtering procedure resulted in the selection of 354 and 352 bigrams in spoken and written registers, respectively. From those two samples, we selected those bigrams that were found in all the sub-corpora in the spoken (215 bigram types) and written (237 bigram types) dataset under scrutiny. Next, using ad hoc scripts written in Python, we checked whether each bigram type occurred in each text, which eventually enabled us to count the number of these highly frequent bigram types in each text in each sub-corpus.

As in this paper we focus on identification of factors/predictors that impact the number of bigram types (count data, i.e. non-negative integer

numbers), we used a Poisson regression model[5]. In short, Poisson regression is a type of a Generalized Linear Mixed Model (GLMM) that is typically employed to model count data and contingency tables (Winter 2019: 247), which in this study are matrices of bigram counts. We hypothesize that these counts depend on multiple independent variables (predictors), e.g. mode of delivery or delivery rate, which are our fixed effects. Importantly, predictors in Poisson regression models can be a mixture of numeric and categorical variables. As in any GLMM model, an individual slope in Poisson regression models provides an estimate of the multiplicative change in the response variable (e.g. the number of bigram types) for a one-unit change in the corresponding predictor (e.g. a delivery rate) (Scherber 2019b). For example, if the slope equals -0.12 then for a one-unit change (1 word per minute) in delivery rate the number of bigram types decreases $e^{-0.12}$ fold. Since we have a potentially large pool of speakers, translators, interpreters and topics of the texts under scrutiny (due to the number of observations we cannot include all of them within our model), we decided to include Text IDs as random effects into our model[6]. Without them we would risk having loads of unaccounted variation. Our analysis is thus based on a mixed-effects model and in order to fit it in R (2013), we used *lme4* package (Bates et al. 2015).

Bentz and Winter (2013) describe assumptions to be met in this type of analysis. For example, random effects in mixed models should have 5 to 6 levels at a minimum, which is a criterion that has been met in all models analyzed here[7]. Similarly, an "important assumption of the Poisson distribution is that the sample mean and the sample variance are identical" (Bentz & Winter 2013) applying to distribution of a response variable, which in this study is a count variable. If sample variance exceeds the mean it indicates overdispersion. In the current analysis, none of datasets (neither spoken

5. For more on statistical modeling (linear models, generalized linear models and mixed models), see, Hastie et al. (2016), Kuhn & Johnson (2018), Scherber (2017, 2019a, 2019b), Winter (2019), the latter focusing primarily on linguistic data.
6. We have only included random intercepts, as the inclusion of random slopes was impossible due to an insufficient number of observations.
7. More precisely, our random effects (Text ID) constitute a factor with *n* levels (particular Text IDs) which come from a probability distribution because, potentially, we had infinite number of levels from which our texts could have come (although the EPTIC corpus is restricted in size and composition).

nor written) showed signs of overdispersion as shown by the overdisp_fun[8]. Another important issue involves zero-inflation, which is the case when there is an excessive number of zero-occurrences in the dataset. None of the reported regression models showed any sign of significant zero-inflation and none of the fixed or random effects were highly correlated.

Model summaries and R2m and R2c point to how much of the variation in the data is explained by the fixed effects accounted for in the model and how much can be explained by the full model including random intercepts (see Appendix 2 and Appendix 3). Marginal and conditional R2 were calculated in R with the MUMIN package (Barton 2019). It is also worthwhile emphasizing that no likelihood ratio tests aiming at establishing the contribution of single effects to the model were carried out as, according to Bolker et al. (2009: 132), "the LR test is not recommended for testing fixed effects in GLMMs, because it is unreliable for small to moderate sample sizes." In such cases, Bolker et al. (2009: 132) "recommend against using the LR test for fixed effects unless the total sample size is and number of blocks are very large", which is not the case in the reported study. Additionally, the variables included in the analysis were theoretically motivated and therefore we did not conduct any model comparison.

3.3. Research questions, hypotheses and study stages

This paper is an attempt to explore formulaicity – operationalized as the number of most frequent bigram types – in constrained communication using the European Parliament discourse as a case in point. The study aims to provide answers to the following research questions:

1. What is the most important factor/predictor (language variety, mode of delivery, delivery rate[9]) that impacts the degree of formulaic language in constrained communication versus native texts?
2. Are the observed patterns the same across spoken and written registers in the case of constrained and native texts?

8. Bolker et al. (2020). GLMM FAQs. (URL: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#overdispersion)
9. Delivery rate as a variable will only be examined in the models regarding the spoken register.

We expect that, due to increased processing constraints related to bilingual processing and/or interpretive language use, interpreters, translators and non-native speakers rely more on the use of formulaic language – measured by the number of bigram types among the most frequent bigrams in the registers under study – than native speakers, and that the mode of delivery of the text and delivery rate, particularly in the case of spoken production, might impact the number of distinct bigram types. Additionally, we hypothesize that increased speed of delivery may contribute to a greater processing effort in the spoken form of constrained communication, which has already been proved in interpreting (Plevoets & Defrancq 2016).

The study will be conducted in a number of stages. First, we will fit Poisson regression models with fixed and random effects to the data obtained from the spoken register[10]. Next, we will repeat the same procedure as applied to the written register. In the last stage, we will compare the results and discuss their implications, paying attention to limitations of this study. In what follows, we present the study findings.

3. Results

As mentioned earlier, in order to provide answers to the research questions, we fitted Poisson regression models with fixed and random effects. The total number of bigram types was modelled as a function of the following predictor variables: text variety, mode of delivery of the source (and speed of delivery in the case of spoken register) adjusted by an exposure variable, which is in this case the number of bigrams in individual text (z-scored). Text-specific random intercepts were also included for the effect of text variety and mode of delivery (and speed of delivery in the case of spoken register) on the number of bigram types (only random intercepts, as the inclusion of random slopes was impossible due to an insufficient number of observations). The source texts of interpretations and translations were in many respects identical[11] and therefore the models for spoken and written registers

10. The dataset, together with the statistical analyses, can be accessed at: https://osf.io/7ktm8/
11. When compared to orthographic transcripts, the verbatim reports analyzed here lack the typical features of orality e.g., repetitions, truncated words etc.; syntactic adjustments are mostly made in the case of discontinued sentences and lexical

were fitted separately. In all models, native English speeches (spoken or written) are used as intercepts.

3.1. Number of bigram types in spoken register

We start by looking at the number of the most frequent bigram types in constrained and non-constrained spoken registers. The first model[12] estimates (1) how the number of bigram types changes as a function of the fixed predictor variables, i.e. text variety, mode of delivery and speed of delivery of the original speech (expressed in words per minute) adjusted by an exposure variable: the number of bigrams in individual text (z-scored) and (2) the variability among the levels of the random effect, i.e. individual texts.

Effect plots (Figure 2) illustrate the general tendencies that can be inferred from the GLMM modelling the patterns in spoken register. First, as visible in the Text Variety effect plot, the number of most frequent distinct bigram types increases with the number of constraints that the users of language have to handle. Thus, said number is the lowest in the spoken native variety, where the speakers are not constrained by either bilingual processing or by the message of the source text; it is higher among the speakers of a foreign language and the highest among interpreters, who transfer someone else's message into a non-native tongue. Second, delivery effect plot shows that when the speakers deliver their speech impromptu, or interpreters interpret a speech that was delivered by the original speaker impromptu, the use of distinct most frequent bigram types in a text increases. The tendency is reverse when the speeches are read out. Finally, the faster the speakers deliver their speeches or the faster the original speakers deliver the speeches that interpreters interpret, the greater the use of distinct

changes are rare. The exact scope of the changes from the spoken parliamentary discourse to the written representation in the analyzed dataset was not measured. It is likely, however, that as in the case of the Hansard, the written representation of the EP debates is not a "hazard" (Mollin 2007) for many linguistic features of interest (Kotze et al. in review).

12. Bigramsspoken <-glmer(CommonBigramTypesNumber~TextVariety+Delivery+ST-WPM+offset(TotalBigramsInText)+(1|TextID)

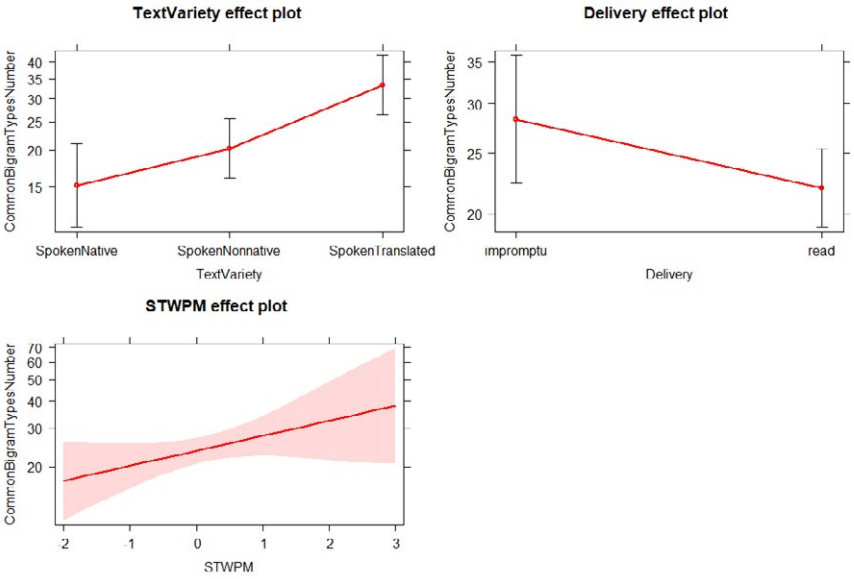

Figure 2. The number of bigram types as a function of fixed and random effects in spoken register

most frequent bigram types, which in this study are used to operationalize formulaicity.

It is worth noticing, however, that not all these trends are statistically significant, which can be inferred from the model summary (see Appendix 2). The outputs of the analysis based on a generalized linear mixed model with Poisson distribution show that within the spoken register both constrained varieties, i.e. non-native language speech and simultaneous interpretations, are characterized by a higher number of distinct most frequent bigram types than spoken native English texts (as illustrated in Figure 2). Still, with estimates at 0.2900 (p=0.14823) in the case of non-native English speeches and 0.7896 (p=0.00127) in the case of interpretations into English, only the difference between the intercept and the latter is statistically significant. The impact of the mode of delivery of the source (i.e. whether the source speech was read out or delivered impromptu) on the number of different bigram

types in a text approaches statistical significance (p=0.07497). Read out speeches, though, seem to contain, in general, a smaller number of different bigram types (estimate -0.2530) than impromptu speeches. In general, the speed of delivery of the (original) speech increases the number of bigram types in a text but its impact is not significant (estimate 0.1579, p=0.11763). It needs to be noted that in the reported regression analysis a large share of variation within the data was explained by the full model including both fixed effects and random intercepts, while fixed effects account only for almost 15% of the variation (as indicated by R^2m). This means that individual text-related effects contributed most to the variation of the number of bigram types (see Appendix 2 for full results). This observation accords with our decision to include the random effects (text ID) into the model as, without it, we would not have been able to capture loads of variation in the model.

3.2. Number of bigram types in written register

Let us now inspect the number of most frequent bigram types in constrained and non-constrained written registers. The second model[13] estimates (1) how the number of bigram types changes as a function of the fixed predictor variables, i.e. text variety and mode of delovery of the original speech adjusted by an exposure variable: number of bigrams in individual text (z-scored) and (2) the variability among the levels of the random effects, i.e. individual texts.

Effect plots (Figure 3) illustrate the general tendencies that can be inferred from the second GLMM. First, the TextVariety effect plot shows that the number of distinct most frequent bigram types increases with the number of constraints that the users of language have to deal with. It can be seen that said number is the lowest in the written native variety, where the authors are constrained by neither bilingual processing nor the message of the source text; it is higher among the authors using a foreign language, and it is the highest among translators, who transfer someone else's message (with the caveat that their native tongue is unknown). Second, the Delivery effect plot shows that when the text represents a verbatim report of a speech that was originally delivered impromptu or when translators translate a speech

13. Bigramswritten <- glmer(CommonBigramTypesNumber~TextVariety+Delivery+offset(TotalBigramsInText)+(1|TextID)

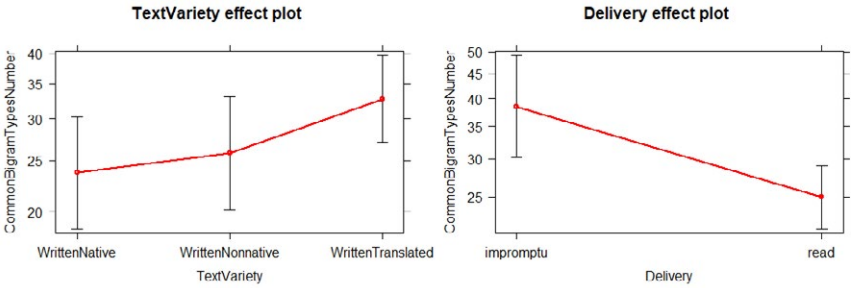

Figure 3. The number of bigram types as a function of fixed and random effects in written register

that was delivered by the original speaker impromptu, the use of distinct most frequent bigram types in a text increases. The observed tendency is reverse when the speeches are read out.

Similarly to the previous model, not all the tendencies are statistically significant. The outputs of the analysis based on a generalized linear mixed model with Poisson distribution (see Appendix 3) show that within the written register both constrained varieties, i.e. non-native language speech and simultaneous interpretations, are characterized by a higher number of different bigram types than written native English texts. Again, with estimates at 0.08581 (p=0.62790) in the case of non-native English texts and 0.32417 (p=0.04154) in the case of translations into English, only the difference between the intercept and the latter is statistically significant. The impact of the mode of delivery of the source (i.e. whether the source speech was read out or delivered impromptu) on the number of different bigram types in a text is statistically significant (p=0.00329) with the read out speeches containing, in general, a smaller number of different bigram types (estimate -0.43241) than impromptu speeches. As in the case of the earlier model, a large share of variation within the data was explained by the full model including both fixed effects and random intercepts, whereas fixed effects account only for approximately 11.5% of the variation (as indicated by R^2m). This means that individual text-related effects contributed most to the variation of the number of bigram types (see Appendix 3 for full results).

Similar to the model described in Section 3.1, the decision to include random effects into the model has been justified.

4. Discussion

In this study we were primarily interested in, first, the identification of predictors of formulaic language in constrained versus native texts and, second, the verification whether the same patterns were observed in native and constrained texts, both spoken and written. Formulaicity was operationalized as the number of the most frequent bigram types, which was our response count variable. As for the potential predictors, we focused on text variety, delivery rate, mode of delivery (fixed effects) and text ID (random effects), which we fit in a mixed-effects model using Poisson regression.

It transpires from both regression models applied to spoken and written registers, respectively, that the translated variety is the main predictor of the number of most frequent bigram types in both registers. A similar trend can be observed in the spoken and written non-native variety, but the estimates do not diverge significantly from the intercept, i.e. the native texts. It may also be argued that the constrained varieties across registers seem to pattern together in a similar way, yet they do not differ from the native varieties to the same extent.

As observed by Kotze (2019: 339), the patterns setting translated language apart from non-translated language, in particular the tendencies relating to "cross-linguistic influence, priming or transfer [are] often of subtle and indirect type". Additionally, written text production in the EU setting is heavily standardized, which can further filter out the nuances which are subtle even in genres not subject to such standardization (e.g. literature or journalistic texts). This is also reflected in the results of the present analysis. Fixed effects in the spoken register account for more variation within the data than in the model fitted for the written register (as indicated by a slightly higher value of R2m and lower value of R2c in the spoken model). One of the potential reasons for such tendencies may be the standardizing effect of the editing and proofreading process at the EU institutions, which might render written texts in the studied constrained and non-constrained varieties more similar. Furthermore, the fixed effects in the spoken model

may be augmented by the "double" cognitive constraint imposed on interpreters relating to the process of language mediation and speaking a foreign language, as all interpretations were carried out into L2. As regards the expectations that speakers might attempt to decrease the higher cognitive load related to higher delivery rate with the use of more formulaic language, these have been only partially confirmed. The rate of delivery, indeed, contributes to a higher number of most frequent bigram types in a text, but the effect is not significant.

We also found that the mode of delivery of the (original) speech is a significant predictor in the written register, and it approaches statistical significance in the regression modelling of the spoken one. In general, the impromptu mode of delivery seems to consistently point to an increased use of most frequent bigram types across all varieties. It is clear that the effect of the mode of delivery in spoken register is significant albeit weaker than in written register. This observation ties in with the one made by Shlesinger (1989, cited in Pym 2007: 178) about the equalizing effect of interpreting, which affects "the position of a text on the oral-literate continuum" and ultimately leads to the reduction of the range of this continuum in simultaneous interpreting. This renders markedly oral texts less oral and markedly literate texts less literal. Such tendencies have also been hinted at in other corpus studies on simultaneous interpreting (Dayter 2018, Kajzer-Wietrzny & Ivaska, 2020). Our findings show that the effect of delivery is weaker in the spoken register, meaning that the "equalizing effect" is stronger in the spoken register than in the written one. This is an important implication for interpreter and translator training: register-specific formulaicity features are transferred with varying degree of difficulty across registers and as such they may require additional attention.

5. Final remarks

The results of a quantitative corpus study like this one should be interpreted with caution. It has to be emphasized that the texts included in the EPTIC corpus are, by their very nature, quite short (100-300 words) and, more importantly, the corpus used in this study contains slightly less than 60,000 words (although it is representative of the registers under scrutiny). Also, it

is noteworthy that the very form of translation and interpreting, and hence the study results, could be influenced by the translators' or interpreters' idiolects. This can be taken up in further research on corpora annotated with such metadata. Another confounding variable could be the effect of L1: in the case of written translations of the MEPs' speeches, it is impossible to establish whether they were produced by native speakers.

There are many ways in which this study could be continued further in order to provide more comprehensive answers to the question of whether constrained communication is by its very nature more formulaic than unconstrained communication. Apart from focusing on count variables such as the number of bigram types, it is possible to adopt other units of analysis that have been used in research on formulaicity. For example, it is possible to focus on frequencies rather than counts of recurrent multi-word items in texts (bigrams, trigrams etc.). Apart from n-grams, one can also explore formulaicity by exploring phrase frames (Fletcher 2002), which are sequences of *n* words identical except for one and which provide a neat generalization of recurrent sequences of words in texts). Hence, it is possible to explore the predictors of formulaic language by focusing on measures of pattern variability applied to phrase frames, e.g. VPR (variant-to-phrase frame ratio proposed by Römer (2010: 105)), Hapaxity, Haprate etc. (Forsyth & Grabowski 2015). As such, these metrics constitute continuous response variables and require the use of linear regression models to identify their predictors. In this study we used a single unit of analysis (bigram types), yet it might be necessary in the future to combine multiple units of analysis to obtain more comprehensive findings since formulaicity is a multi-faceted phenomenon and it cannot be fixed at a single level of analysis only. Also, it might be possible to further explore the causal relation between formulaic language in interpreting, on the one hand, and other text-external variables (e.g. interpreter's status, direction of interpreting), on the other.

Furthermore, in this study we have explored formulaicity in constrained communication using English language material only. However, as pointed out by Buerki (2020), the degree to which languages feature formulaic material remains unclear, notably in the rather underexplored translation/interpreting context, which invites further cross-linguistic (e.g. English-French or

English-Spanish) corpus linguistic research using topically matched corpora with texts representing constrained communication.

Another unexplored avenue of future research on formulaicity in constrained communication, notably in translation/interpreting, is the transfer of discourse functions from the source texts to translations/interpretations, which has implications on how the message of the translation/interpretation is comprehended as compared with the source text. Preliminary exploratory research into this matter, conducted with the use of inter-rater agreement metrics and focusing on recurrent phrases with specific discoursal functions (stance expressions, discourse organizers, including polyfunctional items, e.g. *at the end of the day*), revealed that the discoursal functions are often not conveyed in a fixed and stable way (Grabowski & Groom, accepted). This implies that oftentimes the source and target texts (be it written translations or interpretations) are pragmatically understood differently by respective readers. It seems, however, that further research is required to study the rationale behind the modification of the discoursal functions[14] of recurrent formulas in translation as compared with the original. As this study accounts for an early step in research on formulaicity in constrained communication, we hope that it will pave the way to more comprehensive empirical research into this matter in the future.

References

ALTENBERG, Berndt. (1998) "On the phraseology of spoken English: The evidence of recurrent word combinations". In: Cowie, Anthony (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 101-122.

ASTON, Guy. (2018) "Acquiring the language of interpreters: A Corpus-based Approach". In: Russo, Mariachiara, Claudio Bendazzoli & Bart Defrancq (eds.), *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, pp. 83-96.

14. This is related to more general questions of interest to translation/interpreting practice, namely why translators/interpreters often fail to recognize phrases as a holistic unit and translate them literally, etc.

BAKER, Mona. (1993) "Corpus linguistics and translation studies: Implications and applications". In Baker, Mona, Francis, Gill & Toginini-Bonelli, Elena (eds.), *Text and Technology. In Honor of John Sinclair.* Amsterdam: John Benjamins, pp. 233-250.

BARTON, Kamil. (2019) "MuMIn: Multi-Model Inference". Online version: <https://CRAN.R-project.org/package=MuMIn>

BATES, Douglas, Martin Mächler, Ben Bolker & Steve Walker. (2015) "Fitting linear mixed-effects models using lme4". *Journal of Statistical Software* 67:1, pp. 1-48. Online version: <https://arxiv.org/abs/1406.582>.

BENTZ, Christian & Bodo Winter. (2013) "Languages with More Second Language Learners Tend to Lose Nominal Case." In Wichmann, Soren & Jeff Good (eds.), *Quantifying Language Dynamics: On the Cutting edge of Areal and Phylogenetic Linguistics*. Leiden: Brill, pp. 96-124.

BERNARDINI, Silvia, Adriano Ferraresi & Maja Miličević. (2016) "From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective." *Target. International Journal of Translation Studies* 28:1, p. 61-86.

BIEL, Łucja. (2014) *Lost in the Eurofog: The Textual Fit of Translated Law*. Frankfurt am Main: Peter Lang Verlag.

BOLKER, Benjamin, Mollie Brooks, Connie Clark, Shane Geange, John Poulsen, M. Henry Stevens & Jada-Simone White (2009) "Generalized linear mixed models: A practical guide for ecology and evolution". *Trends in Ecology & Evolution* 24:3, pp. 127-135.

BUERKI, Andreas. (2016) "Formulaic sequences: a drop in the ocean of constructions or something more significant?" *European Journal of English Studies* 20:1, pp. 15-34.

BUERKI, Andreas. (2020) "(How) is Formulaic Language Universal? Insights from Korean, German and English". In Piirainen, Elisabeth, Natalia Filatkina, Sören Stumpf & Christian Pfeiffer (eds.), *Formulaic Language and New Data Theoretical and Methodological Implications*. Berlin: Degruyter, pp. 103-134.

CHESTERMAN, Andrew. (2004) "Hypothesis about translation universals". In: Hansen, Gyde, Kirsten Malmkjaer & Daniel Gile (eds.), *Claims, Changes and Challenges in Translation Studies*. Amsterdam: John Benjamins, pp. 1-13.

DAYTER, Daria. (2018) "Describing Lexical Patterns in Simultaneously Interpreted Discourse in a Parallel Aligned Corpus of Russian-English Interpreting

(SIREN)". *FORUM. Revue Internationale d'interprétation et de Traduction / International Journal of Interpretation and Translation* 16:2, pp. 241-264.

DEFRANCQ, Bart, Koen Plevoets & Cedric Magnifico. (2015) "Connective Items in Interpreting and Translation: Where Do They Come From?". In: Romero-Trillo, Jesus (ed.), *Yearbook of Corpus Linguistics and Pragmatics*. Bern: Springer, pp. 195-222.

DE SUTTER, Gert & Lefer, Marie-Aude (2020) "On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multi-factorial and interdisciplinary approach". *Perspectives*, 28:1, pp. 1-23. https://doi.org/10.1080/0907676X.2019.1611891

EBELING, Jarle & Signe Oksefjell Ebeling. (2018) "Comparing n-gram-based functional categories in original versus translated texts". *Corpora* 13:3, pp. 347-370.

FERRARESI, Adriano, Silvia Bernardini, Maja Miličević & Marie-Aude Lefer. (2019) "Simplified or Not Simplified? The Different Guises of Mediated English at the European Parliament." *Meta: Journal Des Traducteurs / Translators' Journal* 63:3, pp. 717-738.

FORSYTH, Richard. (2015) "Formulib: Formulaic Language Software Library". Online version: http://www.richardsandesforsyth.net/zips/formulib.zip

FORSYTH, Richard & Łukasz Grabowski. (2015) "Is there a formula for formulaic language?" *Poznań Studies in Contemporary Linguistics* 54:1, pp. 511-549.

FOSTER, Pauline. (2001) "Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers". In Bygate, Martin, Peter Skehan & Merill Swain (eds.), *Researching pedagogic tasks: Second language learning, teaching and testing*. London: Longman, pp. 75-93.

FOX, John, Sanford Weisberg, Michael Friendly, Jangman Hong, Robert Andersen, David Firth & Steve Taylor. (2019) Package 'effects'. Online version: https://cran.r-project.org/web/packages/effects/effects.pdf

GRABOWSKI, Łukasz. (2013) "Interfacing corpus linguistics and computational stylistics: translation universals in translational literary Polish". *International Journal of Corpus Linguistics*, 18:2, pp. 254-280.

GRABOWSKI, Łukasz & Nicholas Groom (accepted) "Functionally-defined recurrent multi-word units in English-to-Polish translation: a corpus-based study". *Revista Española de Linguistica Aplicada/Spanish Journal of Applied Linguistics*.

HALVERSON, Sandra. (2003) "The cognitive basis of translation universals". *Target. International Journal of Translation Studies* 15:2, pp. 197-241.

HASTIE, Trevor, Robert Tibshirani & Jerome Friedman. (2016) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. Berlin: Springer.

HU, Xianyao, Richard Xiao & Andrew Hardie. (2016) "How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis". *Corpus Linguistics and Linguistic Theory* 15:2, pp. 347-382.

IVASKA, Ilmari, Adriano Ferraresi & Silvia Bernardini. (Under Review) "Syntactic properties of constrained English: A corpus-driven approach".

KAJZER-WIETRZNY, Marta (2021) "Intermodal approach to cohesion in constrained and unconstrained language" *Target*. https://doi.org/10.1075/target.19186.kaj

KAJZER-WIETRZNY, Marta & Ilmari Ivaska (2020) "A multivariate approach to lexical diversity in constrained language". *Across Languages and Cultures* 21:2, pp. 169-194.

KAJZER-WIETRZNY, Marta. (2012) *Interpreting Universals and Interpreting Style*. Unpublished PhD dissertation. Adam Mickiewicz University, Poznań, Poland.

KAJZER-WIETRZNY, Marta. (2015) "Simplification in interpreting and translation". *Across Languages and Cultures* 16:2, pp. 233-255.

KAJZER-WIETRZNY, Marta. (2018) "Interpretese vs. Non-native Language Use: The Case of Optional That". In Russo, Mariachiara, Claudio Bendazzoli & Bart Defrancq (eds.), *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, pp. 97-113.

KAJZER-WIETRZNY, Marta, Ilmari Ivaska, Adriano Ferraresi & Silvia Bernardini. (2019) "Thanks very much President…" or "Thank you Mr President..." Investigating formality in mediated and non-mediated discourse". Paper delivered at 49th Poznań Linguistics Meeting, 16-19 September, 2019 in Poznań, Poland.

KOTZE, Haidee, Minna Korhonen, Adam Smith and Bertus van Rooy. (under review) "Salient differences between oral parliamentary discourse and its official written records: A comparison of "close" and "distant" analysis methods" In Korhonen, Minna, Kotze Haidee & Tyrkkö Jukka (eds.), *Parliamentary discourse across time and space: Using big data to study language and society. Studies in Corpus Linguistics*. Amsterdam: John Benjamins.

KOTZE, Haidee. (2019) "Converging what and how to find out why: An outlook on empirical translation studies". In Vandevoorde, Lore, Joke Daems & Bart Defranq (eds.), *New Empirical Perspectives on Translation and Interpreting*. London: Routledge, pp. 333-371.

KRUGER, Haidee & Bertus Van Rooy. (2016a) "Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English". *English World-Wide* 37:1, pp. 26-57.

KRUGER, Haidee & Bertus Van Rooy. (2016b) "Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans". *Language Sciences* 56, pp. 118-131.

KRUGER, Haidee, & Bertus Van Rooy. (2018) "Register Variation in Written Contact Varieties of English". *English World-Wide* 39:2, pp/ 214–242. doi:10.1075/eww.00011.kru.

KRUGER, Haidee. (2012) "A corpus-based study of the mediation effect in translated and edited language". *Target* 24:2, pp. 355-388.

KRUGER, Haidee. (2018) "Expanding the third code: Corpus-based studies of constrained communication and language mediation." In Granger, Sylviane, Lefer, Marie-Aude & Penha-Marion, Laura (eds.), *Book of abstracts. Using corpora in contrastive and translation studies conference* (5th edition) CECL papers 1. Louvain-la-Neuve: Centre for English Corpus Linguistics/ Université Catholique de Louvain, pp. 9-12.

KUHN, Max & Kjell Johnson. (2013) *Applied Predictive Modeling*. Berlin: Springer.

LANSTYAK, Istvan & Pal Heltai. (2012) "Universals in Language Contact and Translation". *Across Languages and Cultures* 13:1, pp. 99-121.

LAVIOSA, Sara. (1998) "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose". *Meta* 43:4, pp. 557-570.

LAVIOSA, Sara. (2002) *Corpus-based translation studies: theory, findings, applications*. Amsterdam: Rodopi.

MAURANEN, Anna. (2000) "Strange strings in translated language: A study on corpora". In Olohan, Meave (ed.), *Intercultural Faultlines. Research Models in Translation Studies 1:Textual and Cognitive Aspects*. Manchester: St. Jerome Publishing, pp. 119-141.

MOLLIN, Sandra. (2007) "The Hansard hazard: Gauging the accuracy of British parliamentary transcripts". *Corpora* 2:2, pp. 187-210.

MYLES, Florence & Caroline Cordier. (2017) "Formulaic Sequence(fs) Cannot be an Umbrella Term in SLA: Focusing on Psycholinguistic FSs and Their Identification". *Studies in Second Language Acquisition* 39, pp. 3-28.

NELSON, Robert (2018) "How 'chunky' is language? Some estimates based on Sinclair's Idiom Principle". *Corpora* 13:3, pp. 431-460.

NESI, Hillary. (2012) "ESP and Corpus Studies". In: Paltridge, Brian & Sue Starfield (eds.), *The Handbook of English for Specific Purposes*. London: Wiley, pp. 407-426.

OLOHAN, Meave. (2004) *Introducing Corpora in Translation Studies*. Routledge: London.

PĘZIK, Piotr. (2018) *Facets of prefabrication. Perspectives on modelling and detecting phraseological units*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

PLEVOETS, Koen & Bart Defrancq. (2016) "The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis". *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association,* 11:2, pp. 202-224.

PYM, Anthony. (2007) "On Shlesinger 's proposed equalizing universal for interpreting". In Pochhacker, Franz, Jakobsen, Arnt Lykke & Mees, Inger M. (eds.), *Interpreting studies and beyond: A tribute to Miriam Shlesinger.* Copenhagen: Samfundslitteratur Press, pp. 175-190.

RABINOVICH, Ella, Sergiu Nisioi, Noam Ordan & Shuly Wintner. (2016) "On the Similarities between Native, Non-Native and Translated Texts". In van den Bosch, Antal (ed.) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 7-12 August. Stroudsburg, PA: Association for Computing Machinery.

RÖMER, Ute. (2010) "Establishing the phraseological profile of a text type. The construction of meaning in academic book reviews". *English Text Construction* 3:1, pp. 95-119.

SANDRELLI, Annalisa. & Claudio Bendazzoli. (2005) "Lexical Patterns in Simultaneous Interpreting: a Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)". *Proceedings from the Corpus Linguistics Conference Series*. Online version: < https://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx>

SCHERBER, Christoph. (2019a) "An introduction to mixed-effects models". Online version: <http://www.christoph-scherber.de/content/PDF%20Files/Mixed%20effects%20models.pdf>

SCHERBER, Christoph. (2019b) "An introduction to generalized linear models". Online version: <http://www.christoph-scherber.de/content/PDF%20Files/Generalized%20linear%20models.pdf>

SCHERBER, Christoph. (2017) "Using R to Interpret Interaction Effects in Statistical Models". *Software Developer´s Journal*. Online version: <https://www.researchgate.net/profile/Christoph_Scherber/publication/312093784_Using_R_to_Interpret_Interaction_Effects_in_Statistical_Models/links/586f67ad08ae329d6215fc4c/Using-R-to-Interpret-Interaction-Effects-in-Statistical-Models.pdf>

SCHMITT, Norbert & Ronald Carter. (2004) "Formulaic sequences in action: An introduction". In: Schmitt, Norbert (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, pp. 1-22.

SHLESINGER, Miriam. (1989) *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum*. MA thesis, Tel Aviv University.

SHLESINGER, Miriam & Noam Ordan. (2012) "More Spoken or More Translated?: Exploring a Known Unknown of Simultaneous Interpreting". *Target* 24:1, pp. 43-60.

SIYANOVA-CHANTURIA, Anna & Omidian, Taha. (2019) "Key issues in researching multi-word items". In: Webb, Stewart (ed.), *The Handbook of Vocabulary Studies*. London: Routledge, pp. 511-524.

SZERSZUNOWICZ, Joanna. (2020) "New Pragmatic Idioms in Polish: An Integrated Approach in Pragmateme Research". In: Piirainen, Elisabeth, Natalia Filatkina, Sören Stumpf & Christian Pfeiffer (eds.), *Formulaic Language and New Data Theoretical and Methodological Implications*. Berlin: De Gruyter, pp. 173-196.

SZYMOR, Nina. (2018) "Translation: universals or cognition?". *Target* 30:1, pp. 53-86.

TEAM, R.C. (2013) "R: A language and environment for statistical computing". Online version: <https://www.r-project.org/>

ULRYCH, Margherita & Amanda Murphy. (2008) "Descriptive Translation Studies and the Use of Corpora: Investigating Mediation Universals". In:

Torsello, Carol Taylor, Katherine Ackerley & Erik Castello (eds.), *Corpora for University Language Teachers*. Frankfurt am Main: Peter Lang, pp. 141-166.
WINTER, Bodo. (2019) *Statistics for Linguists: An Introduction Using R*. London: Routledge.
WOOD, David. (2015) *Fundamentals of Formulaic Language*. London: Bloomsbury.
WRAY, Alison. (2002) *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
WRAY, Alison. (2008) *Formulaic language. Pushing the boundaries*. Oxford: Oxford University Press.

Appendix 1. List of frequent bigrams in spoken and written registers

Spoken register (353 bigrams)

of the, in the, it is, the european, to the, and the, on the, for the, we are, we have, i would, mr president, like to, that we, is a, that the, european union, there is, would like, this is, is the, is not, we should, in this, the commission, member states, at the, should be, by the, madam president, with the, will be, has been, the eu, there are, to make, to be, i think, all the, and a, as a, the future, to ensure, the same, from the, we need, of this, i have, i am, president i, this house, the union, that is, and we, and i, need to, must be, the world, not only, does not, but also, that it, do not, and to, and in, european parliament, is important, ensure that, the report, those who, in europe, have been, about the, which is, for a, of a, president the, that this, they are, want to, we do, if we, the situation, that there, it should, the case, what we, we must, it will, is that, have to, we can, in our, a very, in a, in particular, the crisis, and that, who are, which i, what is, i hope, as the, to do, talking about, of european, which will, we cannot, fact that, also like, to thank, the fact, that are, is still, in which, a policy, we will, that in, on this, not the, to ask, let us, have a, but we, but i, the country, between the, would also, think that, the moment, because it, will not, which we, the most, level of, into the, part of, it must, are not, and not, with a, to say, of our, can be, and it, of eu, is to, is no, in my, as we, a new, european commission, the commissioner, this agreement, important to, believe that, of national, for example, are talking, the policy, the people, the common, should not, policy and, which are, to follow, the other, policy of, of course, who have, the time, the need, the euro, say that,

have the, has done, when we, to take, such as, role in, is also, and its, not be, make a, it has, be the, a year, a good, is an, the opposition, important that, and political, the internal, into account, according to, this matter, the council, the second, the recent, the global, percent of, means that, market and, will have, which has, union and, these are, said that, policy we, of people, make sure, future of, for their, cannot be, and there, access to, a certain, union is, to which, the role, the risk, the last, terms of, room for, needs to, jobs and, is about, in terms, in other, based on, and this, who has, we want, we know, we also, was not, to this, to lend, this in, the way, role of, make it, lack of, kind of, is very, a major, a clear, that i, such a, it was, is why, i want, i know, but it, as to, be a, implementation of, this parliament, the possibility, the independent, european budget, cooperation and, very important, to participate, the resolution, parliament has, national level, for innovation, commission has, the president, the elections, the countries, situation and, single market, government of, commission to, agreement and, thousands of, the southern, the republic, the question, the national, the external, research and, president we, member state, in countries, countries in, a resolution, a compromise, where there, welcome the, the subject, the present, the interim, of economic, involved in, included in, in addition, has already, continue to, within the, which were, the number, the budget, states and, social and, for europe, during the, context of, both sides, across the, years ago, under the, today the, to create, the visit, that they, thanks to, thank you, states in, source of, same time, rights in, report on, report is, reform of, policy is, people of, number of, my report, know that, in future, i believe, hope that, have said, for those, first and, debate on, crisis in, case that, are still, and their, after all, you have, years in, when the, were not, visit of, trade in, to avoid, the very, the next, order to

Written register (352 bigrams)

of the, in the, it is, the european, to the, and the, on the, for the, we are, we have, i would, mr president, like to, that we, is a, that the, european union, there is, would like, this is, is the, is not, we should, in this, the commission, member states, at the, should be, by the, madam president, with the, will be, has been, the eu, there are, to make, to be, i think, all the, and a, as

a, the future, to ensure, the same, from the, we need, of this, i have, i am, president i, this house, the union, that is, and we, and i, need to, must be, the world, not only, does not, but also, that it, do not, and to, and in, european parliament, the construction, is important, ensure that, the report, those who, in europe, have been, about the, which is, for a, of a, president the, that this, they are, want to, we do, if we, the situation, that there, it should, the case, what we, we must, it will, is that, have to, we can, in our, a very, in a, in particular, and that, who are, which i, what is, i hope, as the, to do, talking about, of european, which will, we cannot, fact that, also like, to thank, the fact, that are, is still, in which, a policy, we will, that in, on this, not the, to ask, let us, have a, but we, but i, the country, between the, would also, think that, the moment, because it, will not, which we, the most, level of, into the, part of, it must, are not, and not, with a, to say, of our, can be, and it, of eu, is to, is no, in my, as we, a new, european commission, the commissioner, this agreement, important to, believe that, of national, for example, are talking, the policy, the people, the common, should not, policy and, which are, to follow, the other, policy of, of course, who have, the time, the need, the euro, say that, have the, has done, when we, to take, such as, role in, is also, and its, not be, make a, it has, be the, a year, a good, is an, the opposition, important that, and political, into account, according to, this matter, the council, the second, the recent, the global, percent of, means that, market and, will have, which has, union and, these are, said that, policy we, of people, make sure, future of, for their, cannot be, and there, access to, a certain, union is, to which, the role, the risk, the last, terms of, room for, of human, needs to, is about, in terms, in other, based on, and this, who has, we want, we know, we also, was not, to this, to lend, this in, the way, role of, make it, lack of, kind of, is very, a major, a clear, that i, such a, it was, is why, i want, i know, but it, as to, be a, implementation of, this parliament, the possibility, the independent, cooperation and, very important, to participate, the resolution, parliament has, national level, for innovation, commission has, the president, the elections, the countries, situation and, single market, government of, commission to, christians in, agreement and, thousands of, the republic, the question, the national, the external, research and, president we, member state, in countries, countries in, a resolution, a compromise, where there, welcome the, the subject, the present, the interim, republic of,

of economic, involved in, included in, in addition, has already, continue to, within the, which were, the single, the number, states and, social and, for europe, during the, context of, both sides, across the, years ago, under the, today the, to create, the visit, that they, thanks to, thank you, states in, source of, same time, rights in, report on, report is, reform of, policy is, people of, number of, my report, know that, in future, i believe, hope that, have said, for those, first and, debate on, crisis in, case that, are still, and their, after all, you have, years in, when the, were not, visit of, trade in, to avoid, the very, the next, order to

Appendix 2

Details of the model reported in section 3.1 (number of bigram types as a function of predictor variables in spoken register). Marginal and conditional R2 has been calculated in R with the MUMIN package (Barton 2019).

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: CommonBigramTypesNumber ~ TextVariety + Delivery + STWPM + offset(TotalBigramsInText) +      (1 | TextID)
   Data: df
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

     AIC      BIC   logLik deviance df.resid
    1068     1085     -528     1056      119

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.12001 -0.06356  0.06704  0.15995  0.46832

Random effects:
 Groups Name        Variance Std.Dev.
 TextID (Intercept) 0.4378   0.6616
Number of obs: 125, groups:  TextID, 125

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                   2.9014     0.2098  13.830  < 2e-16 ***
TextVarietySpokenNonnative    0.2900     0.2006   1.446  0.14823
TextVarietySpokenTranslated   0.7896     0.2450   3.222  0.00127 **
Deliveryread                 -0.2530     0.1421  -1.781  0.07497 .
STWPM                         0.1579     0.1009   1.565  0.11763
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) TxtVSN TxtVST Dlvryr
TxtVrtySpkN -0.678
TxtVrtySpkT -0.815  0.714
Deliveryred -0.612  0.075  0.193
STWPM       -0.626  0.542  0.784  0.130
> r.squaredGLMM(Bigramsspoken)
                R2m       R2c
delta     0.1491673 0.9325990
lognormal 0.1493523 0.9337552
trigamma  0.1489759 0.9314020
```

Appendix 3

Details of the model reported in section 3.2 (number of bigram types as a function of predictor variables in written register). Marginal and conditional R2 has been calculated in R with the MUMIN package (Barton 2019).

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: CommonBigramTypesNumber ~ TextVariety + Delivery + offset(TotalBigramsInText) +      (1 | TextID)
   Data: df
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

     AIC      BIC   logLik deviance df.resid
  1122.1   1136.3   -556.1   1112.1      120

Scaled residuals:
     Min       1Q   Median       3Q      Max
-0.94026 -0.04801  0.06200  0.15205  0.38361

Random effects:
 Groups Name        Variance Std.Dev.
 TextID (Intercept) 0.4917   0.7008
Number of obs: 125, groups:  TextID, 125

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  3.47663    0.17108  20.322  < 2e-16 ***
TextVarietyWrittenNonnative  0.08581    0.17705   0.485  0.62790
TextVarietyWrittenTranslated 0.32417    0.15905   2.038  0.04154 *
Deliveryread                -0.43241    0.14711  -2.939  0.00329 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) TxtVWN TxtVWT
TxtVrtyWrtN -0.518
TxtVrtyWrtT -0.673  0.554
Deliveryred  0.684  0.005  0.147
> r.squaredGLMM(Bigramswritten)
                R2m       R2c
delta     0.1154960 0.9455923
lognormal 0.1155951 0.9464038
trigamma  0.1153939 0.9447561
```

BIONOTE

MARTA KAJZER-WIETRZNY is an Assistant Professor in the Department of Translation Studies at the Faculty of English, Adam Mickiewicz University in Poznań. Following her PhD dissertation on *Interpreting universals and interpreting style* (2012) she continues with empirical investigations of interpreted, translated and non-native language use, e.g. within the recent TRINFO project carried out in part during an over year-long research stay at the University of Bologna. At times she attempts to combine corpus methods with translation process research such as key-logging and eye-tracking, in

particular while looking into the traits and the process of inter- and intralingual translation.

KURZBIOGRAFIE

MARTA KAJZER-WIETRZNY ist Assistenzprofessorin am Institut für Übersetzungswissenschaft der Fakultät für Englisch der Adam-Mickiewicz-Universität in Posen. Nach ihrer Dissertation über die Universalien und den Stil des Dolmetschens (2012) setzt sie ihre empirischen Forschungen des gedolmetchten, übersetzten und nicht-muttersprachlichen Sprachgebrauchs fort, z.B. im Rahmen des TRINFO-Projekts, das teilweise während eines über einjährigen Forschungsaufenthaltes an der Universität Bologna durchgeführt wurde. Zuweilen versucht sie, Korpusmethoden mit Übersetzungsprozessforschung wie Key-Logging und Eye-Tracking zu kombinieren, insbesondere indem sie sich mit den Merkmalen und dem Prozess der inter- und intralingualen Übersetzung befasst.

ŁUKASZ GRABOWSKI is an Associate Professor at the Institute of Linguistics, University of Opole, Poland. In 2013, he was a post-doctoral research fellow at the University of Birmingham (UK). His main research interests include corpus linguistics, formulaic language and translation studies. He has published internationally in such journals as *International Journal of Corpus Linguistics, International Journal of Lexicography, Across Languages and Cultures* and *English for Specific Purposes*; he has also authored a number of chapters in edited volumes published by John Benjamins, Springer and Emerald, among others.

KURZBIOGRAFIE

ŁUKASZ GRABOWSKI ist außerordentlicher Professor am Institut für Linguistik der Universität Opole, Polen. 2013 war er Postdoktorand an der University of Birmingham (UK). Seine Forschungsschwerpunkte umfasssen Korpuslinguistik, Formelsprache und Übersetzungswissenschaft. Er hat in internationalen Zeitschriften wie *International Journal of Corpus Linguistics, International Journal of Lexicography, Across Languages and Cultures* und

English for Specific Purposes veröffentlicht. Er hat auch eine Reihe von Kapiteln in Sammelbänden verfasst, die unter anderem von John Benjamins, Springer und Emerald veröffentlicht wurden.