Syntactic properties of constrained English: A corpus-driven approach

Ilmari Ivaska (University of Turku)
Adriano Ferraresi (University of Bologna)
Silvia Bernardini (University of Bologna)

Ivaska, Ilmari, Adriano Ferraresi & Silvia Bernardini. 2022. Syntactic properties of constrained English: A corpus-driven approach. In Sylviane Granger & Marie-Aude Lefer (eds.), Extending the Scope of Corpus-Based Translation Studies, 133–157. London: Bloomsbury. http://doi.org/10.5040/9781350143289.0013.

Abstract

This chapter explores the common ground shared by non-native (L2) and translated language (TrL), seen as instances of constrained language use. It has been suggested that these diverge from native non-translated language (L1) in consistent ways. We explore this hypothesis in a corpus-driven manner, comparing written English in its L2 and TrL varieties, setting them against the benchmark of the L1 variety. In an attempt to control for confounding variables, we include two first/source languages for the constrained varieties, as well as three registers (argumentative writing, political speeches, and tourism-related communication), which also allows us to increase representativeness. Methodologically, we look at frequencies of part-of-speech dependency bigrams, adopting keyness analysis and Multi-Dimensional Analysis to detect and interpret differences between the contrasted varieties. The strengths of the approach are that it relies on syntactically parsed data instead of shallow part-of-speech sequences, is fully data-driven and can be easily implemented in different languages. Results indicate a tendency for the constrained varieties to rely on post-nominal modification and common nouns with determiners to a greater extent than non-constrained varieties, and to display a peculiar use of syntactic structures including proper nouns. Registers are found to impact greatly on results, and cross-register differences to be less prominent in the constrained varieties, which might point to a less heightened sensitivity to register conventions when performing language tasks under constraint of another language. Given the vast amount of variation in the data, the contribution ends on a note of caution when generalising over interpretations of constrained language data.

Keywords

Constrained language; Non-native language; Translated language; Random Forests; Multi-Dimensional analysis; Dependency bigrams; Universal Dependencies.

1. Introduction

Second language acquisition (SLA) and translation studies (TS) can be conceptualised as related disciplines, since both study language use under circumstances where more languages than one are inherently present in an act of communication. Despite the conceptual proximity, SLA and TS have until recently operated largely in isolation from each other (however, see Gaspari and Bernardini 2010). In recent years, they have been brought together in a more programmatic fashion in a series of works (Lanstyák and Heltai 2012; Kolehmainen et al. 2014; Kruger and van Rooy 2016; Rabinovich et al. 2016; Kruger and van Rooy 2018) that have attempted to unearth the theoretical or empirical common ground between the two types of language use.

One of the aims of SLA and TS has been to explore whether, to what degree, how, and why, non-native language (L2) or translated language (TrL) diverge from non-translated first language (L1) use. The new line of research aims to explore the common ground shared by L2 and TrL, looking for similarities between their divergences from L1. While L2 and TrL differ in many respects, including the ontological difference of production vs. *re*production (Shlesinger and Ordan 2012), their potential common ground has been hypothesised to stem from similar cognitive and social constraints (Kruger and van Rooy 2016: 27; Kotze in Chapter 3 of this volume), hence the label of *constrained language use* (Lanstyák and Heltai 2012).[1] More precisely, these similarities could be traced back to the activation of a bilingual language mode (Grosjean 2001) and the consequent increase in cognitive load, which could lead to similar patterns in language production, sometimes referred to as "universals".

To investigate this hypothesis, in this contribution we compare written English in its L2 and TrL varieties, setting both against the benchmark of L1 written English. We include two first/source languages (L1/SL: German and Italian), and three registers (argumentative writing, political speeches, and tourism-related communication, such as tourist guides and brochures), in an attempt to limit the impact of a specific constraining language or register and to increase representativeness. Since corpus sourcing is a major challenge for this research design, the choice of registers and L1s/SLs is opportunistic: we include registers and languages for which corpora exist for one or more varieties, and fill gaps with purpose-built ones. Summing up, the corpus used in this study represents three varieties of English: one is native, and two are produced under different kinds of constraints, related to two foreign languages, in three registers, for a total of 16 subcorpora.

Our data are annotated automatically according to the Universal Dependencies scheme (Nivre et al. 2016). Methodologically, we follow the two-phase procedure introduced by Ivaska and Bernardini (2020): we use keyness analysis (Gabrielatos 2018) and Multi-Dimensional Analysis (Biber 1988) as corpus-driven techniques, to detect and interpret differences between the contrasted language varieties related to the part-of-speech (POS) dependency bigrams found in each subcorpus. POS dependency bigrams are pairs of POS linked by a syntactic dependency, whose components might or might not be adjacent. Relying on syntactic dependencies rather than positional adjacency makes it

[1] The same phenomenon has been discussed also under the term *mediated language* (Ulrych and Murphy 2008). Since the term "mediation" could also refer to a form of reproduction, we find *constrained language* to be more transparent.

possible to focus on grammatically defined word pairings irrespective of their structural variation, thus overcoming limitations related to the use of shallow POS structures (Tono 2000; Granger and Bestgen 2014). Furthermore, the approach is fully data-driven and can be implemented in different languages, making it easier to address potential crosslinguistic similarities.

Our specific research questions are: 1) Which POS dependency bigrams best distinguish both L2 and TrL from L1? 2) To what extent are the best distinguishing bigrams register- or language-pair-specific? 3) Can the profiles of constrained language use emerging from the analysis be explained in the light of purported "universals of constrained communication", such as discourse transfer, simplification, normalization, or explicitation (Lanstyák and Heltai 2012)?

The paper is structured as follows: Section 2 reviews previous work on constrained language use focusing on empirical, corpus-based approaches. The setup of the corpus and the methodology used in the study are described in Section 3, and results in Section 4. Section 5 attempts to interpret these results, relating them to those obtained in earlier research. Section 6 concludes by summarising the lessons learnt and sketching directions for further research.

2. The elusive quest for *universals* of constrained language use

Recent research on constrained language (especially Kruger and van Rooy 2016; Rabinovich et al. 2016; Kruger and van Rooy 2018; Ivaska and Bernardini 2020) suggests that activities traditionally seen as unrelated, such as translation and L2 use, in fact seem to display several common features. These include a more explicit and formal style, lower lexical richness, fewer idiomatic expressions, fewer pronouns and greater use of explicit cohesive devices than non-translated or L1 use. These features manifest themselves in interaction with register differences: as in earlier studies within SLA (Ivaska 2015) and TS (Szymor 2018), constrainedness is found to be related to register sensitivity (Kruger and van Rooy 2018), potentially stemming from the register-specificity of the linguistic systems of individual language users (Iwasaki 2015). Besides register, a second important variable in the study of constrained language is crosslinguistic influence (CLI). This has always been a core interest in both SLA (Jarvis 2000) and TS (Toury 2012: 310–15). According to Jarvis (2000; 2010), a comparison-based argument in support of CLI in L2 must show congruity among speakers of the same first language, differences between speakers of different first languages, as well as correlated linguistic phenomena in all the languages involved. Extending the argument to research on constrained language use, to make sure that an observed phenomenon is not due to CLI from any specific language, it must be observed across different L1/SLs. In sum, a research design seeking to find reliable evidence of constrainedness must control for both register and constraining language.

Kruger and van Rooy (2016) include six registers that are comparable across the studied varieties but treat each of the three subcorpora as a unit, effectively conflating the registers in the analyses. The L2 data come from the East Africa component of the International Corpus of English (ICE, Greenbaum 1996), the L1 data come from ICE Great Britain, and translations (from Afrikaans) are collected ad hoc following the ICE guidelines. Due to lack of control on register and constraining language effects, the scope and the generalizability of these otherwise ground-breaking results remain somewhat blurred – especially in light of the remarkable differences between registers observed in a similar (albeit bi-varietal) research design (Hu et al. 2016). Rabinovich et al. (2016) take the opposite approach, as their data only include one register: the European Parliament plenaries. The translated, L2 and L1 varieties are thus fully comparable, and include a large number of constraining languages – yet the generalizability of results is limited, because they represent a single register, which is also linguistically very peculiar (Swallow 2003).

To the best of our knowledge, the most comprehensive study on the topic is Kruger and van Rooy (2018), which elaborates on the previous study by controlling for register and including several constrained varieties, which are categorised according to the amount of language contact. Furthermore, the non-native subcorpora include various L1 backgrounds which, however, are defined on geographical rather than linguistic bases (e.g. Indian English and East African English). Finally, the translation subcorpus only contains one source language (Afrikaans), making it difficult to tease apart phenomena related to single constraining languages from those related to the nature of the constraint (L2 or TrL).

Since data collection and method are of primary importance, as well as particularly complex, in what follows we describe the steps we took to ensure that our results would be reliable *and* generalisable beyond a single register or constraining language.

3. Data and method

3.1. Data collection

Our dataset for the present study consists of 16 English subcorpora in three registers (argumentative writing, political speeches, and tourism-related communication), with two typologically distant languages constraining the TrL and L2 data (Italian and German). Since it was impossible to find published L2 argumentative data, unpublished texts were included for this register, alongside two unconstrained components – one of unpublished texts matching the L2 texts and one of published texts matching the translations, thus controlling for the effect of editing (Kruger 2017).

With sustainability in mind, we obtained data from preexisting sources where possible and filled gaps when necessary and feasible. Table 4.1 provides size and provenance data for the 16 subcorpora: eight are based on existing resources (the Corrected and Structured Europarl Corpus [CoSTEP, Graën et al. 2014], the International Corpus of Learner English [ICLE, Granger et al. 2009] and the Louvain Corpus of Native English Essays [LOCNESS]),[2] and eight were self-compiled (L1 and translated argumentative writing, and all tourism-related communication).[3] CoSTEP comprises verbatim reports of speeches given by members of the European Parliament; we included speeches delivered in English by British, German, and Italian speakers, and translations into English from German and Italian. ICLE includes argumentative essays written by university-level L2 learners of English from various countries; we included the production of L1 German and L1 Italian learners. LOCNESS was compiled to provide English L1 speaker data comparable with ICLE. The translated argumentative texts and the comparable L1 texts were collected from opinion or column sections of online news outlets, including news agencies (e.g. *Reuters*), newspapers (e.g. the *New York Times*) and magazines (e.g. *Internazionale*). The tourism texts come from tourist guides, descriptions of locations found on tourist websites, as well as tourist brochures. We only included texts for which we were able to verify their translated or L2 nature, and their L1/SL. Pre-existing corpora are available for research purposes either freely or via licensing, whereas the self-compiled data were collected from online sources and, due to copyright restrictions, cannot be redistributed. The sizes of the subcorpora vary substantially, both in terms of number of words and number/length of texts. In Section 3.2.2 we describe the procedure we used to match corpus sizes for each register and language combination, in each pairwise comparison.

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/locness.html
[3] Data and scripts are available as an Open Science Framework repository: https://osf.io/8yuz4/.

Table 4.1. Number of words per component, with data provenance.

REGISTER	L1	TrL		L2	
		DE	IT	DE	IT
Argumentative Writing (ARG) *L1: LOCNESS (unpublished) & self-compiled (published)* *TrL: self-compiled* *L2: ICLE*	168,368 *(unpublished)* 93,697 *(published)*	138,236	122,534	108,986	116,355
Political speeches (POL) *All: CoStEP*	1,954,403	2,622,790	2,025,921	18,796	21,454
Tourism (TOU) *All: self-compiled*	35,288	66,051	92,401	86,796	69,042

Language proficiency is not an issue for the translated components, since translators in the language pairs investigated here normally work into their first language or in their language of habitual use, but it does affect the L2 data. Ideally, language proficiency should be as high as possible across the different L2 components. In practice, the English of the CoStEP corpus of transcribed political speeches is likely to be native-like: L2 speakers at the European Parliament who elect to speak English do so although interpreting is available. The proficiency level in ICLE is generally considered from higher intermediate to advanced (Granger et al. 2009), whereas for L2 tourism texts we rely on the fact that the authors are proficient enough and willing to be published in a second language. In other words, all L2 data can be considered to represent relatively high proficiency in English.

Writers' expertise and editorial interventions have also been shown to affect language production and, hence, to influence data comparability (Kotze 2018). In our data, this concerns mainly argumentative writing and tourism-related communication, where we have adopted a two-way solution to mitigate the problem. In argumentative writing, we have two separate L1 datasets: one of non-professional writers' unpublished texts (to be compared with the corresponding L2 datasets), and another of professional writers’ published texts (to be compared with the corresponding TrL datasets). As for the tourism texts, we ensured that all the texts either came from professional publishers or the authors had a documented history of multiple published articles.

3.2. Method

3.2.1. Rationale

Given the exploratory nature of our study, we use a corpus-driven method to detect differences between constrained and unconstrained varieties of English in terms of relatively more/less frequent bigram structures. We chose POS dependency bigrams because of their versatility: they capture information on syntactic functions and POS but also on the constituent order and hierarchy. In this way, our results are comparable with those from earlier studies making use of POS annotation, yet are richer thanks to the syntactically-informed and crosslinguistically comparable insights they provide (cf. Ivaska and Bernardini 2020). In order to balance data comparability and generalizability of results, comparisons are conducted separately for each variety, register and L1/SL. In this way we try to avoid over-interpretations. We then compare the results of these pairwise comparisons to gauge the impact of the key features found in each, and try to tease apart the interactions between type of constraint, register and constraining language.

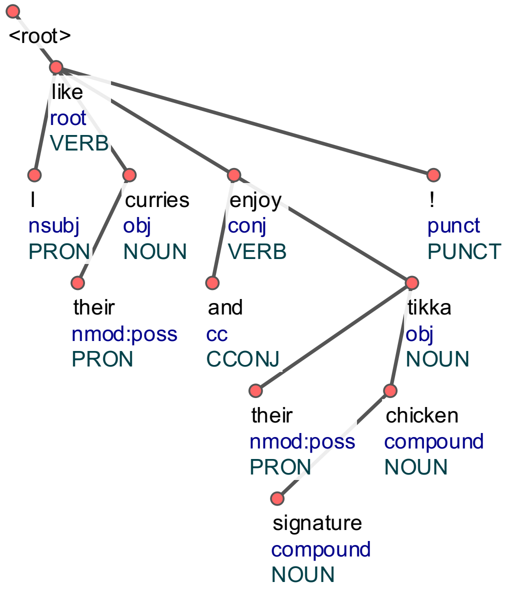
Conceptually, our method is in line with the general notion of keyness analysis (Gabrielatos 2018), in that its aim is to find consistent differences in frequencies of linguistic elements between the compared varieties, which may serve as pointers to potentially interesting generalizations (Scott 2010: 56–57). Our specific procedure is carried out in three steps: 1) data preparation (setting the data in a standard format, parsing them linguistically and extracting frequencies); 2) pairwise keyness analyses and identification of consistent key features; 3) Multi-Dimensional Analysis of consistent key features.

For the purpose of testing the generalizability of our results, each subcorpus is split into a training set and a test set (70%-30% of the data). In other words, all the analyses are done using the training set, and then repeated using the test set, so as to evaluate whether the same results are obtained, and could thus be generalized to other data, too. Key feature identification and Multi-Dimensional Analysis are performed using the training data. We then re-implement the resulting dimensions onto the test set to evaluate the extent to which the keyness-based dimensions also apply to previously unseen data.

3.2.2. Data preparation

After stripping legacy markup, texts were parsed with the UDPipe parser (Straka and Straková 2017), following the Universal Dependencies (UD) annotation scheme (Nivre et al. 2016). The scheme includes information separately for lemmas, POS, morphological features, and syntactic dependencies. In this study, we focus on POS dependency bigrams, exemplified in Figure 4.1. The figure shows a visualization of an annotated sentence (TOU: L2_de-en). Dependency bigrams are, for instance, PRONNODE_nsubj_VERBHEAD (*I–like*), VERBHEAD_obj_NOUNNODE (*like–curries*), PRONNODE_nmod:poss_NOUNHEAD (*their–curries*), and so on.

Figure 4.1. Tree visualization of the sentence *I like their curries and enjoy their signature chicken tikka!*



To maximise comparability and minimise the effects of other variables (e.g. topical and authorial variation), we shuffled the sentences of each subcorpus, reconstructing text blocks of 50 sentences each (for this step, we followed the example of Rabinovich et al. 2016 and Ivaska and Bernardini 2020). We extracted the frequencies of the 1,000 most frequent dependency bigrams and normalised their frequencies per 1,000 tokens for each text block, using ad hoc scripts written in Java.

3.2.3. Pairwise key feature detection

We conducted 12 pairwise keyness analyses, contrasting each constrained subcorpus – one for each variety (L2 and TrL), each constraining language (DE and IT) and each register (ARG, POL, TOU) – with the respective unconstrained subcorpus. Key features shared by all or most constrained subcorpora would support the constrained language hypothesis, whereas features found to be key in single keyness analyses might hint at effects which are specific to a variety, register, or constraining language. In this study we treat key features included in over half of the pairwise comparisons as consistent indicators of constrainedness effects.

All statistical analyses were performed in R (R Core Team 2018). We used Boruta feature selection (Kursa and Rudnicki 2010) to find the most consistent predictors of differences between the datasets. The Boruta algorithm is based on Random Forests (Breiman 2001) and its implementation in R makes use of the Ranger package (Wright and Ziegler 2017). The algorithm adds randomness to the data by creating shadow copies of all variables (here, dependency bigrams) and randomly shuffling their values (here, normalised frequencies). It then runs the Random Forest classifier on all variables and contrasts the actual variables' performance to the shadow variables, to find the ones that consistently outperform the random variables. The process, which is repeated multiple times to avoid over-fitting, delivers a set of dependency bigrams considered important in distinguishing the contrasted varieties. The dependency bigrams deemed important in more than half of the pairwise comparisons are then isolated for further analysis.

3.2.4. Multi-Dimensional Analysis and generalizability evaluation

We used Multi-Dimensional Analysis (MDA) to make sense of the grouping and interaction between the detected key features. MDA is a popular method used to interpret cross-register and cross-varietal quantitative linguistic variation functionally (e.g. Biber 1988; Berber Sardinha and Veirano Pinto 2014). Contrary to the body of work using MDA, however, our variables were not motivated by earlier research but rather defined bottom up in the preceding keyness analysis (Section 3.2.3); otherwise, we followed as closely as possible the description by Egbert and Staples (2019). The factor analysis made use of the functions found in the R package Psych (Revelle 2018).

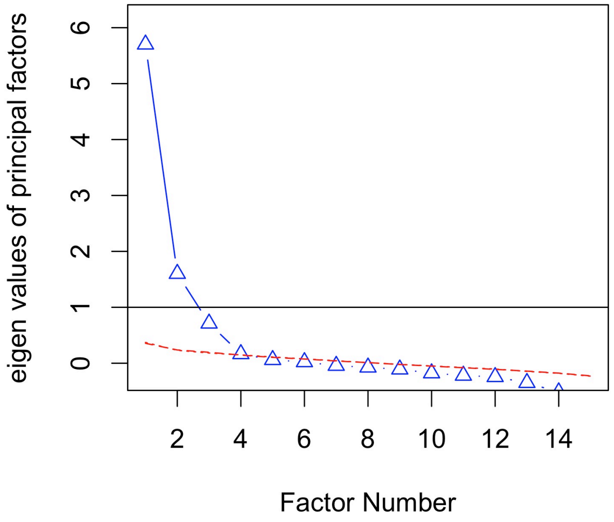
For the evaluation of the generalizability of our results, we took a maximal balanced random sample of test data, calculated the dimension scores, and explored their patterning with regard to the dimensions identified using the training data. The rationale is that patterning similar to the training data would corroborate the generalizability of the obtained results, whereas diverging patterns would indicate that the results are sensitive to uncontrolled underlying variability in the data. Section 4 offers an account of our results, which are then discussed in Section 5.

4. Results

4.1. Detecting candidate key features and choosing the number of dimensions

We identified 15 (out of 1,000) dependency bigram patterns whose frequencies of occurrence indicates systematic differences between constrained and unconstrained data in over half of the 12 pairwise keyness analyses. Our balanced, maximum-sized training set consists of 560 randomly selected text blocks, divided evenly across subsets, with sufficient factorability according to the Kaiser-Meyer-Olkin test, i.e. 0.87 (values above 0.5 generally indicate sufficiently factorable data (Kaiser 1974)). The scree plot of the eigen values (standardised measures of the proportion of variance explained by a given factor) suggests a two- or a three-factor solution (Figure 4.2). In other words, variables whose frequencies seem inter-related could be grouped into either two or three sets, called factors. We opted for a three-factor solution, as there is a clear drop in eigen values between factors 3 and 4, and as factor 4 overlaps with the simulated factors based on randomly permutated data (indicated in dashes in Figure 4.2). This means that the data do not support including a fourth factor – a fourth distinct grouping of inter-related variables. Our interpretation of the three factors in terms of dimensions is described in Sections 4.2-4.4 below.

Figure 4.2. Screeplot of eigen values for dimensions.



4.2. Dimension 1: clausal vs. phrasal elaboration

The first dimension reflects differences between clausal elaboration and phrasal elaboration. In other words, some text blocks display relatively more complex clause structures, while others display relatively more complex phrase structures.[4] Specifically, the dependency bigrams that load positively onto the first dimension (see Table 4.2) reflect relatively frequent use of finite clauses – both main and subordinate. The negatively loading patterns reflect the opposite preference: a relatively frequent use of various forms of phrasal modification. This means that the data can be split into text blocks that make relatively more use of finite clauses and relatively less use of complex nominal phrases – and others where the opposite is the case. In more general terms, the distinction seems to reflect the one proposed by Steiner (2012: 78) between “verbality“ and “nominality”, as both categorizations are based on frequency of verbal vs. nominal word classes as distinctive features of texts.

[4] Since sentences are shuffled, the correlations reflect patterns across datasets, not across individual texts.

Table 4.2. POS dependency bigrams loading onto Dimension 1.

Positive features		**Negative features**	
PROPNNODE_nsubj_VERBHEAD (proper noun acting as subject to verb)	0.856	NOUNNODE_compound_NOUNHEAD (compound noun)	-0.695
VERBHEAD_ccomp_VERBNODE (dependent clause acting as clausal complement)	0.901	ADPNODE_case_PROPNHEAD (preposition acting as case marker to proper noun)	(-0.399)[5]
PARTNODE_mark_VERBHEAD (particle acting as marker for subordinate or infinitival clause)	0.772	PROPNNODE_compound_PROPNHEAD (compound proper noun)	-0.508
VERBHEAD_obj_PRONNODE (pronoun acting as object to verb)	0.445	NOUNHEAD_nmod_PROPNNODE (nominal modifier of proper noun)	(-0.415)
PARTNODE_advmod_VERBHEAD (particle acting as adverbial modifier to verb)	0.636	NUMNODE_nummod_NOUNHEAD (numeral modifer of noun)	-0.634

The dependency bigrams loading positively onto this dimension include proper nouns acting as subjects to verbs (*DEA–legalize*, underlined in example (1a)), finite and non-finite clausal complements (*stated–legalize* in example (1b)), particles acting as adverbial modifiers of verbs (*not–legalize*, in example (1c)), particles acting as markers for infinitival or subordinate clauses (*to–trigger* in example (2)), and pronominal objects of verbs (*face–it* in example (3)).

(1a) Even with all of this being stated the DEA will still not legalize marijuana for medical purposes. [L1 ARG]
(1b) Even with all of this being stated the DEA will still not legalize marijuana for medical purposes.
(1c) Even with all of this being stated the DEA will still not legalize marijuana for medical purposes.
(2) It makes sense to trigger alarm bells rather than be complacent. [L1 ARG]
(3) How do they face it? [L1 ARG]

The negatively loading bigram structures correspond to noun and proper noun compounds (*death–penalty* in example (4)), prepositional modification of proper nouns (*city–Cleveland* in example (5)), and numeral modification (*two–things* in example (6)).

(4) Governments are trying to pass many bills to either opposing the death penalty, or supporting it. [L1 ARG]
(5) Arthur B. Modell sold out the city of Cleveland because he is greedy, pure and simple. [L1 ARG]
(6) Whenever asked why shouldn’t women be admitted there are two things that always come out first. [L1 ARG]

[5] Bigrams in brackets have higher loadings in another dimension: their scores are only included in the calculations for that dimension.

Figure 4.3. Mean scores and standard deviations for Dimension 1.

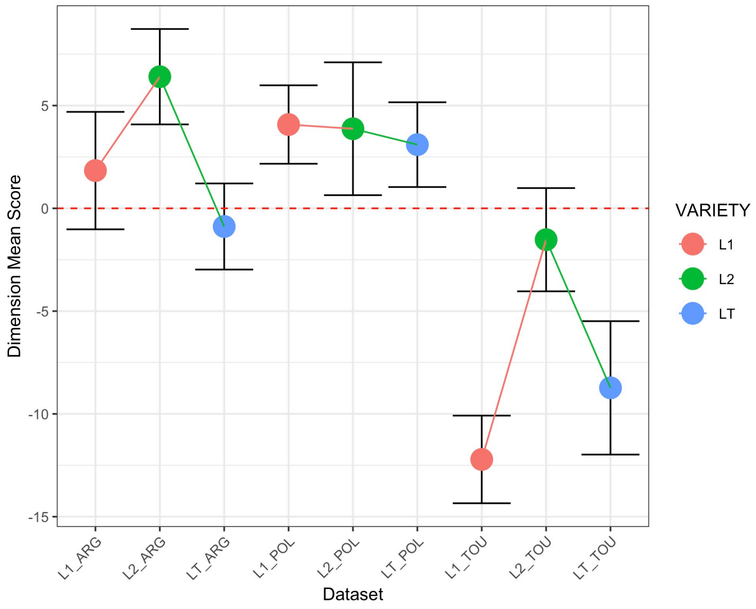


Figure 4.3 shows the mean scores and standard deviations for register groups (from left to right: ARG, POL, and TOU) and variety. Focusing specifically on the register comparison, the dimension scores reveal that tourism texts diverge from the rest due to negative dimension scores across all varieties. When the register-related differences are contrasted with differences related to constrainedness, two distinctive patterns emerge: first, in two of the three registers (i.e., to the exclusion of political speeches), L2 texts have higher scores than L1 texts, suggesting that non-native language users generally make more use of clausal elaboration/verbality than phrasal elaboration/nominality. In L2 and TrL, the mean scores portray consistently smaller differences between the registers than in L1, pointing to reduced register sensitivity.

4.3. Dimension 2: post-nominal modification and use of determiners

The second dimension reflects two distinct syntactic differences, concerning post-nominal modification and determiners. Positive dimension scores reflect higher frequency of these features (see Table 4.3).

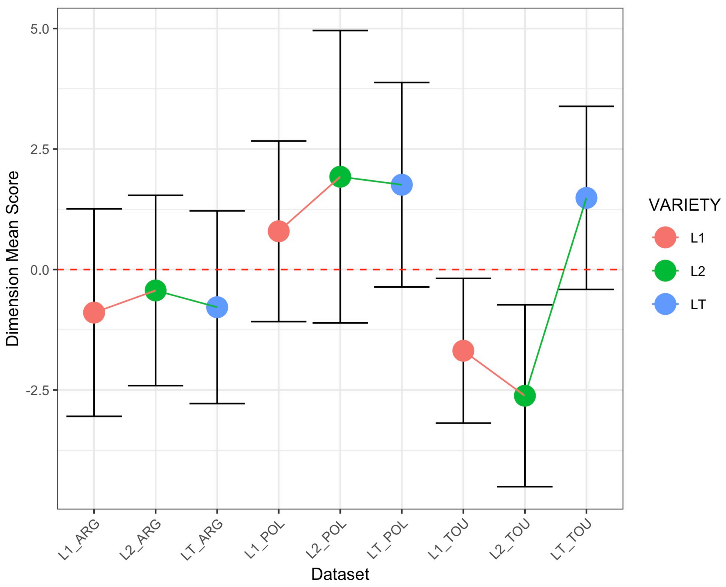
Table 4.3. POS dependency bigrams loading onto Dimension 2.

Positive features		Negative features
NOUNHEAD_nmod_NOUNNODE (post-nominal modifier of a noun)	0.905	–
DETNODE_det_NOUNHEAD (determiner of a noun)	0.648	–
ADPNODE_case_NOUNHEAD (prepositional noun phrase)	0.883	–

The bigrams that reflect post-modification include post-nominal modifiers of nouns (*authority–pictures* in example (7a), *marketplace–fun* in example (8a)) as well as the related prepositional phrases (*of–pictures* in example (7b), *for–fun* in example (8b)). The use of determiners applies both to definite articles (*the–masses* and *the–authority* in example (7c)) and indefinite articles (*a–marketplace*, in example (8c)).

(7a) The masses, however, still do not doubt the authority of pictures. [L2-L1de ARG]
(7b) The masses, however, still do not doubt the authority of pictures.
(7c) The masses, however, still do not doubt the authority of pictures.
(8a) Beer gardens are a marketplace for fun, stories and laughter. [TrL-L1de ARG]
(8b) Beer gardens are a marketplace for fun, stories and laughter.
(8c) Beer gardens are a marketplace for fun, stories and laughter.

Figure 4.4: Mean scores and standard deviations for Dimension 2.



The cross-register patterning of Dimension 2 (see Figure 4.4) reveals that political texts have positive dimension scores and tourism texts have negative scores, while argumentative texts occupy the middle ground. In other words, post-modification and the use of determiners are most common in political discourse and least common in tourism-related discourse. However, this picture gets more blurred when the register tendencies are contrasted with the constrainedness-related patterns. TrL tourism texts diverge drastically from both their L1 and L2 counterparts, displaying positive dimension scores, while the positive dimension scores observed in political discourse are even higher in the constrained varieties than in the unconstrained one.

4.4. Dimension 3: use of proper nouns

The third dimension reflects an array of different uses of proper nouns (see Table 4.4). All the dependency bigrams included in the third dimension have positive factor loadings, meaning that relatively high frequencies of occurrence correspond to positive dimension scores.

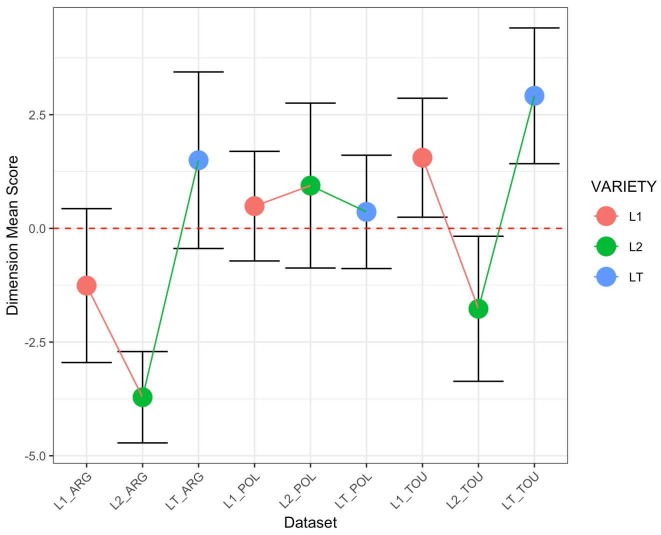
Table 4.4. POS dependency bigrams loading onto Dimension 3.

Positive features		**Negative features**
ADPNODE_case_PROPNHEAD (preposition acting as case marker to proper noun)	0.606	–
PROPNNODE_nsubj_VERBHEAD (proper noun acting as subject to verb)	0.434	–
PROPNNODE_compound_PROPNHEAD (compound proper noun)	(0.471)	–
DETNODE_det_PROPNHEAD (determiner of a proper noun)	0.868	–
NOUNHEAD_nmod_PROPNNODE (post-nominal modifier of proper noun)	0.521	–

The POS dependency bigrams included in Dimension 3 reflect prepositional phrases (*in–UK* in example (9a), *of–regional* in example (11a) and *of–Karl* in example (12a)), proper noun subjects of verbs (*Bernini–used* in example (10)), proper noun compounds (*Regional–Nature*, *Nature–Park* and *Mount–Conero* in example (11b)), determiners of proper nouns (*the–UK* in example (9b) and *the–Regional* in example (11c)) as well as proper nouns acting as post-modifiers (*residence–Karl* in example (12b)). Note that all the syntactic dependencies in Dimension 3 are found also in Dimensions 1 and 2, and the characteristic feature of Dimension 3 thus is the use of proper nouns.

(9a) It is not the only source of law in the UK. [L1 ARG]
(9b) It is not the only source of law in the UK. [L1 ARG]
(10) Here, Bernini used travertine instead of his usual marble. [TrL-L1it TOU]
(11a) The slogan of the Regional Nature Park of Mount Conero sums up the key to 6,011 hectares [TrL-L1it TOU]
(11b) The slogan of the Regional Nature Park of Mount Conero sums up the key to 6,011 hectares [TrL-L1it TOU]
(11c) The slogan of the Regional Nature Park of Mount Conero sums up the key to 6,011 hectares [TrL-L1it TOU]
(12a) much later this same house was the residence of Karl Bohm. [TrL-L1it TOU]
(12b) much later this same house was the residence of Karl Bohm. [TrL-L1it TOU]

Figure 4.5. Mean scores and standard deviations for Dimension 3.



The register-related patterning in Figure 4.5 indicates that, among the unconstrained varieties, proper nouns are less common in argumentative writing than they are in political or tourism-related discourse. Furthermore, the two constrained varieties diverge drastically both from the unconstrained variety and from each other: in both argumentative writing and tourism-related discourse, proper nouns are less common in L2 than in L1, whereas the opposite is the case when comparing TrL to L1. In political discourse, there are no notable differences between unconstrained and constrained varieties.

4.5. Generalizability evaluation

Before moving on to discuss our results, we evaluate their generalizability by mapping the obtained dimensions on separate test data which were not used either in the keyness analysis or in the Factor Analysis. As Figure 4.6-Figure 4.8 show, all three dimensions pattern very similarly in both datasets: the test data reflect all the cross-register and cross-varietal divergences discussed above. This result suggests that the detected key features reliably represent the data of the present study and are likely to be generalizable to similar data.

Figure 4.6. Dimension 1: results from train and test data.

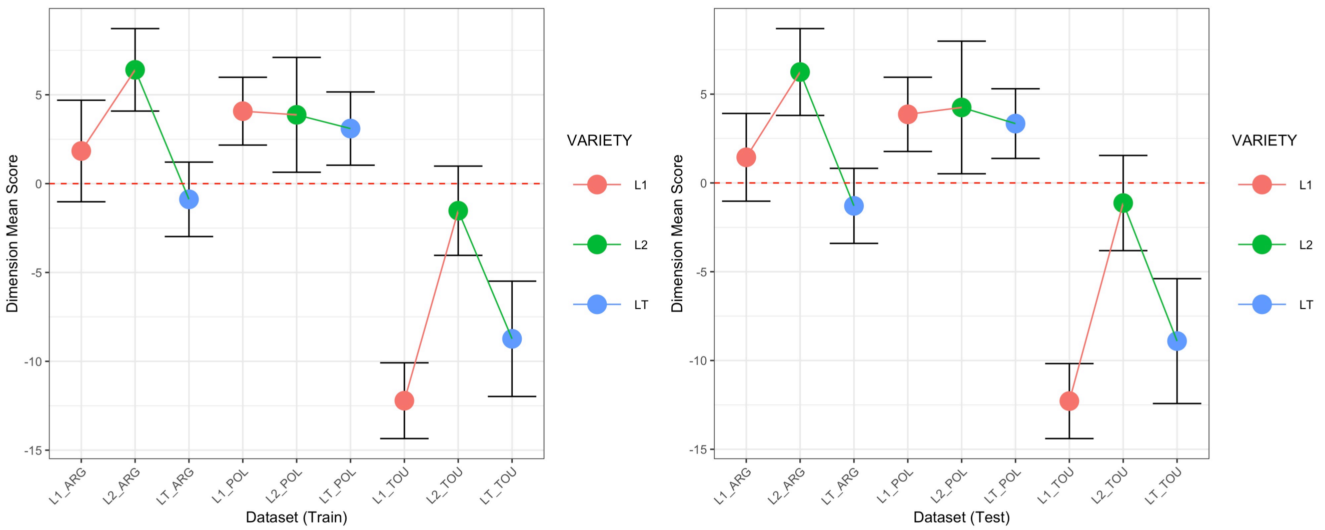


Figure 4.7. Dimension 2: results from train and test data.

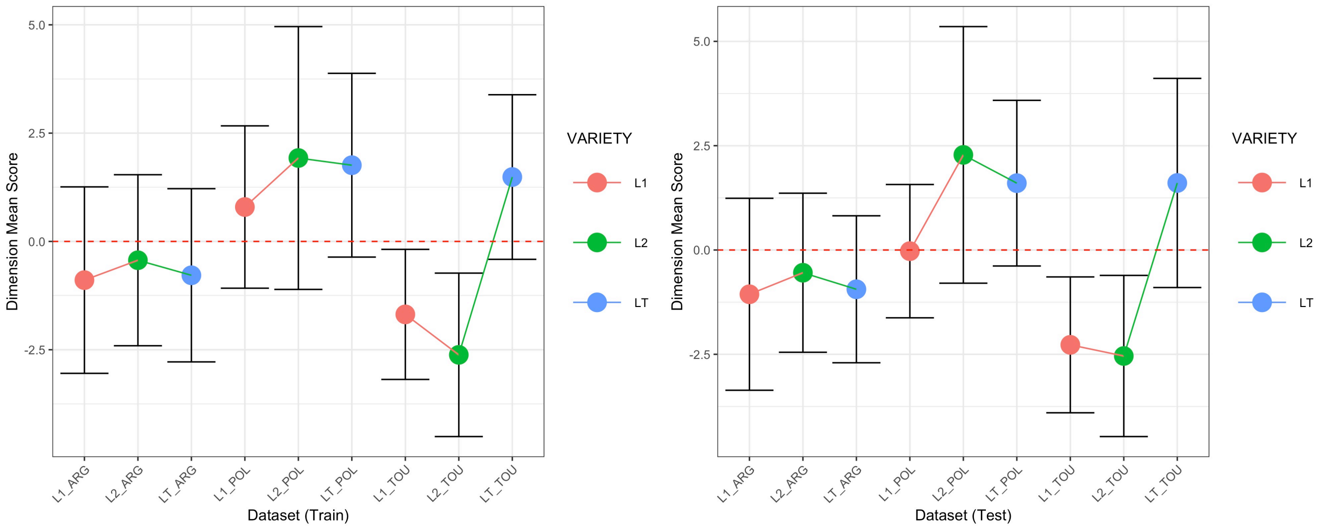
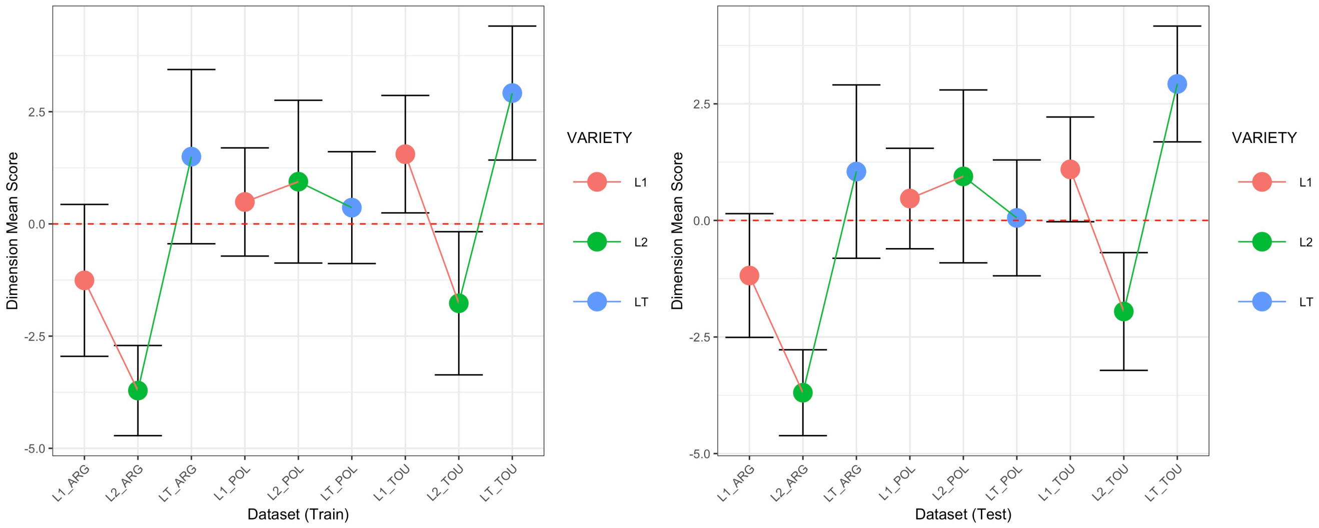


Figure 4.8. Dimension 3: results from train and test data.



5. Discussion

We can now go back to our research questions. The first one concerns the characteristics of constrained language use as evidenced by its distinctive POS dependency bigrams. We detected 15 such bigrams that consistently distinguish both L2 and TrL varieties from their unconstrained counterpart across the studied registers and irrespective of the L1/SLs involved. We grouped these into three underlying dimensions and interpreted them functionally as four distinct, yet closely related, linguistic phenomena. First, clausal complexity and verb-driven style were observed to be typical of the argumentative register, and phrasal complexity and noun-driven style of the tourism register within the L1 variety; differences were much less evident in the constrained (TL and L2) varieties. Second, in the constrained varieties, a preference for post-nominal noun phrase modification by means of prepositional phrases goes hand in hand with a third tendency, namely a more frequent use of nouns and determiners, both definite and indefinite. A fourth, somewhat unexpected result concerns the use of proper nouns in the constrained varieties, which differs from the unconstrained one in opposite ways: proper nouns are used more frequently in TrL than in L1, but less frequently in L2 than in L1.

The first lesson one can learn from our results is a general one: discussing the tendencies of constrained language use without taking into account language-pair and register specificities would unjustifiably simplify a complex phenomenon. This issue is addressed by our second research question, which concerns the impact of registers and language pairs. In this instance we would like to highlight, first and foremost, the amount of variation across language-pairs and registers. Out of the 1,000 most frequent dependency bigrams, 419 were identified as key features in distinguishing a constrained dataset from its unconstrained counterpart in at least one of the 12 pairwise comparisons, but only the 15 dependency bigrams we discuss in this chapter surfaced in over half of the keyness analyses. While differences due to these variables were not specifically focused upon, some were impossible to disregard. Register sensitivity was especially noticeable: in particular, differences between clausal and phrasal elaboration were found not to distinguish constrained from unconstrained language use linearly. Rather, argumentative writing and political discourse favour clausal complexity, while tourism-related discourse favours phrasal complexity. This distinction reflects a functional difference between the registers: argumentative and political texts are characterised by greater use of clausal structures and an involved style, while tourism texts prefer nominal structures and a more informational style. This distinction reflects the first dimension of most MDA studies – the difference between involved and informational production (Biber 1988). This dimension has also been shown to play a central role in distinguishing constrained from unconstrained English, and suggested to be related to heightened cognitive effort in real-time production conditions (Kruger and van Rooy 2016; 2018). Since none of our datasets are characterised by real-time constraints, however, we would suggest that other explanations might also be warranted. The same co-occurring linguistic features might indeed point to a preference for more formal (vs. informal), more written (vs. spoken) choices, in turn related to stylistic/personal preferences rather than functional or cognitive constraints. Register specificities also interact with constrainedness effects in another interesting way: as found in previous studies, cross-register differences are less prominent in the constrained varieties, pointing to a less heightened sensitivity to register conventions when translating into, or writing in, one's L2 (Gilquin and Paquot 2008, Szymor 2018, Ivaska and Bernardini 2020).

As for the second and third phenomena – post-nominal modification and use of common nouns with determiners – they point to a general reliance on noun phrases and on explicit strategies for their modification. The preference for prepositional postmodifiers as a more explicit (and risk reducing) alternative to nominal premodifiers (Quirk et al. 1985: 1330) has been observed in English fiction translation (Bernardini 2011). While in part supporting the earlier results, our study also shows the register-sensitive nature of these phenomena (see also House 2008: 12). The use of common nouns with determiners is more difficult to interpret, as it might hint at greater reliance on nouns as grammatically more complex/metaphorical alternatives to verbs (as our examples 7 and 8 would suggest), or else be due to transfer from the SLs/L1s. Though typologically distant, both Italian and German are generally described as more "nominal" than English (Cragie 2000; Serbina et al. 2017; Heilmann et al. this volume). In this case, and despite our efforts, source/first language influence cannot therefore be ruled out.

Finally, our fourth linguistic phenomenon, the use of various structures including proper nouns, is intriguing inasmuch as it seems to distinguish both L2 and TrL from L1 – but in an opposite fashion. While proper nouns may indeed be sensitive to constrainedness effects, the different constraints behave differently. TL texts come across as richer in precise references to real world entities than same-register L1 texts, while L2 texts, both argumentative and dealing with tourism, come across as less clearly anchored in the external world.

This leads us to our third research question regarding the purported universality and underlying explanation for constrained language use. In our view, the observed characteristics could be interpreted as an interaction between relative complexity (related to language users) and absolute complexity (related to linguistic components involved) (for these notions of complexity, see Miestamo et al. 2008). In other words, different registers are characterised by different types of linguistic complexity, and constrained and unconstrained language users react to these differences in a different manner. The interaction between the different types of complexity, which has been suggested to play a role in L2 acquisition (Bulté and Housen 2012), might also characterise other types of constrained language use. This can, in turn, be seen to stem from the fundamentally usage-based and input-sensitive origin of register variation (Iwasaki 2015), which has been suggested to underlie differences between both L1 and L2 (Ivaska 2015), and L1 and TrL (Szymor 2018). To us, such an interpretation is fully in line with earlier results regarding the centrality of register variation in linguistic systems across languages in general (Biber 2014), and in constrained language use in particular (Kruger and van Rooy 2018).

As for the reliability and generalizability of our results, the two-phase procedure implemented allowed us to strike a balance between the close comparability of the different subsets in the keyness analysis, and the evaluation of the generalizability of the detected key features across different registers and L1/SLs. The use of additional test data in a further sanity and consistency check of the obtained results confirms that our results are reliable. We would suggest that this methodological step, often absent from corpus linguistics studies, is valuable in general, and crucial for research designs that heavily rely on the comparability of the datasets used, as is the case with studies of constrained language, and variationist corpus linguistics at large.

Finally, it is noteworthy that our political discourse data are taken from the EuroParl corpus (Graën, Batinic, and Volk 2014), which has been used as a primary data source in many seminal works on tendencies of translated language in general (e.g. Koppel and Ordan 2011; Volansky, et al 2015), as well as for the purposes of exploring constrained language use in particular (Nisioi et al. 2016; Rabinovich et al. 2016). As our results show, in many cases those data pattern differently from the two other registers. Despite the undeniable value of this dataset, therefore, one should be extremely cautious before generalising interpretations stemming from it to any other communicative setting.

6. Concluding remarks

The definitions and operationalizations of potential universal tendencies of constrained language use are rarely comparable across different studies and languages. At the same time, distinguishing crosslinguistic influences from more general tendencies always requires operating at partially different levels of abstraction – and potentially similar phenomena will unavoidably have different types of realizations across different languages. Using a cross-linguistically comparable framework such as Universal Dependencies facilitates the narrowing down of this gap. In turn, looking at the general tendencies found in terms of the interaction between user-related and systemic complexity allows for the balancing of these different levels of abstraction.

From a corpus linguistics perspective, more research is needed in more language-pairs and including more registers. Furthermore, numerous other linguistic features/structures should be targeted, beyond the dependency bigrams that we report on here, ideally combining corpus-driven and corpus-based perspectives. The tendency for phrasal vs. clausal elaboration, for instance, could be further investigated focusing on nouns and verbs, and structures around them. Despite these limitations, we believe that the methodological procedure we laid out is valuable since it allows one to mediate between general and language-specific interpretations of constrained language use, while relying on solid corpus-driven methods.

Beyond corpus linguistics, the obtained results on the interaction between different types of complexity should be explored further using different methods – shifting the focus from the product to the process by means of, e.g. eye-tracking, key-logging, informant judgements, interviews and focus groups. This kind of methodological triangulation could provide evidence that none of these approaches will ever be able to provide in isolation.

Key readings

Lanstyák, I. and P. Heltai (2012), ‘Universals in Language Contact and Translation’, *Across Languages and Cultures* 13 (1): 99–121.

This article introduces the concept of *constrained communication* and surveys parallels between the proposed translation universals and similar phenomena pertaining to bilingual communication, by means of a meta-analysis of relevant literature in translation studies and contact linguistics. Hence, it serves as a cornerstone of the programmatic study of constrained language use.

Kolehmainen, L., L. Meriläinen and H. Riionheimo (2014), ‘Interlingual Reduction: Evidence from Language Contacts, Translation and Second Language Acquisition’, in H. Paulasto, L. Meriläinen, H. Riionheimo and M. Kok (eds), *Language Contacts at the Crossroads of Disciplines*, 3–32, Cambridge: Cambridge Scholars Publishing.

This article provides a very thorough theoretical meta-analysis of frequency-related language contact effects as instances of constrained language use, bringing together results from the fields of second language acquisition, translation studies, and contact linguistics. Besides its thoroughness, of particular value is the fact that this article looks at results from multiple languages.

Rabinovich, E., S. Nisioi, N. Ordan and S. Wintner (2016), ‘On the Similarities between Native, Non-Native and Translated Texts’, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1870–81. Berlin: Association for Computational Linguistics.

This article explores constrained language use in the speeches delivered in English at the European Parliament, making use of advanced methods of statistics and computational linguistics. While potentially limited in their generalizability to other data, the results indicate clear and also linguistically interpretable tendencies of constrained language use in general, while also pointing out the effect of the language-pairs involved.

Kruger, H. and B. van Rooy (2018), ‘Register Variation in Written Contact Varieties of English’, *English World-Wide* 39 (2): 214–42.

This article is probably the first to address the role of register in constrained language use. It compares native, non-native and translated English by means of a Multi-Dimensional Analysis, and comes to the conclusion that, in terms of register variation, constrained and unconstrained varieties pattern very similarly. Furthermore, the study shows that both register and variety contribute significantly to the observed variance.

Bibliography

Berber Sardinha, T. and M. Veirano Pinto, eds (2014), *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins.

Bernardini, S. (2011), ‘Monolingual Comparable Corpora and Parallel Corpora in the Search for Features of Translated Language’, *SYNAPS* 26: 2–13.

Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (2014), ‘Using Multi-Dimensional Analysis to Explore Cross-Linguistic Universals of Register Variation’, *Languages in Contrast* 14 (1): 7–34.

Breiman, L. (2001), ‘Random Forests’. *Machine Learning* 45 (1): 5–32.

Bulté, B. and A. Housen (2012), ‘Defining and Operationalising L2 Complexity’, In A. Housen, F. Kuiken and I. Vedder (eds), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, 21–46, Amsterdam: John Benjamins.

Cragie, S. (2000), *Thinking Italian Translation,* London: Routledge.

Egbert, J. and S. Staples (2019), ‘Doing Multi-Dimensional Analysis in SPSS, SAS, and R’, in T. Berber Sardinha and M. Veirano Pinto (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 99–114, London: Bloomsbury Academic.

Gabrielatos, C. (2018), ‘Keyness Analysis: Nature, Metrics and Techniques’, in C. Taylor and A. Marchi (eds), *Corpus Approaches to Discourse: A Critical Review*, 225–58, Oxford: Routledge.

Gaspari, F. and S. Bernardini (2010), ‘Comparing Non-Native and Translated Language: Monolingual Comparable Corpora with a Twist’. In R. Xiao (ed.), *Using Corpora in Contrastive and Translation Studies*, 215–34. Newcastle: Cambridge Scholars.

Gilquin, G. and M. Paquot (2008), ‘Too chatty: learner academic writing and register variation’. *English Text Construction* 1(1): 41–61.

Granger, S., E. Dagneaux, F. Meunier and M. Paquot (2009), *The International Corpus of Learner English. Version 2*. Handbook and CD-ROM, Louvain-la-Neuve: Presses universitaires de Louvain.

Granger S. and Y. Bestgen (2014), 'The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study', *International Review of Applied Linguistics in Language Teaching* 52: 229–52.

Greenbaum, S. ed. (1996), *Comparing English worldwide: the international corpus of English.* Oxford: Clarendon press.

Graën, J., D. Batinic and M. Volk (2014), 'Cleaning the Europarl Corpus for Linguistic Applications', in *KONVENS*, 222–227.

Grosjean, F. (2001), 'The Bilingual's Language Modes'. In J. Nicol (ed.), *One Mind, Two Languages*, 1–22. Oxford: Blackwell Publishers.

House, J. (2008), 'Beyond Intervention: Universals in Translation?' *Trans-Kom* 1 (1): 6–19.

Hu, X., R. Xiao and A. 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), 347–82.

Ivaska, I. (2015), 'Longitudinal Changes in Academic Learner Finnish: A Key Structure Analysis', *International Journal of Learner Corpus Research* 1 (2): 210–41.

Ivaska, I. and S. Bernardini (2020), 'Constrained language use in Finnish: A corpus-driven approach', *Nordic Journal of Linguistics* 43(1). 33–57.

Iwasaki, S. (2015), 'A Multiple-Grammar Model of Speakers' Linguistic Knowledge'. *Cognitive Linguistics* 26 (2): 161–210.

Jarvis, S. (2000), 'Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon', *Language Learning* 50 (2): 245–309.

Jarvis, S. (2010), ‘Comparison-Based and Detection-Based Approaches to Transfer Research’. *EUROSLA Yearbook* 10: 169–92.

Kaiser, H. F. (1974), ‘An Index of Factorial Simplicity’. *Psychometrika* 39 (1): 31–36.

Kolehmainen, L., L. Meriläinen and H. Riionheimo (2014), ‘Interlingual Reduction: Evidence from Language Contacts, Translation and Second Language Acquisition’, in H. Paulasto, L. Meriläinen, H. Riionheimo and M. Kok (eds), *Language Contacts at the Crossroads of Disciplines*, 3–32, Cambridge: Cambridge Scholars Publishing.

Koppel, M. and N. Ordan (2011), ‘Translationese and Its Dialects’. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1318–26, Portland, Oregon: Association for Computational Linguistics.

Kotze, H. (2018), ‘Expanding the third code: Corpus-based studies of constrained communication and language mediation’, in S. Granger, M.-A. Lefer and L. A. de Souza Penha Marion (eds), *Book of Abstracts, Using Corpora in Contrastive and Translation Studies Conference (5th edition)*, 9–12, CECL Papers 1: Louvain-la-Neuve.

Kruger, H. (2017), ‘The Effects of Editorial Intervention. Implications for Studies of the Features of Translated Language’, in G. de Sutter, M.-A. Lefer and I. Delaere (eds), *Empirical Translation Studies: New Methodological and Theoretical Traditions*, 113–55, Berlin: De Gruyter.

Kruger, H. and B. van Rooy (2016), ‘Constrained Language: A Multidimensional Analysis of Translated English and a Non-Native Indigenised Variety of English’, *English World-Wide* 37 (1): 26–57.

Kruger, H. and B. van Rooy (2018), ‘Register Variation in Written Contact Varieties of English’, *English World-Wide* 39 (2): 214–42.

Kursa, M. and W. Rudnicki (2010), ‘Feature Selection with the Boruta Package’, *Journal of Statistical Software, Articles* 36 (11): 1–13.

Lanstyák, I. and P. Heltai (2012), ‘Universals in Language Contact and Translation’, *Across Languages and Cultures* 13 (1): 99–121.

Miestamo, M., K. Sinnemäki and F. Karlsson, eds (2008), *Language Complexity: Typology, Contact, Change,* Amsterdam: John Benjamins.

Nisioi, S., E. Rabinovich, L. P. Dinu and S. Wintner (2016), ‘A Corpus of Native, Non-Native and Translated Texts’, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4197–4201.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, et al. (2016), ‘Universal Dependencies v1: A Multilingual Treebank Collection’. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 1659–66.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*, London: Longman.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rabinovich, E., S. Nisioi, N. Ordan and S. Wintner, (2016), ‘On the Similarities between Native, Non-Native and Translated Texts’, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1870–81. Berlin: Association for Computational Linguistics.

Revelle, W. (2018), *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Scott, M. (2010), ‘Problems in Investigating Keyness, or Clearing the Undergrowth and Marking out Trails...’, in M. Bondi and M. Scott (eds), *Keyness in Texts*, 43–57. Amsterdam: John Benjamins.

Serbina, T., S. Hintzen, P. Niemietz and S. Neumann (2017), ‘Changes of Word Class during Translation – Insights from a Combined Analysis of Corpus, Keystroke Logging and Eye-Tracking Data’, in S.

Hansen-Schirra, O. Czulo and S. Hofmann (eds), *Empirical Modelling of Translation and Interpreting*, 177–208, Berlin: Language Science Press.

Shlesinger, M. and N. Ordan (2012), ‘More Spoken or More Translated? Exploring a Known Unknown of Simultaneous Interpreting’, *Target* 24 (1): 43–60.

Steiner, E. (2012) ‘A characterization of the resource based on shallow statistics’, in S. Hansen-Schirra, S. Neumann and E. Steiner (eds), Cross-Linguistic Corpora for the Study of Translations, 71–90, Berlin/Boston: de Gruyter.

Straka, M. and J. Straková (2017), ‘Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe’, in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99, Vancouver: Association for Computational Linguistics.

Swallow, H. (2003), ‘Linguistic Interpenetration or Cultural Contamination’, in A. Tosi (ed.), *Crossing Barriers and Bridging Cultures: The Challenges of Multilingual Translation for the European Union*, 104–10, Clevedon: Multilingual Matters.

Szymor, N. (2018), ‘Translation: Universals or Cognition? A Usage-Based Perspective’, *Target* 30 (1): 53–86.

Tono, Y. (2000), ‘A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora’, in B. Lewandowska-Tomaszczyk and J.P. Melia (eds), *PALC'99: Practical Applications in Language Corpora,* 323-340, Frankfurt am Main: Peter Lang.

Toury, G. (2012), *Descriptive Translation Studies – and beyond: Revised Edition,* Philadelphia: John Benjamins.

Ulrych, M. and A. Murphy (2008), ‘Descriptive Translation Studies and the Use of Corpora: Investigating Mediation Universals’, in C. T. Torsello, K. Ackerley, and E. Castello (eds), *Corpora for University Language Teachers*, 141–166, Bern: Peter Lang.

Volansky, V., N. Ordan and S. Wintner (2015), 'On the Features of Translationese', *Digital Scholarship in the Humanities* 30 (1): 98–118.

Wright, M. N. and A. Ziegler (2017), 'Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R', *Journal of Statistical Software* 77 (1): 1–17.