AUTHOR	Ilmari Ivaska, Silvia Bernardini & Adriano Ferraresi
TITLE	The complex case of constrained communication. A corpus-driven, multilingual and multi-register search for the common ground between non-native and translated language
YEAR	2024
DOI	https://doi.org/10.1075/coll.60.07iva
VERSION	Author accepted version
CITATION	Ivaska, I., Bernardini, S. & Ferraresi, A. (2024). The complex case of constrained communication. A corpus-driven, multilingual and multi-register search for the common ground between non-native and translated language. In van Rooy, B. & Kotze, H. (eds). Constraints on Language Variation and Change in Complex Multilingual Contact Settings. (Contact Language Library; Vol. 60). John Benjamins. pp. 191-222. Doi: 10.1075/coll.60.07iva

Chapter 7[1]

The complex case of constrained communication: A corpus-driven, multilingual and multi-register search for the common ground between non-native and translated language

Ilmari Ivaska,[1] Silvia Bernardini[2] & Adriano Ferraresi[2]

[1] University of Turku | [2] University of Bologna

Abstract

In this study we explore the common ground between second-language writing and translated language as instances of constrained language use. Our research design involves three languages (English, Finnish, Italian), two constraining languages and two different registers in each of the three languages. These are compared in terms of frequency of syntactic structures (part-of-speech [POS] bigrams), adopting a corpus-driven method combining keyness analysis and multidimensional analysis. No general constrainedness effects that apply irrespective of languages and registers were observed, but our results point to the centrality of the opposition between verbal and nominal orientation for distinguishing constrained from unconstrained varieties. We conclude with suggestions on how our method and findings could lead to a deeper understanding of constrained language use, and be extended to different modes of language production and to language contact research in general.

Keywords: constrained language, second language, translated language, random forests, multidimensional analysis, dependency bigrams, universal Dependencies

[1] Published in: Ivaska, Ilmari, Silvia Bernardini & Adriano Ferraresi. 2024. The complex case of constrained communication: A corpus-driven, multilingual and multi-register search for the common ground between non-native and translated language. In Bertus Van Rooy & Haidee Kotze (eds.), *Contact Language Library*, vol. 60, 191–222. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/coll.60.07iva. Please refer to the published version.

1. Introduction

The study of constrained language use has been identified as a distinct line of research relatively recently (for an overview see Kotze, 2020, pp. 344–48). In their seminal contribution on the subject, Lanstyák and Heltai (2012, p. 100) acknowledge that communication is always constrained by physical, cognitive and social factors, but describe as "constrained" in a narrow sense any act of communication "taking place under conditions where one or several of the potential limiting factors play a greater than average role". Scholars in second language acquisition (SLA) and translation studies (TS) have since set out to search for similarities between these two types of communication, seen as instances of constrained communication (Ivaska & Bernardini, 2020; Ivaska et al., 2022; Kolehmainen et al., 2014; Kruger & Van Rooy, 2016, 2018; Rabinovich et al., 2016). The underlying idea is that the two disciplines share a common interest in language use where several linguistic systems become simultaneously activated (on bilingual activation in general, see Grosjean, 2001). This kind of activation could constrain second language (L2) and translated language (LT) use, and make texts produced by second-language users and translators diverge in similar ways from comparable native, non-translated texts (L1).

So far, the studies on constrained language use have focused only on single languages, almost exclusively English. Furthermore, the effects of different registers and of constraining languages – the L1s of the L2 users and the source languages of the translations – have not always been exhaustively controlled for, making it impossible to evaluate the generalisability of results. In this paper, we address this research gap by studying three languages (English, Finnish and Italian); for each of them, we consider two constraining languages (German and Italian for English; German and Russian for Finnish; German and English for Italian) and two registers (argumentative and tourism texts for English; academic and narrative texts for Finnish; narrative and news texts for Italian). However much we would have liked to analyse the same registers across all language pairs

and across both types of constraint, such data are not readily available, unless one forces texts into artificial categories. This complexity is inherent to comparisons of different varieties of constrained language, and dealing with it is one of the major methodological challenges to be addressed in empirical investigations like ours. Crucially, this inherent complexity is very much in line with multilingualism as a global norm rather than exception (e.g. Romaine 2001). Finding ways to grapple with such complexity will therefore also allow for building bridges between research on constrained language use and the broader field of language contact research in general. To identify the bigrams that best distinguish constrained (L2 and LT) and unconstrained (L1) varieties, we combine keyness analysis (Gabrielatos, 2018) using Ranger classifiers (Wright & Ziegler, 2017), and multidimensional analysis (Biber, 1988). Our research questions are:

1. Are there any differences in the frequency of syntactic structures distinguishing both L2 and LT from L1? If so, are such differences observable in all studied registers, and amenable to generalisations across the different languages?
2. Are any of these differences, detected in a data-driven manner, interpretable in the light of hypothesised constraints inherent to both L2 and LT communication (e.g., cross-linguistic influence, processing strain, norm orientation)?

The paper is structured as follows. First, we introduce the challenges involved in investigating constrained language use from a multilingual perspective, and explain the rationale for our methodological decisions. On these bases, we then describe our dataset and the ways in which we try to ensure it is representative, and explain our methodological procedure for carrying out cross-linguistically comparable analyses. We then analyse constrained and unconstrained varieties in each of the three languages separately, and discuss similarities and differences between them, commenting on the wider implications of these results. We conclude with an outline of what we see

as the most fruitful future directions in the quest for a deeper understanding of constrained language use.

2. Collocations as signals of constrained language use from a multilingual perspective: Challenges and how to address them

2.1 Variation and cross-linguistic comparability

The study of constrained language has framed itself as addressing phenomena that happen in language in general – rather than in any particular language. This aim is in many ways in line with research paradigms in both SLA and in TS, where the goal is to unearth phenomena emerging irrespective of the combinations of languages involved, such as a more explicit and formal style, fewer involvement features, lower lexical richness, and fewer idiomatic expressions than in L1 (Filipović & Hawkins, 2013; Mauranen & Kujamäki, 2004; Nawal, 2018; O'Brien, 2006). At the same time, both SLA and TS have always shown interest in cross-linguistic influence, underlining the fact that the linguistic systems of both first/source language and second/target language may contribute to the surfacing of differences opposing constrained and unconstrained varieties (for SLA see Jarvis, 2000; for TS see Teich, 2003). The "universality" hypothesis would posit that similar features are to be found in all instances of constrained communication. The "cross-linguistic influence" hypothesis would instead suggest that all instances of constrained communication show traces of the interaction of at least two linguistic systems; the actual traces left in texts – properties of lexicogrammatical encoding in Steiner's words (Steiner, 2005) – should, by definition, differ depending on the language pairs under investigation. Studies of constrained language should therefore reconcile the universalist and the variationist, both in terms of dataset and, consequently,

in terms of feature selection: More than one pair of constraining/constrained languages should be included in the corpus, and the lexicogrammatical features selected for analysis should not be specific, or better suited, to any given language or language pair.

A second challenge concerns register variation and text comparability, in particular at the cross-linguistic level. Since "each register has its own grammar of use" (Biber & Conrad, 2009, p. 216), taking into account cross-register variation is an indispensable methodological step when evaluating the generalisability of a study's results. The same argument has been put forward within cognitive linguistics by Iwasaki's multiple-grammar model (Iwasaki, 2015). In situations that inherently include several languages, such as translation and second-language use, an interplay is typically observed between the different languages and their register variation. While some register conventions are common to different linguacultures, others are likely to be language-specific (e.g., Lefer & Vogeleer, 2013; Upton & Connor, 2001). Research on bilingualism has shown how registers in the language users' different language systems influence each other (e.g., Gentil, 2011; Kobayashi & Rinnert, 2013; Mein, 2012); similarly, in TS Szymor (2018) has shown that translations in a highly specialised register display patterns typical of other, more general registers, suggesting that register knowledge is transferred in a usage-based manner. Stemming from quite separate traditions, these studies point to interactions between different languages, different language varieties, and different registers within language varieties. One should therefore be careful, when analysing constrained language, to include separate registers and not conflate them, so that results are both valid and generalisable.

Given the nature of the data required for constrained language analysis, featuring multiple languages and multiple registers, methods are needed to cope with substantial (cross-)linguistic variation in text comparison. To this effect, Biber (2014) calls for implementing multidimensional analysis, a statistical dimension-reduction technique for condensing multiple linguistic variables into a few basic parameters of linguistic variation. This is applied separately for each language, as a

basis for “comparing the patterns of register variation across languages: linguistically [...], functionally [...], and situationally” (Biber, 2014, p. 9). The fact that similar dimensions are found to describe register variation in multiple languages suggests that general tendencies of linguistic variation exist across languages (see Biber, 2014). Yet how to select the variables to be used in the first place, independently for each language, remains an open question, which is addressed in the next section.

2.2 Locating constrainedness effects: What features should be used?

Review articles addressing bi-/multilingual language use and including translation as a mode of language contact (though not referring explicitly to constrainedness) have generally privileged theoretically motivated and language-pair-specific features. Focusing on source-language transfer leading to contact-induced language change, Kranich (2014, p. 104) mentions personal pronouns and modal markers (English to German), features for managing author-reader interaction (English to Hungarian), possessive determiners, demonstrative pronouns and coordinating vs subordinating structures (English to Italian), all pointing to an influence of source-language conventions at the pragmatic level. Kolehmainen et al. (2014, p. 3) focus on interlingual reduction instead, described as “the lower frequency of target language linguistic items or patterns not shared by both of the languages involved in the language contact situation”. This phenomenon has been observed in language contact, SLA and translation, though it has been referred to by different names (the more common term used in TS being “under-representation of unique items”; Tirkkonen-Condit, 2004). As in the case of Kranich (2014), the selection of items and structures is directional (and thus hardly generalisable), as it takes into account (dis)preferences and gaps between source and target language at the lexical or morpho-syntactic levels. Examples cited in Kolehmainen et al. (2014, pp. 8–9) include different types of Finnish infinitive constructions in translations from English and

Russian (Eskola, 2002, 2004), English manner of motion verbs in translation from French (vs German) (Cappelle, 2012), and use of English phrasal verbs by Finnish and Swedish learners Sjöholm, 1995). For a research setup that includes different language pairs, language-pair-specific features are clearly inadequate.

Turning to the few empirical studies that contrast translation and other kinds of constrained language use, a broad range of linguistic phenomena have been targeted so far. In their search for similarities between native, non-native and translated English, Rabinovich et al. (2016) rely on features traditionally used for text classification tasks and language modelling, such as function words, part-of-speech (POS) triplets, and tokens in initial and final sentence positions, as well as theoretically motivated metrics from previous studies on translationese and learner language, such as type-token ratio, frequency of idiomatic expressions, and frequency of personal and possessive pronouns. Most of the metrics used confirm the existence of similarities between non-native and translated English: Both are found to be less lexically rich and to use fewer idiomatic expressions and pronouns, and more explicit cohesive devices, than native English.

Kruger and Van Rooy (2016, 2018) use the feature set and methodology originally introduced by Biber (1988) for register analysis, applying it to the comparison of native, non-native and translated English texts. This feature set consists of 67 features whose co-occurrence can be interpreted in terms of functional variation, such as tense and aspect markers (past, present, perfect), place and time adverbials, pronouns and pro-verbs, and so on. In the first study, the three varieties are set against Biber's original dimensions of variation, while in the second a new factor analysis of six registers and 16 varieties (seven native, eight non-native, one translated) is performed. Both studies find a higher degree of formality and a lower degree of involvement in constrained than in unconstrained English.

Most of the features used in the constrained language analyses reviewed in this section are language-specific, and thus not readily generalisable to languages other than English. The only

exception are the features used for text classification and language modelling by Rabinovich et al. (2016), which, however, are hardly human-interpretable. The approach taken in this paper (as well as in previous work by Ivaska & Bernardini, 2020 on constrained Finnish, and Ivaska et al., 2022 on constrained English) aims to explore the possibility of using the Universal Dependencies (UD) scheme to identify cross-linguistically comparable lexico-syntactic structures, using these as input to a multidimensional analysis. UD is an initiative which seeks to bring the linguistic description of many languages under a common framework. The comparability between data in different languages is maximised by common data formatting, common principles and common core feature sets for the annotation guidelines of word segmentation, morphology and syntax (Nivre et al., 2016).[2] Furthermore, documentation is provided in an accessible manner to facilitate cross-linguistic comparisons. At the same time, the scheme is flexible and allows for the tagging of language-specific phenomena that can be introduced in different annotation layers as necessary.

The procedure adopted in this contribution is corpus-driven, in the sense that "recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories" (Tognini-Bonelli, 2001, p. 84), rather than the opposite course of action, in which theoretically motivated features are preselected to the exclusion of other, potentially more relevant features not previously observed. At the same time, the structures are both interpretable by humans (since we consider syntactically-defined POS bigrams), and language-specific yet cross-linguistically comparable (thanks to UD). Relying on keyness analyses to identify the most salient bigram structures, and on multidimensional analyses to find out if and how they group together into functionally interpretable dimensions of variation (see Section 3.2), our method is radically exploratory. Since the features used in the factor analysis are not a priori motivated, there is no certainty that they are effective in clustering texts; at the same time, any regularities observed are particularly informative. One last point that needs to be addressed is the choice to focus on

[2] See the project website for the latest version: https://universaldependencies.org (visited 1 June 2021).

syntactically-defined POS bigrams as collocation-yielding structures. This is taken up in the next section.

2.3 Collocations in focus

The choice to limit the scope to syntactically-defined POS bigrams is essentially an arbitrary one, yet it allows us to capture information about collocations, or recurrent word combinations. Substantial work on collocation, and phraseology in general, has been carried out both in TS and SLA/learner language studies. In both fields, the use of phraseological units is assumed to signal (non-)native-like or (un)conventional language. Within TS, differences in frequency of conventional phrases in translated vs non-translated language have been interpreted against the background of purported translation norms or universals, such as source-language interference or standardisation/normalisation (for an overview, see e.g., Marco, 2009, pp. 844–847). Within SLA and learner language studies, collocation use has been linked to L2 proficiency, and specific aspects of phraseological competence have been observed in L2 language production, such as the greater sensitivity to frequency effects with respect to L1 speakers (Durrant & Schmitt, 2009; Wray, 2002).

In both TS and SLA/learner language studies, research on collocations adopting corpus methods has traditionally focused on bigrams defined on the basis of word forms (e.g., in Durrant & Schmitt, 2009) or shallow POS annotation (e.g., Bernardini, 2011; Ferraresi & Miličević, 2017; Granger & Bestgen, 2014). Relying on word proximity or adjacency, paired with frequency and/or statistical significance testing, researchers try to maximise chances that the bigrams are structurally connected. However, in this way more complex structures and longer-distance relations may be artificially excluded from the count, while incomplete or badly formed pairs may be inadvertently included. With our method, which extracts bigrams from parsed corpora on the basis of syntactic

dependencies, we are able to disregard distance between the words in the bigram, their constituent order and hierarchy.

For this exploratory study, we do not restrict our study to lexical words, nor do we predetermine the POS or the structures. Instead, we use corpus-driven statistical techniques to highlight the POS patterns whose frequencies differ the most across constrained and unconstrained sub-corpora. This method has the merit of making the fewest possible assumptions about the form that phrases take, and is thus adequate for research frameworks in which multiple languages are included. Section 3 provides a step-by-step account of our analysis of typical phrasal structures across constrained and unconstrained English, Finnish and Italian.

3. Data and method

3.1 Data used

In our view, the questions of language-specificity on the one hand, and generalisability on the other, operate on different levels of abstraction – and the great challenge for the study of an extremely complex phenomenon like constrained language use is to acknowledge that not everything can be fully accommodated, and yet try to do what is possible (Leech, 2006). We would suggest that there are a minimum of three data-related criteria a contrastive study needs to address in order to support a generalisable, language-agnostic argument for typicalities of constrained language use. Data need to include:

1. multiple constrained varieties and an unconstrained reference variety, to tease apart constraint-specific tendencies;
2. multiple language combinations with typologically different languages, to tease apart direct

cross-linguistic influence;

3. multiple registers, to tease apart register effects.

Furthermore, the data should be maximally comparable across all these variables, making it possible to isolate the likely reasons for the observed variation. In reality, such a dataset is almost impossible to collect, and one has to make do with what is feasible. Table 7.1 summarises the size and provenance of the data used in this study. We contrast three languages (English ["en"], Finnish ["fi"], Italian ["it"]), with one unconstrained variety ("L1") and two constrained varieties (non-native ["L2"] and translated ["LT"]). Each of these includes two constraining languages (German ["de"] and Italian for English; German and Russian ["ru"] for Finnish; English and German for Italian) and two registers (argumentative ["arg"] and tourism ["tou"] for English; academic ["aca"] and narrative ["nar"] for Finnish; narrative and news ["new"] for Italian). Additionally, we also control for editing effects (following Kruger, 2017), keeping published and unpublished texts separate in the pairwise comparisons. We use existing resources where available (full references are provided in the Appendix) and complement them with purpose-built ones where necessary (these are marked as "ad hoc" in Table 7.1). We thus end up with 34 subcorpora of very different sizes. As will be described below, we match the corpora used to make maximal use of existing resources while relying on balanced data samples.

	English [en]		Finnish [fi]		Italian [it]	
	argumentative [arg]	**tourism [tou]**	**academic [aca]**	**narrative [nar]**	**narrative [nar]**	**news [new]**
L1	published: 93 697 (ad hoc) unpublished: 168 368 (LOCNESS)	35 288 (ad hoc)	published: 1 180 564 (CTF & LAS1) unpublished: 392 588 (LAS1)	published: 2 769 603 (CTF & InterCorp) unpublished: 32 748 (ad hoc)	published: 1 781 038 (PEC) unpublished: 129 162 (VINCA)	791 978 (PEC)
L2	de-unpublished: 108 986 (ICLE) it-unpublished: 116 355 (ICLE)	de: 86 796 (ad hoc) it: 69 042 (ad hoc)	de-unpublished: 32 670 (ICLFI & LAS2) ru-unpublished: 166 905 (ICLFI & LAS2)	de-unpublished: 10 015 (ICLFI) ru-unpublished: 20 085 (ICLFI)	de-unpublished: 15 255 (VALICO) en-unpublished: 13 123 (VALICO)	de: 31 725 (ad hoc) en: 28 275 (ad hoc)
LT	de-published: 138 236 (ad hoc) it-published: 122 534 (ad hoc)	de: 66 051 (ad hoc) it: 92 401 (ad hoc)	de-published: 114 751 (CTF & FinDe) ru-published: 51 285 (CTF)	de-published: 334 448 (CTF & InterCorp) ru-published: 650 880 (CTF)	de-published: 771 976 (InterCorp) en-published: 2 572 269 (InterCorp)	de: 44 6971 (InterCorp) en: 64 0560 (InterCorp)

Table 7.1. Corpus size (in number of words) and provenance. For corpus abbreviations, see the Appendix

3.2 Method

We adopt the same method as Ivaska and Bernardini (2020) and Ivaska et al. (2022), as described in Sections 2.2 and 3.2.2. All data handling is done through Java scripts, and all statistical analyses in

R (R Core Team, 2018). The R scripts and frequency data are available for consultation from the Open Science Framework repository.[3]

3.2.1 A cross-linguistically comparable feature set

We use syntactically defined POS bigrams as features in the analysis; in other words, combinations of two POS connected to one another by a syntactic dependency, and not necessarily adjacent. Each considered variable is a bigram of two POS, together with the syntactic relationship between the two, their order and hierarchy. For instance, the English sentence in Figure 7.1 begins with bigrams PRONNODE_nsubj_VERBHEAD (*I–like*), VERBHEAD_obj_NOUNNODE (*like–curries*) and PRONNODE_nmod:poss_NOUNHEAD (*their–curries*). The Finnish sentence in Figure 7.2 begins with bigrams PROPNNODE_nsubj_VERBHEAD (*Auron–asetti* 'Auron placed'), VERBHEAD_obj_NOUNNODE (*asetti–kätensä* 'placed hand') and ADJNODE_amod_NOUNHEAD (*vasemman–kätensä* 'left hand'). The Italian sentence in Figure 7.3 begins with bigrams PRONHEAD_acl:relcl_VERBNODE (Q*uello–vuole* 'that [thing] wants'), PRONNODE_nsubj_VERBHEAD (*che–vuole* 'that [s/he] wants') and PRONNODE_nsubj_PRONHEAD (*Quello–qualcosa* 'that [thing] something]').

[3] At https://osf.io/vayb5/ (visited 1 June 2021).

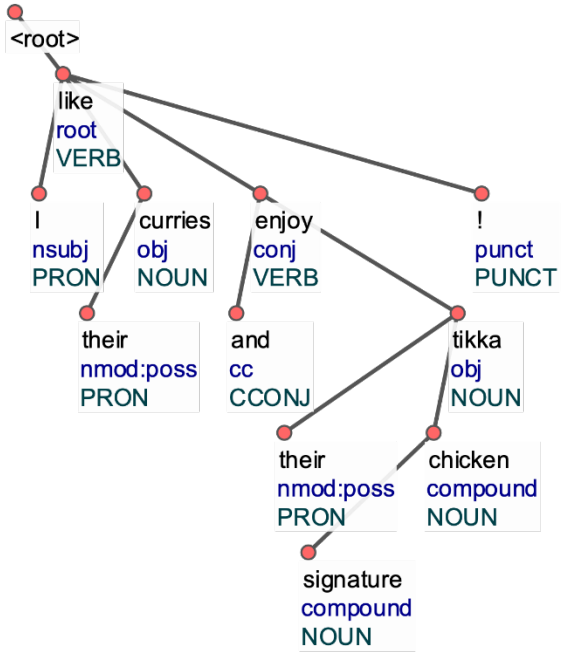

Figure 7.1. Example sentence tree for the English sentence *I like their curries and enjoy their signature chicken tikka!*

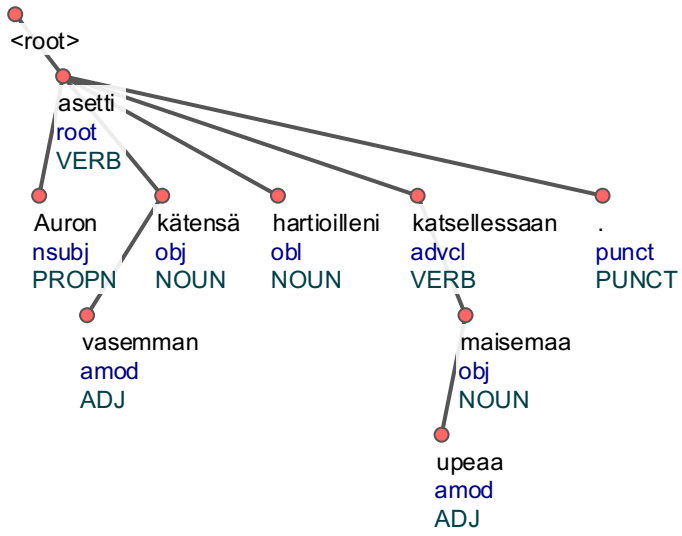

Figure 7.2. Example sentence tree for the Finnish sentence *Auron asetti vasemman kätensä hartioilleni katsellessaan upeaa maisemaa* 'Auron placed his/her left arm on my shoulders while watching the marvellous landscape'

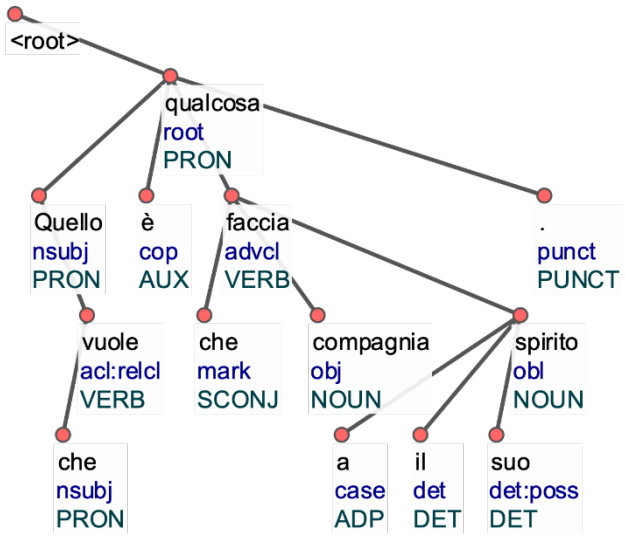

Figure 7.3. Example sentence tree for the Italian sentence *Quello che vuole è qualcosa che faccia compagnia al suo spirito* 'What (s)he wants is something that keeps company to his/her soul'

Such POS bigrams allow for cross-linguistic comparability: The POS tags[4] as well as the dependency relations[5] for all languages stem from lists of universal tags used in all UD-based annotations. On the other hand, each language implements only the relevant tags, with additional language-specific subtypes of syntactic relations as needed.[6] As the bigrams also capture word order and syntactic hierarchy, the feature set can be seen to maximise comparability without losing its flexibility in terms of cross-linguistic differences.

3.2.2 Methodological workflow

The comparison has five steps, as depicted in Figure 7.4.

[4] See https://universaldependencies.org/u/pos/index.html (visited 1 June 2021).

[5] See https://universaldependencies.org/u/dep/index.html (visited 1 June 2021).

[6] See https://universaldependencies.org/fi/ for Finnish, https://universaldependencies.org/it/ for Italian (visited 1 June 2021).

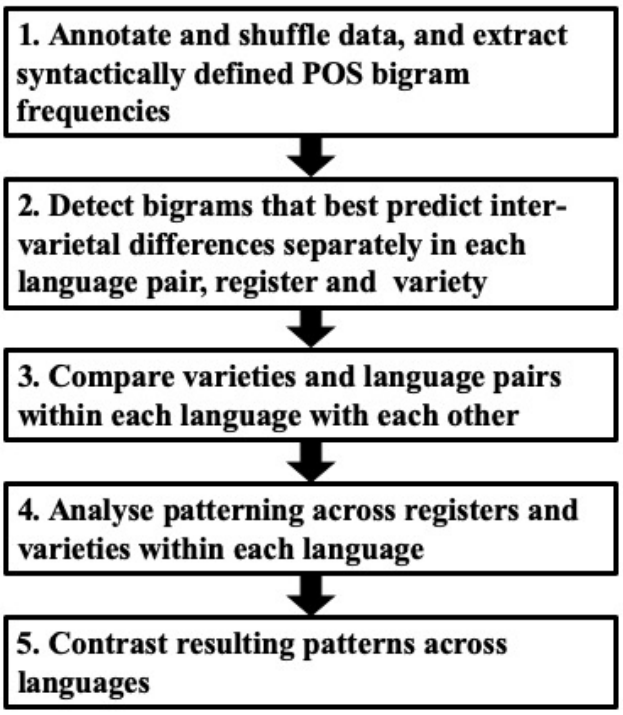

Figure 7.4. Methodological workflow

In Step 1, all data are parsed using the UDPipe parser version 1.2 (Straka & Straková, 2017); to avoid the effect of text length and topical differences, we randomly shuffle the sentences of each subcorpus and re-structure them in blocks of 50 sentences; we then extract the normalised frequencies of all POS bigrams from each block. The reported performance of the parser varies a little across the three languages, but can be considered relatively comparable, with POS F1-scores of 93.50 for English, 94.49 for Finnish, and 97.31 for Italian, and with dependency accuracies of 77.25 for English, 77.26 for Finnish, and 86.11 for Italian.

Step 2 consists of a keyness analysis: The 1 000 most frequent bigrams in each language are used as features in eight pairwise comparisons – between each register-, variety- and L1-/source-language-specific constrained subcorpus and the corresponding unconstrained subcorpus. For feature selection we use the Boruta wrapper algorithm (Kursa & Rudnicki, 2010). The algorithm creates copies of all variables and permutates their values randomly. It then runs a Ranger classifier (Wright & Ziegler, 2017) – a binary classifier between constrained and unconstrained – using both

the actual and the randomised variables and identifies as important variables those that consistently outperform the randomised ones. After multiple iterative repetitions, the method produces a list of variables that can be considered relevant for the classification task – the key features of that pairwise comparison. Different subcorpora are of different size, and to ensure the equal contribution of each subcorpus, we match the data by the amount of included text blocks in the compared subcorpora, so that as many blocks as possible are used in each pairwise comparison. That is, each pairwise comparison has the same amount of text blocks from both corpora, and the amount is defined by the smaller corpus.

Next, in Step 3, we compare the results of the pairwise comparisons in each language to find the POS bigrams that contribute to distinguishing between constrained and unconstrained varieties. We treat as consistent key features those that are identified as important in over half of the pairwise comparisons for each language.

In Step 4, we conduct, for each language, a multidimensional analysis using the consistent key features as variables. For this, we *z*-standardise the frequencies and take a maximum balanced sample from all subcorpora, so that the same amounts of text blocks are included from each subcorpus. We follow the process described by Egbert and Staples (2019), using functions from the *psych* package (Revelle, 2018). Multidimensional analysis employs exploratory factor analysis to group together multiple variables, in our case the key features, according to their co-occurrence patterns. The resulting groupings (dimensions), can be interpreted in relation to their patterning across registers and varieties.

The fifth and final step consists of comparing qualitatively the results between the three studied languages to identify whether the resulting dimensions in the different languages – and the bigrams contributing to these dimensions – have anything in common. This, in turn, allows for evaluating the potential general characteristics of constrained language use across the studied

languages, as well as any regularities that apply to SLA only or translation only, consistent with Lanstyák and Heltai's (2012, p. 117) original recommendations.

3.2.3 Keyness analysis and factor solutions

Out of the 1 000 most common bigram patterns, we identified 19 consistent key features for constrained English, 32 for constrained Finnish, and 29 for constrained Italian (shown in Tables 7.3, 7.4 and 7.5 in Section 4). These POS bigrams indicate systematic differences between constrained and unconstrained data in over half of the eight pairwise keyness analyses of each respective language. This threshold ensures that none of the key features stems from comparisons of any single register, any single type of constraint, or any single constraining language. For the subsequent analyses, we took a balanced maximum-sized data subset of text blocks sampled evenly across varieties, registers, and L1s/source languages in each language. This was done to ensure equal contribution of each data subset. For English, the sampled data include 1 133 text blocks, for Finnish 768 text blocks, and for Italian 506 text blocks. The Kaiser-Meyer-Olkin (KMO) measures of sampling adequacy in Table 7.2, ranging between 0.88 and 0.90, indicate that all samples are factorable: Following Kaiser's suggestions (1974, p. 35), values above 0.5 are acceptable and those above 0.8 meritorious.

	English	**Finnish**	**Italian**
KMO	0.88	0.90	0.88

Table 7.2. Overall measure of sampling adequacy

We then proceeded to carry out multidimensional analyses for the three languages. As is customary in multidimensional analysis studies (Egbert & Staples, 2019), we chose the number of dimensions based on factors whose eigenvalues are above one (the eigenvalue is a standardised measure of the proportion of variance explained by the factor). As the scree plots in Figure 7.5

show, this means a two-factor solution for English and a three-factor solution for Finnish and Italian. Again, consistent with previous studies, we use a threshold of 0.35 for factor loadings for features to be included in the analysis (factor loadings are estimates of the contribution the factor makes to the variance of the feature in question). Essentially, this phase makes it possible to bundle up and interpret together those identified key features that are inter-related. Features reaching the threshold for more than one dimension were analysed in each of them, but included in the dimension score calculations only in the highest loading dimension.

Due to the exploratory nature of this chapter and in the interest of space, we limit the analysis to the first dimension for each language. It should be remembered that multidimensional analysis is used here to provide a quick way of clustering potential constrainedness features identified bottom up. As shown by the plots in Figure 7.5, the amount of variation accounted for by the first dimension is by far the largest one: Any constrainedness effects in the data are likely to show at their clearest there.

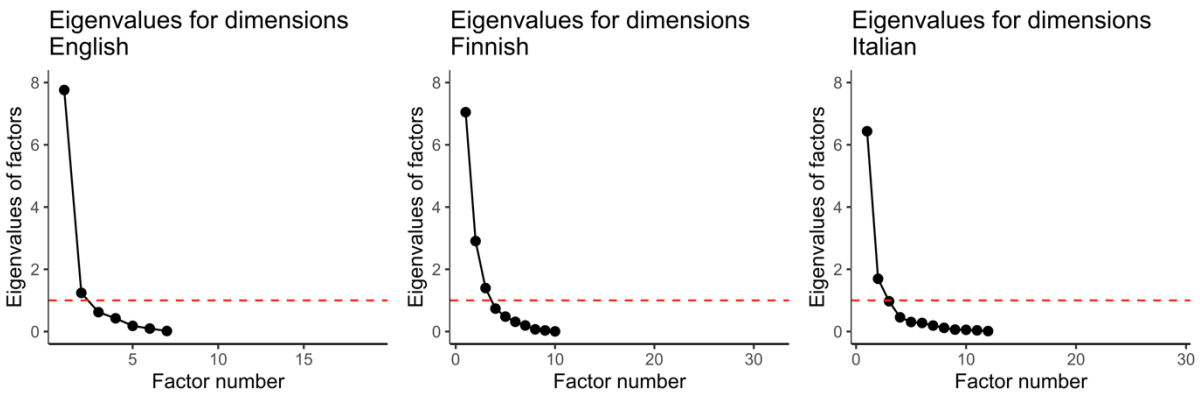

Figure 7.5. Scree plot of eigenvalues in the different data sets.

4. Results

4.1 English Dimension 1: Nominal vs verbal orientation

The first dimension in English reflects differences related to the frequency of syntactic structures involving nouns or verbs (see Table 7.3), a distinction that is reminiscent of Steiner's *nominal* vs *verbal* orientation (2008, pp. 328–329). Features loading positively onto this dimension reflect a preference for nominal word classes and various forms of nominal modification, whereas negatively loading features indicate a preference for verbs in different clause-level constructions.

Positive features		**Negative features**	
PROPNNODE_compound_PROPNHEAD (proper noun compound)	0.793	SCONJNODE_mark_VERBHEAD (subordinating conjunction introducing verb)	–0.939
ADPNODE_case_PROPNHEAD (preposition as case marker of proper noun)	0.787	VERBHEAD_ccomp_VERBNODE (verb as clausal complement of verb)	–0.928
NOUNHEAD_nmod_PROPNNODE (noun post-modified by proper noun)	0.750	NOUNNODE_nsubj_VERBHEAD (noun as subject of verb)	–0.804
NOUNNODE_compound_NOUNHEAD (noun compound)	0.594	PARTNODE_advmod_VERBHEAD (particle as adverbial modifier of verb)	–0.701
DETNODE_det_PROPNHEAD (determiner introducing proper noun)	0.579	PRONNODE_nsubj_VERBHEAD (pronoun as subject of verb)	–0.510
VERBNODE_amod_NOUNHEAD (verb as adjectival modifier of noun)	0.385		

Table 7.3. POS bigrams loading onto English Dimension 1

Bigrams associated with positive dimension scores include on the one hand compounds (see *sea–views*, in Example 1) and nouns pre-modified by verbs used as adjectival modifiers

(*surrounding–woods* in Example 2), and on the other hand phrasal constructions involving proper nouns. These include prepositional phrases (*on–Sundays* in Example 3a) which might be headed by a common noun (*brunch–Sunday* in Example 3b), as well as proper nouns introduced by a determiner (*the–'Ndrangheta* in Example 0) or combined into multi-word proper nouns (*Outer–Hebrides* in Example 0).

(1) Be inspired by our stunning sea views.

(2) Wander through the surrounding woods and visit the Woodland Centre.

(3a) The brunch on Sundays (09:00-16:00) is excellent!

(3b) The brunch on Sundays (09:00-16:00) is excellent!

(4) Images of the Madonna di Polsi […] (revered by members of the 'Ndrangheta) hung on the wall.

(5) Many people enjoy dedicated active breaks in the Outer Hebrides.

The structures loading negatively on Dimension 1 include verbs introduced by nominal subjects (*Detectives–obtained* in Example 6) or pronominal ones (*I–think*, *he–invite* in Example 7) and verbs modified by a particle, in most cases the negation particle *not* (as in Example 8, *not–healed*). They also include bigrams associated with clause complexity, including subordinating conjunctions introducing a verb in a clause (*although–seen* in Example 9) and verbs acting as clausal complements (*make–feel* in Example 10).

(6) Detectives obtained the evidence.

(7) What I think is that he should also invite people to make a good use of it.

(8) A sign, perhaps, that the wound has not yet healed.

(9) Although time, in the central section of the book, is actually seen as a supernatural force.

(10) This would make students feel less alienated and more at home.

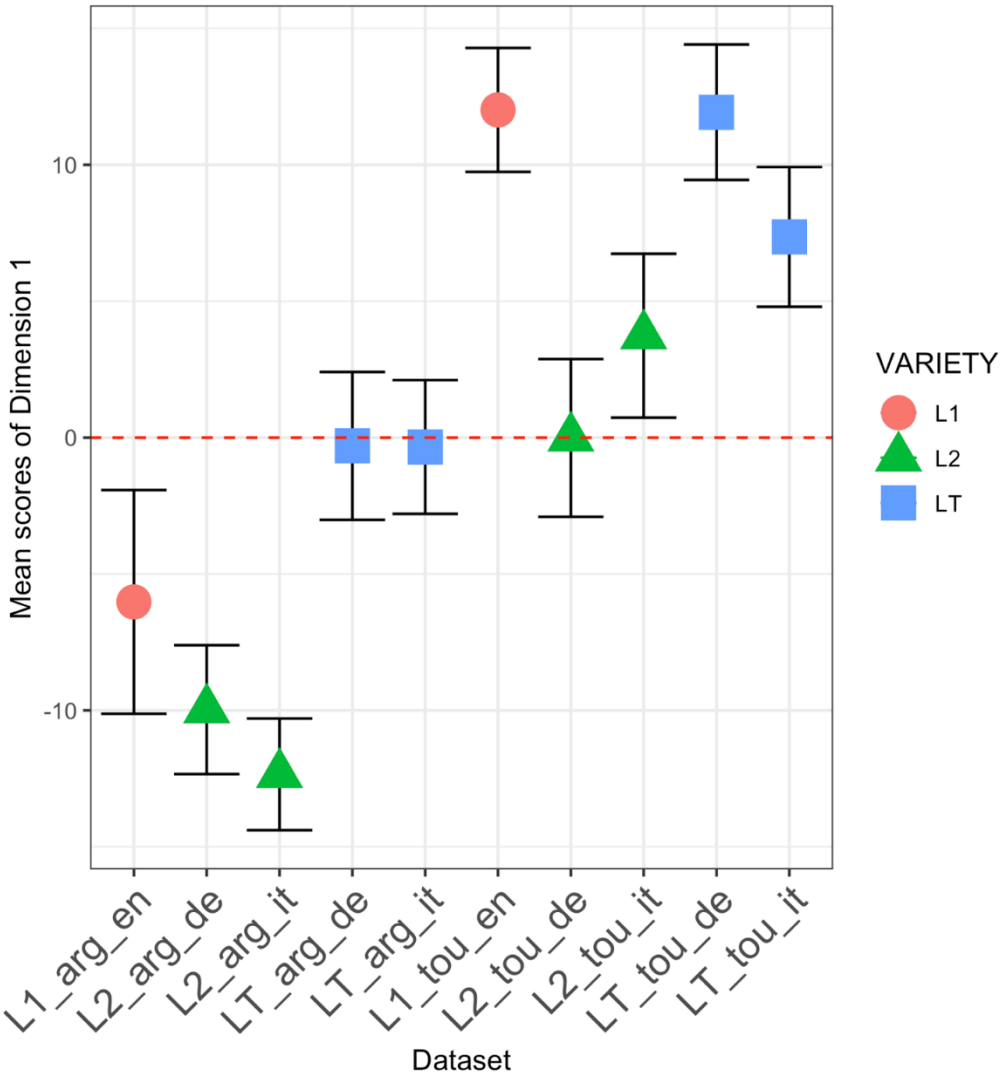

Figure 7.6. Mean scores and standard deviations for English Dimension 1

Figure 7.6 displays averages and standard deviations for scores on Dimension 1 for English, separately for each subcorpus representing a different combination of variety (unconstrained, non-native and translated, named "L1", "L2" and "LT" respectively in the plot), register (argumentative and tourist, i.e., "arg" and "tou"), and, for the constrained sub-corpora, the first language of non-native writers or the source language of the translated texts (German and Italian, i.e., "de" and "it"). This dimension clearly distinguishes the two registers, with argumentative texts being characterised

by negative scores, reflecting verbal orientation, and tourist texts by positive scores, reflecting an opposite tendency towards nominal orientation. When focusing on the comparison between constrained and unconstrained varieties, translated texts have consistently less extreme scores than L1 texts: They seem to display a less marked nominal orientation in tourist texts (especially those translated from Italian), and a correspondingly less marked verbal orientation in argumentative ones. By contrast, L2 displays lower scores than L1 and LT in both tourist and argumentative texts, indicating heavier reliance on verbs than nouns in the two registers. Interestingly, these patterns of reduced register variation in translated texts and verbal orientation in L2 texts seem to apply irrespective of different constraining languages. Finally, it is evident that the constrained subcorpora cluster together within the same register based on their variety more than their L1 background or source language.

4.2 Finnish Dimension 1: Verbal vs nominal orientation

As with English, the first dimension in the Finnish data can also be seen to reflect a difference related to verbal vs nominal orientation (see Table 7.4). Features loading positively onto this dimension reflect a verbal/clausal orientation, while features loading negatively reflect a nominal orientation.

Positive features		**Negative features**	
VERBHEAD_xcomp_VERBNODE (verb as clausal complement of verb, same subject)	0.634	NOUNNODE_nsubj.cop_NOUNHEAD (noun subject of nominal copula clause)	–0.643
VERBHEAD_advcl_VERBNODE (clausal modifier of verb)	0.612	NOUNNODE_nsubj.cop_ADJHEAD (noun subject of adjectival copula clause)	–0.564

PROPNNODE_nsubj_VERBHEAD (proper noun as subject of verb)	0.584	NOUNHEAD_conj_NOUNNODE (noun phrase coordination)	–0.518
VERBHEAD_obj_PRONNODE (pronoun as object of verb)	0.466	CCONJNODE_cc_NOUNHEAD (coordination with noun as conjunct)	(–0.488)
VERBHEAD_parataxis_VERBNODE (coordinated non-copular clauses with no explicitly marked coordination)	0.430	NOUNNODE_nmod.poss_NOUNHEAD (pre-nominal noun as genitive modifier)	(–0.356)
PRONNODE_nsubj_VERBHEAD (pronoun as subject of verb)	(0.403)		
VERBHEAD_xcomp.ds_VERBNODE (verb as clausal complement of verb, different subjects)	0.382		
VERBHEAD_obl_NOUNNODE (post-verbal nominal adjunct)	0.382		

Table 7.4. POS bigrams loading onto Finnish Dimension 1; factor loadings in parentheses indicate features that have a higher loading on another dimension

When the bigrams with positive dimension scores – reflecting a verbal orientation – are analysed in greater detail, they can be seen to reflect two partially hierarchical phenomena: Complete non-copular clauses in general and complex verbal constructions in particular. Bigrams reflecting typical non-copular clauses include verbs followed by pronoun objects and nominal adjuncts (*lannistanut–häntä* 'discouraged–him/her' in Example 11 and *piti–juhlatilaisuudessa* 'gave–at the ceremony' in Example 12) as well as verbs preceded by pronoun or proper noun subjects (*hän–pakotti* '(s)he–forced' in Example 13 and *Aragorn–sanoi* 'Aragorn–said' in Example 14a). Paratactic coordination (*herätä–sanoi* 'wake–said' in Example 14b) can also be grouped with these features reflecting non-copular clauses. Features associated with complexity at the clause level include (often non-finite) clausal modifiers (*asetti–katsellessaan* 'placed–while watching' in

Example 15), as well as various clausal complements (*kertoi–porottavan* ‘indicated–was shining’ in Example 16 and *ryhtyi–viheltelemään* ‘began–to whistle’ in Example 17).

(11) *Se ei ollut suinkaan lannistanut häntä*.

‘It most certainly hadn’t discouraged him/her.’

(12) *Amiraali piti juhlatilaisuudessa puheen.*

‘Admiral gave a speech at the ceremony.’

(13) *Hän pakotti itsensä kysymään.*

‘(S)he forced him/herself to ask.’

(14a) *Herätä sitten minut, Aragorn sanoi.*

‘Wake me up then, Aragorn said.’

(14b) *Herätä sitten minut, Aragorn sanoi.*

‘Wake me up then, Aragorn said.’

(15) *Auron asetti vasemman kätensä hartioilleni katsellessaan upeaa maisemaa.*

‘Auron placed his/her left arm on my shoulders while watching the gorgeous landscape.’

(16) *Syvän sininen väri kertoi auringon porottavan keskipäivää.*

‘The deep blue color indicated that the sun was shining the noon.’

(17) *Hän ryhtyi viheltelemään viattomasti*.

‘(S)he began to whistle innocently.’

The negatively loading bigrams include copula structures (*osa–runoutta* ‘part–poetry’ in Example 18 and *validiteetti–mahdollista* ‘validity–possible’ in Example 19), phrasal coordination (*kieli–kulttuuri* ‘language–culture’ in Example 20a and *ja–kulttuuri* ‘and–culture’ in Example 20b) as well as nominal phrasal modification (*kysynnän–puute* ‘lack of demand’ in Example 21), all pertaining to nominal orientation.

(18) *Suurin osa naisten kirjoittamasta kirjallisuudesta oli runoutta.*

‘The major part of the literature written by women was poetry.

(19) *Validiteetti ei ole mahdollista ilman reliabiliteettia.*

‘Validity is not possible without reliability.’

(20a) *Lyydin kieli ja lyydiläisten kulttuuri vähitellen häviävät.*

‘The Ludic language and the Ludic culture slowly disappear.’

(20b) *Lyydin kieli ja lyydiläisten kulttuuri vähitellen häviävät.*

‘The Ludic language and the Ludic culture slowly disappear.’

(21) *Kysynnän puute lisää yrityksien velkaa.*

‘The lack of demand increases the debt of the companies.’

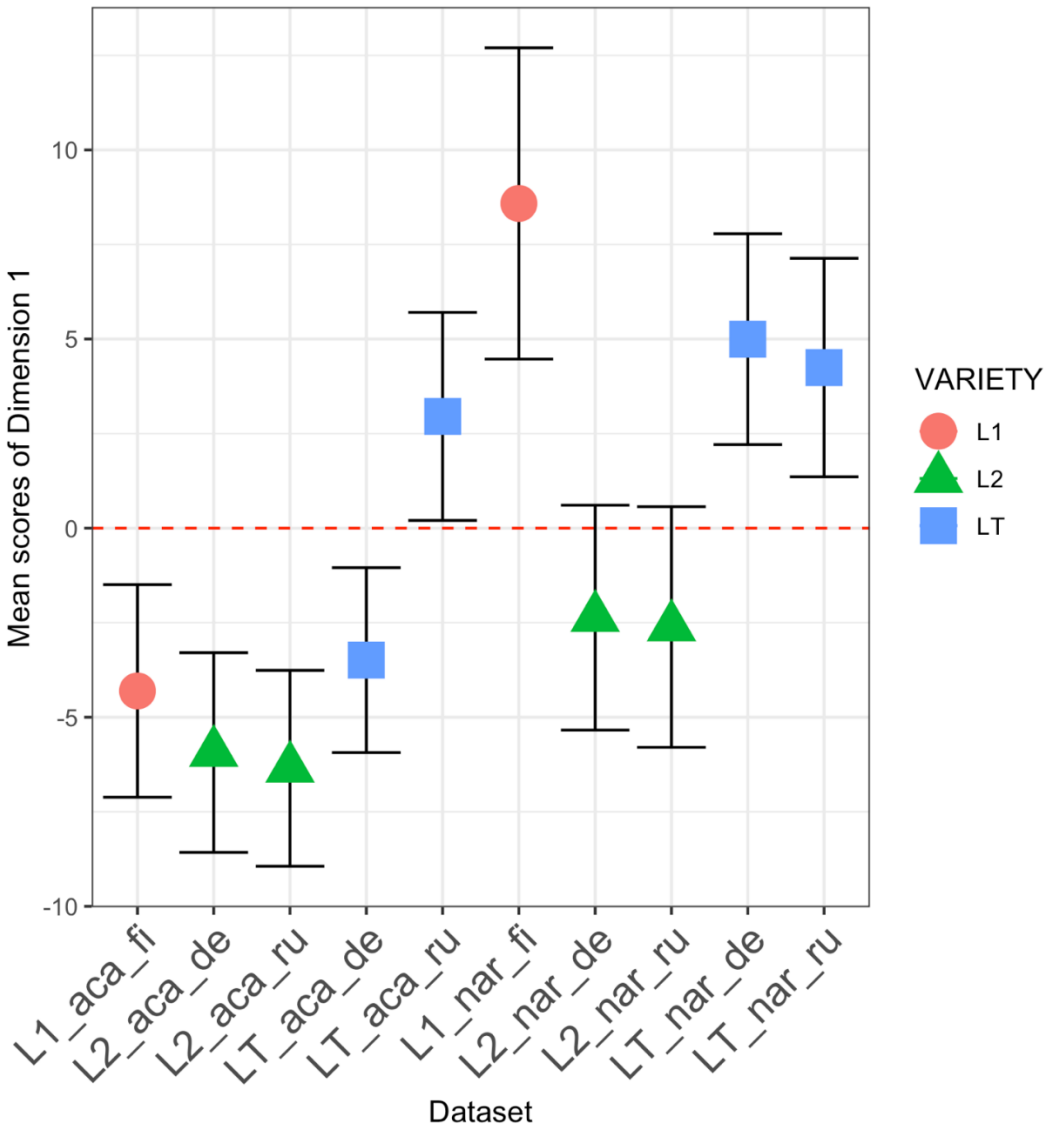

Figure 7.7. Mean scores and standard deviations for Finnish Dimension 1

Figure 7.7 visualises dimension score averages and standard deviations for Finnish Dimension 1 for each variety-, register-, and L1-/source-language-specific subcorpus. In L1, this dimension distinguishes the academic ("aca") from the narrative ("nar") register, with academic texts characterised by negative scores and a preference for nominal style, and narrative texts characterised by positive scores and a more verbal orientation. As in English, translated texts display less extreme (negative or positive) scores than L1 texts, converging towards the centre of the plot. By contrast, L2 texts show a clear overall preference for nominal style, irrespective of

register. Interestingly, with the exception of academic texts translated from Russian, constrained subcorpora cluster in each register according to the variety rather than the respective L1/source language. As was the case with English constrained by Italian and German, the results suggest that the observed phenomena are unlikely to be due to the constraining languages, German and Russian, as they do not lead to diverging tendencies on Finnish.

4.3 Italian Dimension 1: Verbal vs nominal orientation

Dimension 1 in Italian closely reflects the corresponding dimensions in English and Finnish. Here, positive scores are associated with more frequent use of verbal structures, whereas negative scores reflect a preference for nominal structures, two of which involve proper nouns.

Positive features		**Negative features**	
PRONNODE_iobj_VERBHEAD (pronoun as indirect object of verb)	0.762	NOUNHEAD_nmod_NOUNNODE (noun post-modified by noun)	–0.904
CCONJNODE_cc_VERBHEAD (coordinating conjunction introducing verb)	0.744	ADPNODE_case_NOUNHEAD (preposition as case marker of proper noun)	–0.748
PRONNODE_obj_VERBHEAD (pronoun as object of verb)	0.706	PROPNHEAD_flat.name_PROPNNODE (proper noun compound)	–0.725
ADVNODE_advmod_VERBHEAD (adverb modifying verb)	0.613	DETNODE_det_PROPNHEAD (determiner introducing proper noun)	–0.703
DETNODE_det.poss_NOUNHEAD (possessive determiner introducing noun)	0.545	NOUNNODE_nsubj.pass_VERBHEAD (noun as subject of verb in passive form)	–0.566
SCONJNODE_mark_VERBHEAD (subordinating conjunction introducing verb)	0.513	ADJNODE_amod_NOUNHEAD (adjective pre-modifying noun)	–0.434
VERBHEAD_xcomp_VERBNODE	0.356	NOUNNODE_nsubj_VERBHEAD	–0.411

(verb as open clausal complement of verb)		(noun as subject of verb in active form)	

Table 7.5. POS bigrams loading onto Italian Dimension 1

As shown in Table 7.5, bigrams that load positively onto Dimension 1 in Italian include different forms of complementation by pronouns, used as either direct objects of verbs (*l'–visto* 'saw–him' and *che–dimenticherò* 'which–forget' in Example 22) or indirect ones (*ci–appartiene* 'belongs–us' in Example 23), verbs modified by an adverb (*finalmente–firmato* 'finally–signed' in Example 24), as well as verbs in coordinate clauses (*ma–rinsaldato* 'but–strenghened' in Example 25) and subordinate ones. The latter set of bigrams comprises subordinating conjunctions introducing a verb (*quando–sembra* 'when–seems' in Example 26) and verbs used as open clausal complements of other verbs (*inizia–collaborare* 'begins–collaborate' in Example 27). The only non-verbal bigram showing positive scores on Dimension 1 is formed by a noun modified by a possessive determiner (*sua–creatura* 'his–creature' in Example 28). This bigram is special in more than one sense: Apart from being the only nominal structure in an overwhelmingly verbal dimension, it is also the only feature that is key in distinguishing constrained from unconstrained communication in 8/8 comparisons. Overuse of possessives in Italian constrained by English and German can be related to structural differences: In the latter source language/L1, possessives act as determiners, whereas in Italian they are more akin to (optional) adjectives that accompany determiners (Perridon & Sleeman, 2011). Therefore, the consistent overrepresentation of possessive determiners in constrained Italian can most likely be related to cross-linguistic influence, although our design does not allow us to rule out explicitation as an alternative explanation.

(22) Quando l'ho visto era già a terra: una scena che non dimenticherò mai.

'When I saw him, he was already on the ground: a scene which I'll never forget.'

(23) *Questo non è un termine che ci appartiene.*

'This is not a term that belongs to us.'

(24) *Il 9 dicembre la Croazia ha finalmente firmato il Trattato di adesione.*

'On December 9, Croatia has finally signed the Accession Treaty.'

(25) *Ufficialmente l'alleanza è rotta, ma il voto su Cosentino di ieri ha rinsaldato i legami.*

'Officially the alliance is broken, but yesterday's vote on Cosentino has strengthened ties.'

(26) *I miracoli ogni tanto accadono, anche quando tutto sembra destinato al peggio.* 'Miracles happen every now and then, even when everything seems destined for the worst.'

(27) *Nel 1950 ritorna a Roma, e inizia a collaborare al "Mondo".*

'In 1950 he returns to Rome, and begins to collaborate at the "Mondo" newspaper.'

(28) *Vedeva la sua creatura, la nave, affondare davanti a lui.*

'He saw his creature, the ship, sink before him.'

Negatively loading features include nouns modified by adjectives (*precedente–governo* 'previous–government' in Example 29), or by other nouns (*intervento–premier* 'speech–premier' in Example 30a) and prepositions, sometimes within the same phrase (*del–premier* 'of (the)–premier' in Example 30b). Other bigrams with negative loadings involve nouns as subjects of verbs in the active form (*sindaco–rischia* 'mayor–risks' in Example 31) or in the passive form (*operazione–notificata* 'transaction–notified' in Example 32). As in English, the nominal pole of Dimension 1 also includes structures with proper nouns: Proper nouns preceded by a determiner (*la–Polonia* 'the–Poland' and *la–Finlandia* 'the–Finland' in Example 33) and multi-word proper nouns (*Wall–Street* in Example 34).

(29) *Questo problema era ben noto anche al precedente governo.*

'The problem was well known to the previous government, too.'

(30a) *L'intervento del premier a Londra tre giorni fa ha impressionato molti grandi investitori.*

'The speech by the premier in London three days ago impressed many big investors.'

(30b) *L'intervento del premier a Londra tre giorni fa ha impressionato molti grandi investitori.*

'The speech of the premier in London three days ago impressed many big investors.'

(31) *Il sindaco del PdL rischia una parte del suo consenso per un provvedimento ambientale.*

'The PdL mayor risks losing popularity due to an environmental measure.'

(32) *A breve l'operazione dovrebbe essere notificata all'Antitrust.*

'The transaction should shortly be notified to the Antitrust Authority.'

(33) *In alcuni paesi, come la Polonia o la Finlandia, la quota sale oltre l'80 per cento.*

'In some countries, such as Poland or Finland, the share rises to over 80 percent.'

(34) *Sulla scia di Wall Street salirà, alla fine, dell'1,23%.*

'In the wake of Wall Street it will rise, in the end, by 1.23%.'

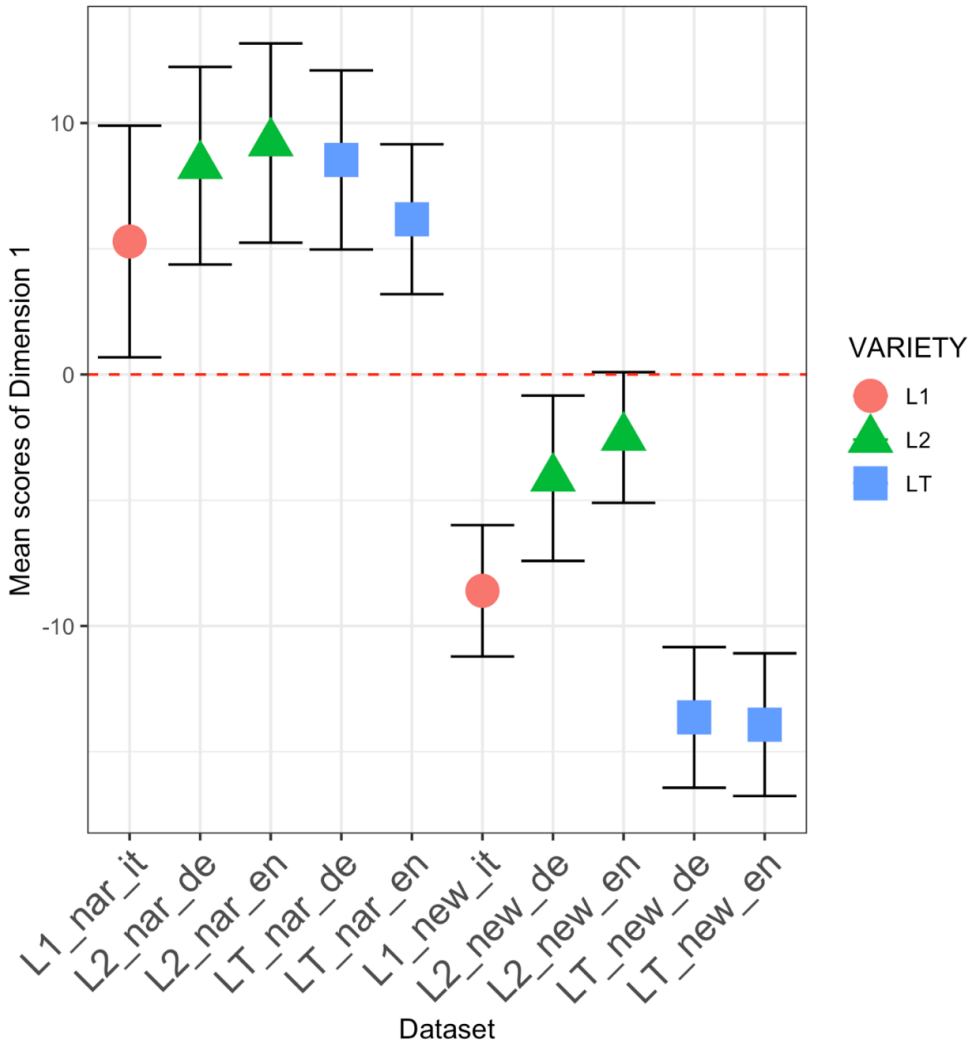

Figure 7.8. Mean scores and standard deviations for Italian Dimension 1

The distribution of mean scores (see Figure 7.8) indicates that, as is the case for English, Dimension 1 splits the subcorpora into two distinct groups based on registers. Specifically, narrative texts are characterised by a marked verbal orientation, reflected in their positive dimension scores, while the opposite holds true for news; all subcorpora in this case are characterised by negative scores and hence a nominal orientation. Within each register, differences emerge relative to how the constrained varieties position themselves with respect to L1 and to each other. In the narrative register, comparable scores are observed for all subcorpora, suggesting an overall tendency to rely on verbal elaboration in both constrained and unconstrained varieties. The same does not hold for

news, where L1 strikes a middle ground between LT and L2: Translated newswriting is characterised by a stronger preference for nominal structures than both L1 and L2 news, while L2 newswriting is at the opposite end of the spectrum, indicating a comparatively less frequent use of nouns and phrasal elaboration than in both other varieties. Again, as is the case for English, and to some extent Finnish, no clear difference emerges related to L1s/SLs of the constrained varieties: Translated and L2 texts pattern together based on their variety more than their constraining language.

4. Discussion: What constrainedness effects are common across languages?

Our first research question aimed to establish whether a set of POS bigrams could be identified that pointed to syntactic structures distinguishing both L2 and LT from L1 in all studied registers, ideally with parallels in the three languages under study. Indeed, we were able to identify, for each language, a number of such POS bigrams (19 for English, 32 for Finnish, and 29 for Italian) that distinguish the constrained varieties consistently – although not univocally – across all registers and constraining languages. Interestingly, when the relationships and the patterns of these features were explored by means of the multidimensional analysis, the centrality of register effects in constrained (and unconstrained, or less-constrained) language was the single most noticeable result. The first dimension of all the multidimensional analyses is characterised first and foremost by a clear distinction between the registers analysed in all languages, even though the analysis was conducted using features that were shown to consistently distinguish between the constrained and unconstrained varieties. In many ways, this result corroborates Kruger and Van Rooy's (2018) observation that register is always the most significant variable when explaining data variance in constrained English. Note, however, that in contrast with Kruger and Van Rooy (2018), here all the

variables were included in the multidimensional analysis based on their statistical keyness between constrained and unconstrained varieties of the same register; thus, variance between different registers did not play a role in the variable selection. The fact that register plays such a central role in the data patterning suggests, contrary to Kruger and Van Rooy (2018, p. 237), that register sensitivity lies at the very core of general constrainedness effects. The difference stems in part from the different aims of the two studies: Kruger and Van Rooy (2018) started off with variables that are known to reflect register differences, whereas we began by identifying constrainedness effects and relating them only afterwards to register variation. To us, the fact that even these variables point to register differences supports the inherent interrelatedness of constrainedness and register.

Our second research question aimed to find out if hypothesised typical features of constrained varieties would emerge bottom-up. Our results point in two directions. First, there is a lack of consistent effects related to constraining languages. In all but one comparison (i.e., Finnish academic texts translated from German and Russian), the constrained texts pattern together within the same register based on their variety (L2 texts with L2 texts, and LT texts with LT texts), better than their constraining languages. We take this to indicate that L1 background or source language plays a more marginal role compared to constrainedness or register. Second, we observe an interplay between complexity features and register-based nominal vs verbal orientation. While the registers differ between languages, the patterning is similar: Those registers that are profoundly informational in nature (academic, news, tourism) prefer a noun- and phrase-driven style, whereas the registers where personal stance and involvement are paramount (argumentative, narrative) are characterised by a verb- and clause-driven style. In many ways, this distinction reflects the first dimension of the body of earlier multidimensional analysis studies (Biber, 1988, 2014), observed also in the context of constrained language use (Kruger & Van Rooy, 2016, 2018). L2 English and L2 Italian generally prefer verbal elaboration, which is in line with earlier results (Kruger & Van Rooy, 2016; Williams, 1987). This is in part due to heavier reliance on subordination as the primary

type of structural elaboration, but also due to the use of clausal complements and, hence, more complex verbal predicates within clauses. Note, however, that even though nominal vs verbal preferences also distinguish L2 from L1 Finnish, the Finnish L2 variety is in fact characterised by a nominal style. Many of the verbal constructions that are considered in this comparison display a high degree of morphosyntactic complexity and lack an equivalent in the constraining languages (Ivaska & Bernardini, 2020; see also Eskola, 2004; Ivaska, 2014), which might explain their underuse in L2 Finnish.

As to the underlying reasons for the observed differences, this study does not lend direct support to "the constraints of real-time production" (Biber, 2014, p. 12) as an explanation. This has earlier been hypothesised to interact in constrained varieties with bilingual activation and to cause a cognitively more demanding environment seen in consequent linguistic choices (Kruger & Van Rooy, 2016, pp. 36–37). However, the fact that a similar effect is visible in the data of this study – all stemming from written varieties with no real-time constraints in the production phase – suggests that the difference is not related to real-time effects as such but rather to stylistic devices and register sensitivity. The observation of effects of bilingual activation that are agnostic to the production mode are relevant not only to research on constrained language use, but also to the broader field of language contact research.

Methodologically, the results demonstrate the value of bringing together cross-linguistically comparable tools and a language-agnostic corpus-driven methodological procedure. As also mentioned in Section 1, while ideally one would wish to have access to texts in the same registers across all languages, types of constraint and constraining language, data of this type may simply not exist, unless they are elicited in experimental or semi-experimental settings. Yet, we were able to address the same questions in parallel in three different languages despite the inherent asymmetries across the datasets, and without losing access to typological differences related to, for example, word order, use of pre- or postpositions, as well as language-specific unique items. We believe that

the possibility to ease the requirement for cross-linguistically comparable registers without losing access to register effects might be a fruitful avenue of enquiry in other types of contact language research designs, too.

Another advantage of the approach is that it is easily scalable to any language included in the UD family with openly accessible data and a relatively reliable language model. Here, we merely scratched the surface of all the possibilities offered by the approach by focusing on syntactic features. The results could – and should – be backed up by complementary analyses looking at lexical and morphological variation. What is more, extending this approach to spoken modes would be a welcome addition, to test its applicability in more diverse types of enquiry. Here, too, the maximised comparability of the UD scheme would allow for meaningful contrastive analyses.

6. Conclusion

Our study has adopted a composite corpus design and a corpus-driven methodology to search for features that distinguish constrained from unconstrained language varieties. In an attempt to control for the major variables involved (register, constraining languages, amount of editing), and at the same time produce generalisable results, we carried out the analysis in three languages, each constrained by two languages, with two registers per language, controlling for publication status. We used a wrapper built around the Ranger classifier algorithm to identify the POS bigrams that best distinguish L1 vs L2 and L1 vs translated texts and then carried out a multidimensional analysis based on these features. The ultimate aim was the bottom-up identification of a "constrainedness dimension", if one could be found in the data.

As discussed in Section 5, our results do not corroborate the existence of such a dimension. At the same time, several promising directions for further research emerged. First, we

would point to the centrality of the opposition between verbal and nominal orientation for distinguishing constrained from unconstrained varieties, different kinds of constrained varieties from each other, and different registers across varieties. Second, single (syntactically connected) POS bigrams, and different configurations of such bigrams, should be analysed in more detail, as they may point to local regularities that are obscured by the overpowering effect of register contrasts dominating the multidimensional analysis. Third, atypical register-related patterning of linguistic features could be interpreted as an effect of constrainedness in and of itself, irrespective of the nature of the linguistic features that contribute to such patterning.

It should not be forgotten that the decision to focus on syntactically dependent POS bigrams was arbitrary. The study should be replicated with differently sized structures (particularly unigrams), other types of data made available by the UD annotation scheme, as well as in different modes of language production. Morphological features might be especially relevant as complexity indices for some languages, like Finnish, that in our study appeared to behave somewhat differently from Italian and English. The interaction between morphological features and syntactic dependencies of different sizes could also be investigated. Furthermore, it would be interesting to see whether different registers of spoken interaction in typologically diverging languages would also witness the centrality of sensitivity of register effects as predictors of constrained language use. As we add layer by layer of linguistic data to an already complex research design, finding a way to combine local perspectives and general tendencies, depth and breadth, will become ever more challenging, yet inescapable as we pursue the complex case of constrained communication.

References

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 University Press.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast, 14*(1), 7–34. <https://doi.org/doi:10.1075/lic.14.1.02bib>

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Cappelle, B. (2012). English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures, 13*(2), 173–195.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL: International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177.

Egbert, J., & Staples, S. (2019). Doing multi-dimensional analysis in SPSS, SAS, and R. In T. Berber Sardinha, & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 99–114). Bloomsbury Academic.

Eskola, S. (2002). *Syntetisoivat rakenteet käännössuomessa: Suomennetun kaunokirjallisuuden ominaispiirteiden tarkastelua korpusmenetelmillä* [Synthesising structures in translated Finnish: A corpus-based analysis of the special features of Finnish literary translations]. Joensuu: University of Joensuu.

Eskola, S. (2004). Untypical frequencies in translated language: A corpus-based study on a literary corpus of translated and non-translated Finnish. In A. Mauranen, & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 83–99). John Benjamins.

Ferraresi, A., & Miličević, M. (2017). Phraseological patterns in interpreting and translation: Similar or different? In G. de Sutter, & M.-A. Lefer (Eds.), *New ways of analysing translational behaviour* (pp. 157–82). Mouton de Gruyter.

Filipović, L., & Hawkins, J. A. (2013). Multiple factors in second language acquisition: The CASP model. *Linguistics, 51*, 145–176. <https://doi.org/10.1515/ling-2013-0005>

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor, & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 225–258). Routledge.

Gentil, G. (2011). A biliteracy agenda for genre research. *Journal of Second Language Writing*, *20*(1), 6–23. <https://doi.org/10.1016/j.jslw.2010.12.006>

Granger, S., & Bestgen, Y. (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*, 229–252. <https://doi.org//10.1515/iral-2014-0011>

Grosjean, F. (2001). The bilingual's language modes. In J. Nicol (Ed.), *One mind, two languages: Bilingual language processing* (pp. 1–22). Blackwell Publishers.

Ivaska, I. (2014). Edistyneen oppijansuomen avainrakenteita: Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin [Key structures in advanced learner Finnish: Corpus approach towards structural differences between two language forms]. *Virittäjä 118*, 161–193.

Ivaska, I., & Bernardini, S. (2020). Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics 43*(1), 33–57. <https://doi.org//10.1017/S0332586520000013>

Ivaska, I., Ferraresi, A., & Bernardini, S. (2022). Syntactic properties of constrained English: A corpus-driven approach. In S. Granger, & M.-A. Lefer (Eds.), *Extending the scope of corpus-based translation studies* (pp. 133–157). Bloomsbury. <https://doi.org/10.5040/9781350143289.0013>

Iwasaki, S. (2015). A multiple-grammar model of speakers' linguistic knowledge. *Cognitive Linguistics*, *26*(2), 161–210. <https://doi.org/10.1515/cog-2014-0101>

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning*, *50*(2), 245–309. <https://doi.org/10.1111/0023-8333.00118>

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika 39*, 31–36. <https://doi.org/10.1007/BF02291575>

Kobayashi, H., & Rinnert, C. (2013). L1/L2/L3 writing development: Longitudinal case study of a Japanese multicompetent writer. *Journal of Second Language Writing*, *22*(1), 4–33. <https://doi.org/10.1016/j.jslw.2012.11.001>

Kolehmainen, L., Meriläinen, L., & Riionheimo, H. (2014). Interlingual reduction: Evidence from language contacts, translation and second language acquisition. In H. Paulasto, L. Meriläinen, H. Riionheimo, & M. Kok (Eds.), *Language contacts at the crossroads of disciplines* (pp. 3–32). Cambridge Scholars Publishing.

Kotze, H. (2020). Converging *what* and *how* to find out *why*. In L. Vandevoorde, J. Daems, & B. Defrancq (Eds.), *New empirical perspectives on translation and interpreting* (pp. 333–371). Routledge.

Kranich, S. (2014). Translations as a locus of languge contact. In J. House (Ed.), *Translation: A multidisciplinary approach* (pp. 96–115). Palgrave Macmillan.

Kruger, H. (2017). The effects of editorial intervention: Implications for studies of the features of translated language. In G. de Sutter, M. -A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (pp. 113–155). De Gruyter.

Kruger, H., & Van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide, 37*(1), 26–57. <https://doi.org/10.1075/eww.37.1.02kru>

Kruger, H., & Van Rooy, B. (2018). Register variation in written contact varieties of English. *English World-Wide, 39*(2): 214–242. <https://doi.org/doi:10.1075/eww.00011.kru>

Kursa, M., & Rudnicki, W. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, Articles, 36*, 1–13. <https://doi.org/10.18637/jss.v036.i11>

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, *13*(1), 99–121. <https://doi.org/10.1556/Acr.13.2012.1.6>

Leech, G. (2006). New resources, or just better old ones? The holy grail of representativeness. In N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 133–149). Language and Computers 59. Brill.

Lefer, M.-A., & Vogeleer, S. (2013). Interference and normalization in genre-controlled multilingual corpora: Introduction. *Belgian Journal of Linguistics*, *27*(1), 1–21. <https://doi.org/doi:10.1075/bjl.27.01lef>

Marco, J. (2009). Normalisation and the translation of phraseology in the COVALT corpus. *Meta*, *54*(4), 842–856. <https://doi.org/10.7202/038907ar>

Mauranen, A., & Kujamäki, P. (2004). *Translation universals: Do they exist?* John Benjamins.

Mein, E. (2012). Biliteracy in context: The use of L1/L2 genre knowledge in graduate studies. *International Journal of Bilingual Education and Bilingualism*, *15*(6), 653–667. <https://doi.org/10.1080/13670050.2012.699946>

Nawal, A. F. (2018). Cognitive load theory in the context of second language academic writing. *Higher Education Pedagogies, 3,* 385–402. <https://doi.org/10.1080/23752696.2018.1513812>

Nivre, J., Marneffe, M. -C. de, Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., et al. (2016). Universal Dependencies v1: A multilingual treebank collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

O'Brien, S. (2006). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, *14*(3), 185–205. <https://doi.org/10.1080/09076760708669037>

Perridon, H., & Sleeman, P. (2011). The noun phrase in Germanic and Romance: Common developments and differences. *Linguistik Aktuell, 171*, 1–22.

R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Rabinovich, E., Nisioi, S., Ordan, N., & Wintner, S. (2016). On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1870–1881). Association for Computational Linguistics.

Revelle, W. (2018). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. <https://CRAN.R-project.org/package=psych>

Romaine, S. (2001). Multilingualism. In M. Aronoff & J. Rees-Miller (eds.), *The Handbook of Linguistics* (pp. 541–556). Blackwell.

Sjöholm, K. (1995). *The influence of cross-linguistic, semantic and input factors on the acquisition of English phrasal verbs*. Åbo Akademi University Press.

Steiner, E. (2005). Explicitation, its lexicogrammatical realisation, and its determining (independent) variables: Towards an empirical and corpus-based methodology. *SPRIKreports, 36*, 1–42.

Steiner, E. (2008). Empirical studies of translations as a mode of language contact: "Explicitness" of lexicogrammatical encoding as a relevant dimension. In P. Siemund, & N. Kintana (Eds.), *Language contact and contact languages* (pp. 317–341). John Benjamins. <https://doi.org/10.1075/hsm.7.18ste>

Straka, M., & Straková, J. (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 (pp. 88–99). Association for Computational Linguistics. <http://www.aclweb.org/anthology/K/K17/K17-3009.pdf>

Szymor, N. (2018). Translation: Universals or cognition? A usage-based perspective. *Target*, *30*(1), 53–86. <https://doi.org/10.1075/target.15155.szy>

Teich, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Text, Translation, Computational Processing. De Gruyter.

Tirkkonen-Condit, S. (2004). Unique items – over- or under-represented in translated language? In A. Mauranen, & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 177–184). John Benjamins.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Upton, T. A., & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, *20*(4), 313–329. <https://doi.org/10.1016/S0889-4906(00)00022-3>

Williams, J. (1987). Non-native varieties of English: A special case of language acquisition. *English World-Wide, 8*(2), 161–199. <https://doi.org/10.1075/eww.8.2.02wil>

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*, 1–17. <https://doi.org/10.18637/jss.v077.i01>

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511519772>

Appendix: Corpora used

CTF = Corpus of Translated Finnish. Mauranen, A. (2000). Strange strings in translated language: A study on corpora. In M. Olohan (Ed.), *Intercultural faultlines: Research models in translation studies* (pp. 119–141). St Jerome.

EFCamDat = EF-Cambridge Open Language Database. Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics, 23*(1), 28–54.

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

ICLFI = International Corpus of Learner Finnish. Jantunen, J. (2011). Kansainvälinen oppijansuomen korpus (ICLFI): Typologia, taustamuuttujat ja annotointi [International Learner Finnish Corpus (ICLFI): Tpology, background variables and annotation]. *Lähivõrdlusi. Lähivertailuja* [Close comparisons]*, 21*, 86–105.

InterCorp. Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics, 17*(3), 411–427.

LAS1 = The Corpus of Academic Finnish. (2018). The Department of Finnish and Finno-Ugric Languages, University of Turku.

LAS2 = The Corpus of Advanced Learner Finnish. Ivaska, I. (2014). The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples – Journal of Applied Language Studies, 8*(3), 21–38.

LOCNESS = Louvain Corpus of Native English Essays. Centre for English Corpus Linguistics (CECL), Université Catholique de Louvain.

PEC = Perugia Corpus. Spina, S. (2014). Il Perugia Corpus: Una risorsa di riferimento per l'italiano.

Composizione, annotazione e valutazione [The Perugia Corpus: A reference resource for Italian. Composition, annotation and evaluation]. In R. Basili, A. Lenci, & B. Magnini (Eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it, vol. 1* (pp. 354–359). Pisa University Press.

VALICO. Allora, A., Colombo, S., & Marello, C. (2011). I corpora VALICO e VINCA: Stranieri e italiani alle prese con le stesse attività scritte [The VALICO and VINCA corpora: Foreigners and Italians struggling with the same written tasks]. In N. Maraschio, & D. de Martino (Eds.), *La Piazza delle lingue: L'italiano degli altri. Firenze, 27–31 maggio 2010* [The square of languages: The Italian of others. Florence, 27–31 May 2010] (pp. 49–61). Accademia della Crusca.