# Constrained Language Use in Finnish: A corpus-driven approach

## 1. INTRODUCTION[1]

Second language acquisition (SLA) and translation studies (TS) share a common interest in investigating language use where several linguistic systems are at play simultaneously – in one way or another. Whilst the two disciplines have operated somewhat independently in the past, growing interest in similarities between the two has recently led to the emergence of a new line of research described as *constrained language use* (Lanstyák & Heltai 2012; Kolehmainen, Meriläinen & Riionheimo 2014; Kruger & van Rooy 2016; Rabinovich et al. 2016; Kruger & van Rooy 2018)[2] that explores whether and to what degree different types of constrained language use share common characteristics.

Research designs in both SLA and TS tend to follow two seemingly opposed approaches. On the one hand the study of crosslinguistic influence (CLI) is a central research focus in both disciplines (Jarvis 2000 for SLA; Toury 2012: 310–315 for TS). On the other, both disciplines are interested in phenomena that function similarly irrespective of the languages or language-pairs involved (Granger 2015 for SLA; Mauranen & Kujamäki 2004 for TS). These approaches operate on different levels of abstraction – the potential general features will inevitably have language-specific realizations, which make a fully comparable operationalization between languages challenging (House 2008; Becher 2010). The problem is amplified by the fact that in the case of both SLA and TS, the body of research and the subsequent theorising have been based on few languages and language pairs, predominantly including English. To our knowledge, this is the case also in studies that explicitly explore constrained language use, although the meta-analysis by Kolehmainen et al. (2014) includes studies on various languages.

The focus on English limits the generalizability of the results gained thus far (for a discussion, see Volansky, Ordan & Wintner 2013). Given the extraordinary situation of English as a de facto lingua franca and the typological diversity of the languages of the world, further evidence on constrained language use is needed from languages of different typological and geopolitical status. The present paper addresses this research gap by exploring the typicalities of constrained language use in Finnish – a language that diverges from English considerably both typologically and geopolitically. We report a quantitative bottom-up study on similarities between Finnish as a second language (F2) and translated Finnish (FT). More specifically, we contrast F2 and FT with their non-constrained counterpart – non-translated Finnish as a first language (F1) – to see whether and to what extent the features that distinguish F2 and FT from F1 are shared by the constrained varieties. Our specific research questions are: 1) Which linguistic phenomena consistently distinguish both F2 and FT from F1? 2) Are there variety-, language-pair-, or register-specific differences? 3) How do the detected features of constrained Finnish line up with earlier suggestions of general phenomena?

Our data come from various existing corpora, which we have complemented with data collected ad hoc to increase data comparability. The data represent two registers (academic and narrative texts) and two first/source languages (German and Russian). We implement a two-phase methodological procedure to detect and analyze differences across the compared varieties. First, we conduct a keyness analysis (for the concept, see Gabrielatos 2018), to detect syntactically defined part-of-speech bigrams whose frequency consistently distinguishes constrained and non-constrained Finnish across the majority of data subsets. Then, we carry out a Multi-Dimensional Analysis (e.g. Biber 1988; Biber 2014) based on these key features, to interpret their linguistic and textual distributional patternings in the light of constrained language use in Finnish.

The paper is structured as follows: Section 2 discusses the core theoretical constructs and earlier research and Section 3 introduces the data and method used. Section 4 reports on our results, linking them to earlier findings and Section 5 discusses their theoretical impact, offering concluding remarks on future work.

## 2. VARIATION IN CONSTRAINED LANGUAGE USE

### *2.1 Varieties, crosslinguistic influences and registers*

In order to reliably relate any given linguistic phenomenon to constrained language use, there are a minimum of three factors to take into account. First, two or more constrained varieties must diverge from the non-constrained variety. Second, this divergence should be observable across two or more language-pairs, so as to tease apart the general tendencies from CLI. Third, the study should be conducted across multiple registers, to control for the potential interaction between the observed divergence and register-specific particularities (see also Kruger & van Rooy 2018).

Lanstyák and Heltai (2012) point out that all communication is somehow constrained but that the research on constrained communication refers to varieties where constraining factors "play a greater than average role" (Lanstyák & Heltai 2012: 100). The underlying reason for these similarities has been suggested to stem from similar cognitive and social environments between such varieties (Kruger & van Rooy 2016: 27). The situation resembles bilingual language activation in bilingualism studies (Grosjean 2001). Constrained language use can be seen to increase the cognitive load which can lead to increasing explicitness (Rohdenburg 1996). In SLA, this mechanism has been suggested to take place in terms of interacting learning principles (Filipović & Hawkins 2013). Finally, differences between constrained and unconstrained varieties may also stem from strategic choices: as Lefer &

Vogeleer (2013: 10) point out, “it can be hypothesized that translators stop searching for a better solution when they believe their current solution to be sufficiently relevant”.

In subsequent studies, the focus has mostly been on communication constrained by (at least) two languages. In a meta-analysis of constrained language use (in their terms, *interlingual reduction*), Kolehmainen et al. (2014) compare studies on translated, contact, and second language. Kruger and van Rooy (2016) contrast translated English with indigenized varieties, whereas Rabinovich et al. (2016) look at similarities between translated and second languages. To our knowledge, the most fine-grained take on the constrained varieties can be found in Kruger and van Rooy (2018), as they do not only distinguish native, L2 and translated varieties, but also the amount of language contact in both native and L2 varieties. In the present study, the constrained varieties are Finnish written by L2 speakers and texts translated into Finnish from another language by professional translators (typically L1 speakers of Finnish). The reference variety consists of non-translated texts written by L1 speakers of Finnish.

CLI is among the most widely studied phenomena in both L2 and TS contexts. At the core of the inquiry is the identification of linguistic phenomena that can be attributed to specific first/source languages. Following Jarvis (2000; 2010), reliable support for CLI should be based on the following aspects: 1) congruent linguistic behavior in L2 by speakers of the same L1; 2) diverging linguistic behavior in L2 by speakers of different L1s; 3) congruence between the L2 behavior of speakers of the same L1 and the linguistic system of that language; 4) divergence between the linguistic systems of the L1s represented.

Despite the centrality of CLI within L2 research and TS, it has received only limited attention in research on constrained language use, and often the data used do not even allow for distinguishing general and language-pair specific phenomena. Perhaps the most notable exception is Rabinovich et al. (2016), who approach CLI from a computational perspective:

they use texts translated into English from Germanic and Romance language families and build separate language models for each family. They then use these models to predict whether L2-English texts were written by L1 speakers of a Germanic or Romance language. The results show that the model based on data from the same language family as the author's L1 indeed outperforms the model based on the other language family – effectively corroborating the similarity of CLI in L2 and translated English.

Different registers of one language may vary drastically both qualitatively and quantitatively in their grammatical properties (for the concept, see Biber & Conrad 2009; for examples in different languages, see Biber 2014). The notion of register-specific linguistic systems is theorized also to apply to the level of individual language users in Iwasaki (2015), suggesting that the grammars of different registers are learned partially in isolation – in a usage-based fashion – and inter-connected only at a later stage. This suggestion is supported by Ivaska (2015) in the observation that academic F2 users adjust their tense use to that of F1 academic texts over the course of 16 months – together with their more extensive exposure to that register. Also Szymor's (2018) results support the usage-based nature of register-specific individual grammars, showing that the use of aspect in modal contexts in Polish translations of legal texts reflect distributions found in non-translated Polish in general rather than those of non-translated Polish legal texts. This can be related to the translators' lesser exposure to that particular register and the ensuing reliance on their knowledge of other registers.

Finally, registers and CLI do not work in isolation. Although CLI is typically approached from a structural perspective – lexical, morphological or syntactic –registers are to some degree also specific to different linguacultures. As Lefer and Vogeleer (2013: 15) put it, "some genre conventions are common to different cultures, while some others are culture-specific […]. As a result, in translation, localization […] and non-native writing, interference and normalization can be traced not only at the micro-level of linguistic features but also at the macro-level of genre conventions."

### *2.2 Contrastive quantitative research on constrained varieties in Finnish*

We are not aware of any studies that explicitly address constrained language use in Finnish (although Jantunen 2008 draws from TS when hypothesizing possible L2 universal tendencies). Furthermore, systematic studies on register variation in either F2 or FT are rare. In what follows, we review earlier research on F2 and FT relevant for the present study[3]. Given the methodological foundations of our work, we limit ourselves to quantitative approaches. Furthermore, as our study is corpus-driven, we focus on bottom-up observations of potentially interesting phenomena, and then relate our findings to earlier studies. Finally, with a view to attributing any given observed linguistic phenomenon to a certain variety, we compare the said variety and other varieties (Szmrecsanyi 2017). As a consequence, we only consider studies with a contrastive component (although earlier results discussed in Section 4 are not necessarily contrastive in nature).

Most quantitative contrastive work on F2 has been conducted on two corpora: the International Corpus of Learner Finnish (ICLFI, Jantunen 2011a) or the Corpus of Advanced Learner Finnish (Ivaska 2014a). Using an automatically annotated version of ICLFI, Jantunen (2011b) detected grammatical key features that distinguish F2 from F1 at different proficiency levels. While groundbreaking as a research design, the results are very preliminary in nature with a number of methodological concerns related to CLI and register effects, which makes it difficult to attribute the findings to the F2 variable. Ivaska (2014b) detected four types of differences between F2 and F1: 1) F2 texts make less use of morphosyntactically complex constructions (see also Ivaska, Reunanen & Siitonen 2016); 2) F2 texts express modal possibility less frequently (see also Ivaska 2014c); 3) F2 texts are more narrative, as reflected in their higher frequency of the simple past tense (see also Ivaska 2015); 4) certain register-specific, lexically limited constructions are less frequent in F2 texts. The results are methodologically reliable, in that all data belong to the same register, although this has the downside of limiting their generalizability. Potential CLI could not be evaluated, as the speakers L1 backgrounds were not controlled for.

As for possible CLI in F2, Spoelman (2013) studied the use of partitive case by Dutch, Estonian and German F2 users and showed clear differences between Estonian and the two other learner groups – likely due to typological similarities between Finnish and Estonian. Ivaska and Siitonen (2017a) detected morphological features that distinguished Chinese, Estonian, German, Polish, Russian and Swedish F2 users in a data-driven manner. They showed that Swedish F2 users express pronominal subjects overtly more often than the other learner groups and linked it to the fact that Swedish is the only L1 included with mandatory pronominal subject-marking and without conjugational subject marking. Similarly, Estonian F2 users use more numerals, possibly due to lexical and phrasal similarities between Finnish and Estonian.

Most quantitative contrastive approaches to FT use the Corpus of Translated Finnish (Mauranen 2000), including a number of studies on unique items – features that are typical for the target language but not for the source language. Tirkkonen-Condit (2004; 2005) identified certain language-specific constructions of Finnish –verbs expressing sufficiency and the clitic pragmatic particles -*kin* and -*hAn* – to be less common in FT than in F1. Similarly, Kujamäki (2004) reported on a Finnish-English/German-Finnish back-translation experiment and concluded that unique items are consistently used less in the back-translations than in the original F1. Jantunen and Eskola (2002) (see also Eskola 2004; Jantunen 2004) showed that unique items affect FT at various levels: the non-finite referative construction (e.g. *Tiedän hänen* ***tulleen*** 'I know (that) (s)he has **come'**) is consistently more frequent in F1 than in FT from English or Russian. On the other hand, TL constructions with a formal equivalent in the SL tend to be used more often in FT than in F1. In Jantunen and Eskola's case, such is the case of a non-finite construction expressing motivation for a certain behavior (e.g. *Kiirehdin* ***ehtiäkseni*** *junaan* 'I hurry (in order) to **catch** the train'), and the degree modifier *hyvin* 'very'. Finally, Mauranen and Tiittula (2005) compared the use of subjectless impersonal constructions (e.g. *Jos tupakoi…* 'if [one] smokes…') in FT from English and German and F1 and concluded that the construction – which lacks a similar counterpart in Germanic languages – is more common in F1, while FT relies on solutions that are more aligned with those in the original German texts. According to Mauranen and Tiittula (ibid), there is also a related explicitation tendency: FT contains more first person nominative pronouns (*minä*) than F1. This is probably due to constructional differences: both English and German have obligatorily overt subject marking, whereas in Finnish subject marking is obligatory only in the third person.

CLI in FT has also been studied outside the unique item hypothesis. Mauranen (2004) compared word frequency data to measure differences between FT and F1 and between SL-specific subcorpora and a subcorpus with a range of SLs. She (2004: 78) concluded that "translations resemble each other more than original target language texts, but a clear source language effect is also discernible. This implies that transfer is one of the causes behind the special features of translated language." Mauranen's observation underlines the often-disregarded fact that general features of translated language and CLI are two facets of the same phenomenon, differing in their level of abstraction. L. Ivaska (2019), in turn followed the line of research unearthed by Baroni and Bernardini (2006) and Koppel and Ordan (2011) – with the goal to distinguish FT from F1 and to tease apart the different source languages of FT by means of core vocabulary frequencies and machine learning techniques. She was able to accurately distinguish FT and F1 texts; translations from different source languages, while diverging from each other in a systematic manner, were more difficult to categorise reliably. The results corroborate Mauranen's point on the intertwined nature of CLI and the more general tendencies of translated language.

## 3. DATA AND METHODS[4]

### *3.1 Data: Composite corpus of existing and ad hoc materials*

The data in this study consist of 12 subcorpora, and include Finnish texts constrained by second language use and by translation in two registers (academic and narrative), together with comparable non-constrained texts. The first languages included are German and Russian. We attempted to maximize the replicability of the study results/design by choosing registers and first/source languages included in existing corpora. The existing corpora used are: Corpus of Translated Finnish (CTF, including both F1 and FT), Contrastive Corpus of Finnish and German (FinDe), International Corpus of Learner Finnish (ICLFI), InterCorp (including both F1 and FT), Corpus of Advanced Learner Finnish (LAS2, including both F1 and F2), Corpus of Academic Finnish (LAS1). In order to ensure the comparability of each constrained subcorpus and the respective unconstrained subcorpus, we use two unconstrained subcorpora for both registers – one of unpublished texts to match F2 data and another of published texts to match FT data (on the effects of editorial intervention, see Kruger 2017). To this end, we collected an additional ad hoc corpus of unpublished F1 narrative texts. Finally, in the context of F2, the proficiency level also plays a role, therefore we included only texts representing advanced proficiency level. That is, all the texts from ICLFI have been evaluated to reflect C level in the Common European Framework of Reference. In LAS2, a minimum of 2 texts from each informant were evaluated, and only texts by informants where both evaluations reflect C level were included.

After stripping off all legacy annotation, we annotated the data according to the Universal Dependencies scheme (Bohnet et al. 2013) using the Turku neural parser (Kanerva et al. 2018). To minimize the effects of other artifacts (e.g. text length, topic, authorship), we then shuffled the sentences of each subcorpus and reconstructed text blocks of 50 sentences (Rabinovich et al. 2016). This step also ensures that the subcorpora are directly comparable while still maintaining the general variation inherent in them. Table 1 sums up the data provenance and size for the 12 subcorpora, 11 of which come from pre-existing resources and 1 is collected ad hoc.

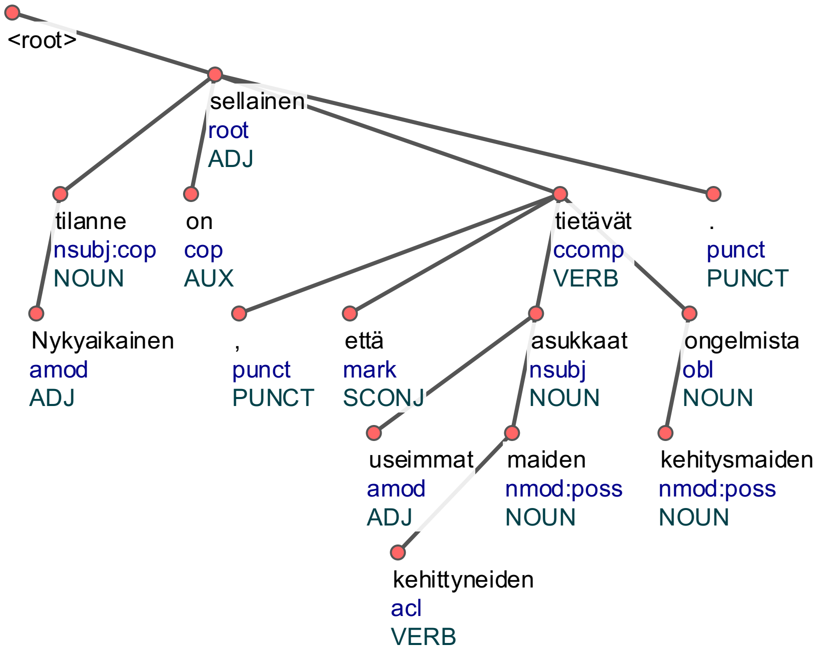
| REG | F1-published | FT-de | FT-ru | F1-unpublished | F2-de | F2-ru |
|---|---|---|---|---|---|---|
| ACAD | CTF & LAS1<br>4 733 blocks<br>1 180 564 w | CTF & FinDe<br>470 blocks<br>114 751 w | CTF<br>164 blocks<br>51 285 w | LAS1 & LAS2<br>1398 blocks<br>392 588 w | ICLFI & LAS2<br>152 blocks<br>32 670 w | ICLFI & LAS2<br>806 blocks<br>166 905 w |
| NARR | CTF & InterCorp<br>11 622 blocks<br>2 769 603 w | CTF & InterCorp<br>1 400 blocks<br>334 448 w | CTF<br>4 259 blocks<br>650 880 w | *ad hoc*<br>214 blocks<br>32 748 w | ICLFI<br>64 blocks<br>10 015 w | ICLFI<br>121 blocks<br>20 085 w |
| **ALL** | 3 950 167 w | 449 199 w | 702 165 w | 425 336 w | 64 797 w | 263 424 w |

**Table 1. Data provenance and size information.**

### *3.2 Keyness analysis*

The first step in the analysis aims to detect consistent quantitative differences between the varieties that emerge from the data bottom up. First, we extracted the normalized frequencies of all syntactically defined part-of-speech (POS) bigrams from each text block. Each bigram provides information on POS, syntactic relationship, as well as the constituent order and hierarchy. Figure 1 exemplifies the feature set with the following POS-bigrams: ADJNODE_amod_NOUNHEAD (*nykyaikainen–tilanne* ‘contemporary–situation’), NOUNNODE_nsubj:cop_ADJHEAD (*tilanne–sellainen* ‘situation–such’), AUXNODE_cop_ADJHEAD (*on–sellainen* ‘is–such’), ADJHEAD_ccomp_VERBNODE (*sellainen–tietävät* ‘such–know’), and so on. In this first phase, we include the 1,000 most frequent bigrams.

**Figure 1. Tree visualization of POS bigrams.**



Next, we detected consistent differences between constrained and non-constrained texts. Each constrained data subset and its non-constrained counterpart were contrasted to find the bigrams whose frequency differences contribute to distinguishing the two. The resulting sets of bigrams were then compared to single out those bigrams that best correlate with the constrained vs. non-constrained distinction. This two-phase approach has two advantages related to the opposite pulls of comparability and generalizability (for a discussion, see Leech 2006): each pairwise comparison can be conducted with maximally comparable data, and each pairwise comparison is of equal importance in the overall comparison.

In the pairwise comparisons, we use Boruta feature selection (Kursa & Rudnicki 2010) as a statistical technique to find the bigrams contributing to the difference. Boruta makes use of Random Forests (Breiman 2001), a machine learning algorithm for automatic data classification. Boruta adds randomness to the data by creating a duplicate variable for each actual variable (here, bigram) and randomly permutating its values (here, normalized frequencies). Then, it creates a decision-tree-based forest model to find the variables that are best at the classification task at hand (here, distinguishing between constrained and non-constrained texts). It then compares the actual variables' performance to the duplicate variables and deems as important those variables that consistently outperform the permutated duplicates. The procedure is repeated iteratively – leaving out in each run variables that are clearly less significant than the duplicates – resulting in a list of bigrams that consistently contribute to the classification. After obtaining such lists from each pairwise comparison, we compare them to find the bigrams that are included in over half of the lists and can thus be considered key features.

32 of the 1,000 syntactic POS-bigrams analyzed were considered as key in over half of the pairwise comparisons (Appendix 1). While these comparisons make it possible to capture the bigrams that consistently contribute to distinguishing constrained from non-constrained language, they do not tell us anything about the underlying functional reasons. As pointed out by Gries (2019), applying tree-based methods without addressing the potential interactions between variables may indeed lead to suboptimal or even misleading interpretations. Thus, in order to analyse the variance and interactions between the detected key bigrams, we used them as variables in the subsequent analysis.

### *3.3 Multi-Dimensional Analysis of consistent key features*

To understand how the 32 bigrams resulting from the keyness analysis behave and interact, we conducted a Multi-Dimensional Analysis (MDA, e.g. Biber 1988; Berber Sardinha & Veirano Pinto 2014) using them as variables. We then conducted a factor analysis based on these key features, interpreted the resulting dimensions functionally, computed dimension scores and compared them across datasets to interpret the relationships between the dimensions and the type of constraint, register, and first/source language.

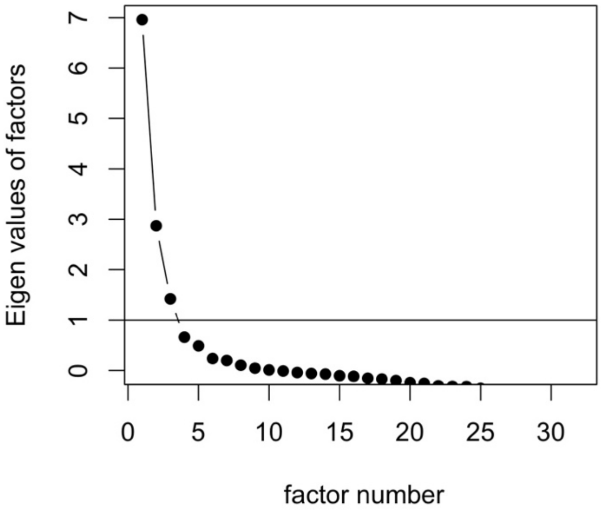
For the factor analysis, we used the functions found in the R package *psych* (Revelle 2018). Factor analysis provides a calculation of eigenvalues and a corresponding scree plot, which represents a standardized measure of the proportion of variance explained by a given factor. Having chosen the number of factors, we run the final analysis to obtain the factor loadings of each variable for each factor. Factor loadings are weights (ranging from –1 to 1) that reflect how much a given variable contributes to a given factor. As in many MDAs, we use a threshold level of 0.35 (absolute values) for factor loadings in order for a variable to be included in the functional interpretations and dimension score computations. For these computations, we followed the procedure described in Biber (1989: 11) and standardized the variables' normalized frequencies in terms of how many standard deviations the value is above or below the overall mean for that variable. We then computed the dimension scores for each dimension by adding (for each text block) the standarized values of the variables with positive loadings and subtracting those with negative loadings.

## 4. RESULTS

### *4.1 Choosing the number of dimensions*

The balanced, maximum-sized random sample used in the study consists of 64 text blocks of 12 subsets each, totaling 768 blocks. Before the actual factor analysis, we measured the factorability of the 32 included variables by means of the KMO measure of sampling adequacy (Kaiser 1974). As this resulted in an overall value of 0.9 – considered by Kaiser (1974: 35) as "marvelous" – we felt confident proceeding with our analysis. The scree plot of the factor analysis (figure 2) shows three factors with eigenvalues above 1, clearly distinct from the subsequent factors. Our final model is thus a rotated three-factor solution that allows for intercorrelated factors (with all factor correlations 0.37 or lower).

**figure 2. Scree plot of eigenvalues for dimensions.**



### *4.2 Dimension 1: Clausal complexity and dialogue*

The bigrams that load onto dimension 1 (see table 2) portray various clause-level constructions that reflect two diverging phenomena: clausal elaboration/verbal complexity, and linguistic means of narration[5]. Complexity features include (potentially non-finite) adverbial clauses that act as modifiers (1), and various non-finite clausal complements (2 and 3). The negatively loading bigrams can be seen to mirror this as they include different copular clauses, phrasal coordination and genitive modification, all of which reflect nominal complexity. In terms of the complexity taxonomy of Bulté and Housen (2012), these features can be said reflect linguistic structure complexity – more specifically an interaction between formal and functional complexity. As examples (1)–(3) show, these features portray structural complexity in terms of multiple morphological elements that interact with sentence, clause, and phrasal complexity (Bulté & Housen 2012: 27).

(1) *Auron* ***asett-i*** *vasemman kätenä hartioilleni* ***katsell-e-ssa-an*** *upeaa maisemaa*[6].
[…] VERBHEAD_ […] advcl_VERBNODE
[…] place-PRET.3SG […] watch-INF2-INE-PX3
’Auron **placed** his/her left arm on my shoulders **while watching** the gorgeous landscape.

(2) *syvän sininen väri* ***kerto-i*** *auringon* ***porotta-va-n*** *keskipäivää.*
[…] VERBHEAD_ […] xcomp.ds_VERBNODE […]
[…] tell-PRET.3SG […] shine-PTCP1-GEN […]
’the deep blue color **indicated** that the sun was **shining** the noon.

(3) *Hän* ***ryhty-i*** ***viheltele-mä-än*** *viattomasti.*
[…] VERBHEAD_ xcomp_VERBNODE […]
[…] begin-PRET.3SG whistle-INF3-ILL […]
‘(S)he **began to whistle** innocently.

Paratactic coordination and proper noun subjects (4) are used to switch between direct speech and narration, and pronominal subjects (5) and objects (6) are typical cohesive devices used in linear narration.

(4) ***Herätä*** *sitten minut,* ***Aragorn*** ***sano-i****.*
VERBHEAD_ […] parataxis_VERBNODE
[…] PROPNODE_nsubj _VERBHEAD
wake.IMP.2SG […] Aragorn say-PRET.3SG
‘Wake me up then, **Aragorn said**.’

(5) ***Hän*** ***pakott-i*** *itsensä kysymään.*
PRONNODE_nsubj _VERBHEAD […]
(s)he.NOM force-PRET.3SG […]
**‘(S)he forced** him/herself to ask.

(6) *Se ei ollut suinkaan* ***lannista-nut*** ***hän-tä****.*
[…] VERBHEAD_ obj_PRONNODE
[…] discourage-PTCP2 (s)he-PTV
‘It most certainly hadn’t **discouraged him/her**’

| Positive features | | Negative features | |
|---|---|---|---|
| PROPNNODE_nsubj_VERBHEAD | 0.584 | NOUNNODE_nsubj:cop_ADJHEAD | -0.564 |
| PRONNODE_nsubj_VERBHEAD | (0.403) | CCONJNODE_cc_NOUNHEAD | (-0.488) |
| VERBHEAD_obj_PRONNODE | 0.466 | NOUNNODE_nsubj:cop_NOUNHEAD | -0.643 |
| VERBHEAD_advcl_VERBNODE | 0.612 | NOUNNODE_nmod:poss_NOUNHEAD | (-0.356) |
| VERBHEAD_xcomp:ds_VERBNODE | 0.382 | NOUNHEAD_conj_NOUNNODE | -0.518 |
| VERBHEAD_obl_NOUNNODE | 0.382 | | |
| VERBHEAD_parataxis_VERBNODE | 0.430 | | |
| VERBHEAD_xcomp_VERBNODE | 0.634 | | |

**Table 2. POS bigrams loading onto dimension 1.**

Cross-corpus comparison of the dimension scores reveals an interesting patterning. As seen in the text chunk groupings in figures 3 (register), 4 (variety), and 5 (first/source language), the L1-narrative corpora have clearly positive mean scores (L1-narrative_pub: 5.7; L1-narrative_nonpub: 11.7), meaning that the above-described features are typical for the said corpora. On the other hand, the F1-academic corpora (figures 3 and 4) have clearly negative mean scores (L1-academic_pub: –5.3; L1-academic_nonpub: –3.6) meaning that the features are less typical. In other words, in L1-Finnish, narration elements and switches between direct and indirect speech are typical for narrative texts and atypical for academic texts. The mean score differences are clearly smaller in LT-corpora (LT-de_narrative: 4.8; LT-ru_narrative: 4.2 VS. LT-de_academic: –3.0; LT-ru_academic: 3.0) and missing from the L2-corpora (L2-de_narrative: –2.4; L2-ru_narrative: –2.8 VS. L2-de_academic: –6.0; L2-ru_academic: –6.3).

Our findings thus partly corroborate those of Kruger and van Rooy (2018: 235–236) who show that, in constrained English, reported speech often distinguishes non-native and native varieties. In a broader sense, this may in turn be related to the first dimension of most multidimensional analyses – involved versus informational production (e.g. Biber 1988; Biber 2014). The difference is also related to non-finite verbal constructions, some of which are unique to Finnish, and were studied under the unique item hypothesis (Eskola 2004). Besides their uniqueness, many of them are also morphosyntactically complex, and shown to distinguish even advanced L2- from L1-Finnish (Ivaska 2014b; Ivaska & Siitonen 2017b).

**Figure 3**. **Scatterplot of dimensions 1 and 2 in relation to register.**

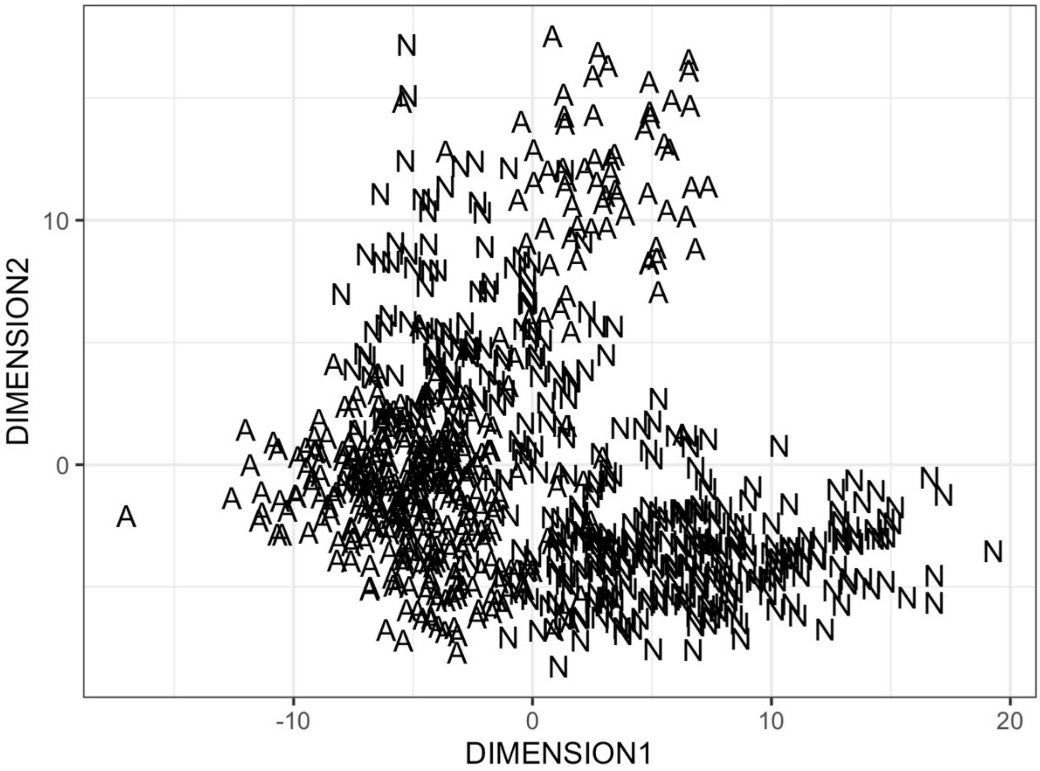


**Figure 4**. **Scatterplot of dimensions 1 and 2 in relation to variety.**

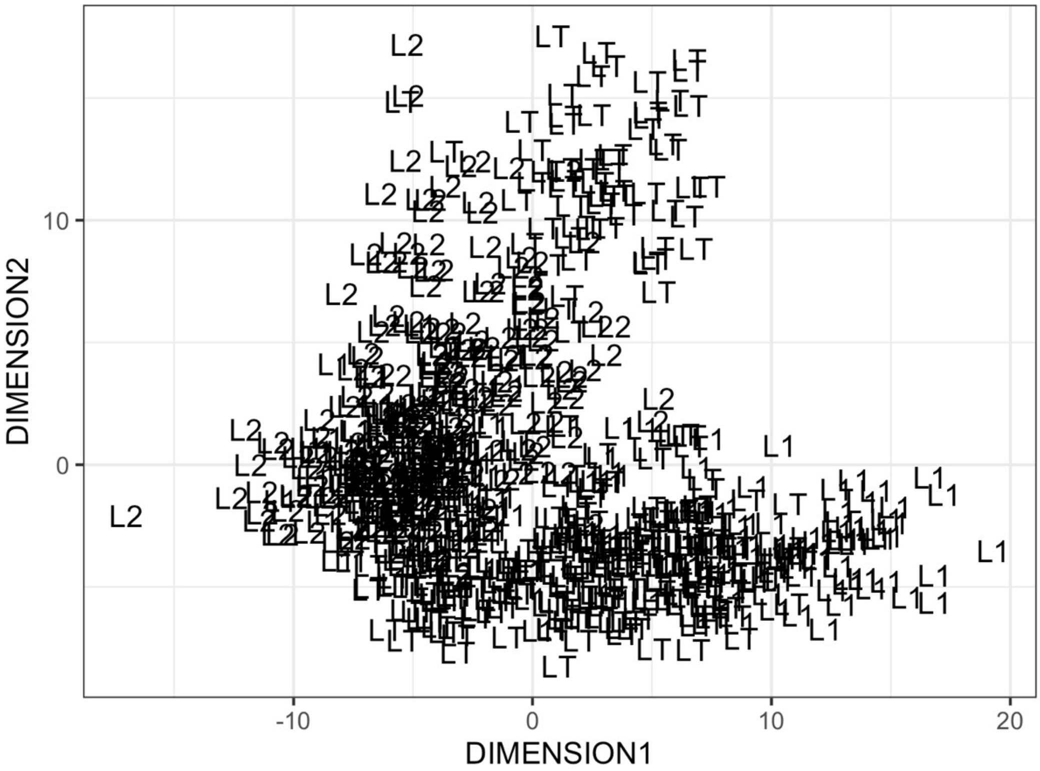
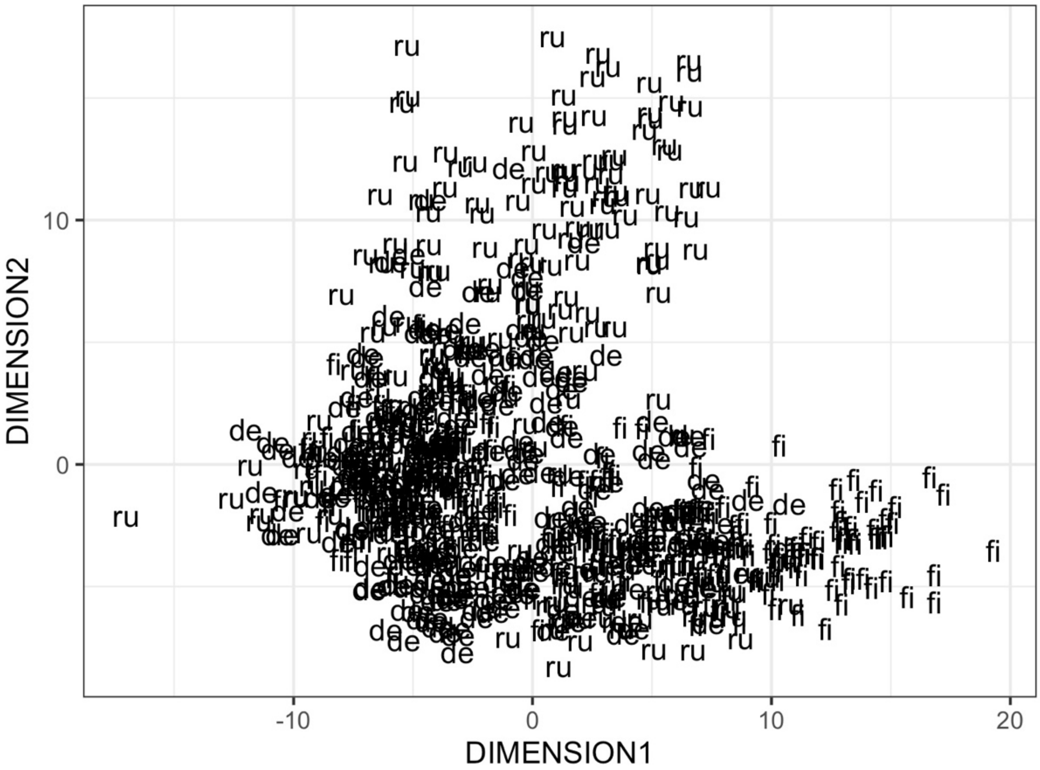


**Figure 5**. **Scatterplot of dimensions 1 and 2 in relation to first/source language.**



***4.3 Dimension 2: Proper name explicitation***

As seen in table 3, most of the bigrams that load positively onto dimension 2 include proper nouns in various syntactic roles, such as pre- and post-verbal adjuncts (7), phrasal coordination (8), parts of compounds (9), pre-nominal possessive modifiers (10) and post-nominal modifiers (11). This dimension is thus predominantly characterized by the use of proper nouns.

(7) ***Liettua-ssa*** *vastuu* ***sälyte-ttiin*** *lähetystöneuvoksille ja -sihteereille*
PROPNNODE_obl[…] _VERBHEAD […]
Lithuania-INE […] pass-PASS.PRET […]
**'In Lithuania**, the responsibility […] **was passed** to the counsellors and secretaries.'

(8) *Kokouksen lopussa ihmiset kysyivät Pentiltä* ***ja*** ***Tuula-lta***.
[…] CCONJNODE_cc _PROPNHEAD

[…] and Tuula-ABL

’At the end of the meeting the people asked Pentti **and Tuula**.’

(9) *Ohjaaja omistaa* ***yhtiö*** ***Sputnik-in***.

[…] NOUNNODE_compound:nn _PROPNHEAD

[…] company Sputnik-ACC

’The director owns **a company [named] Sputnik**.

(10) ***Karjala-n*** ***metsäteollisuus*** *on kärsinyt tappiota*

PROPONNODE_nmod:poss _NOUNHEAD […]

Karelia-GEN forest_industry.NOM […]

’The **forest industries in Karelia** have suffered losses.

(11) ***Tuottaja-t*** *ja näyttelijät* ***Bollywoodi-stä*** *voivat kuvata rauhassa*.

NOUNHEAD_ […] nmod_PROPNHEAD […]

producer-PL.NOM […] Bollywood-ELA […]

’**Producers** and actors **from Bollywood** can film in peace.

| **Positive features** | | **Negative features** | |
|---|---|---|---|
| VERBHEAD_obl_PROPNNODE | 0.678 | ADVNODE_advmod_VERBHEAD | -0.502 |
| CCONJNODE_cc_PROPNHEAD | 0.602 | | |
| NOUNNODE_compound:nn_PROPNHEAD | 0.497 | | |
| PROPNNODE_nmod:poss_NOUNHEAD | 0.831 | | |
| PROPNNODE_obl_VERBHEAD | 0.625 | | |
| NOUNHEAD_nummod_NUMNODE | 0.480 | | |
| NOUNHEAD_nmod_PROPNNODE | 0.671 | | |

**Table 3**. POS bigrams loading onto dimension 2.

While there may be various factors influencing the use of proper nouns, it has been hypothesized (Baker 1993: 243–244), and to a degree proved (Volansky, Ordan & Wintner 2013), to distinguish translated and non-translated texts. Explicitation by means of proper nouns is in general supported by data patterning regarding dimension 2. As can be seen in the upper part of figures 3, 4, and 5, positive mean scores for dimension 2 are almost exclusive to LT-corpora and L2-corpora. This could be due to the lack of gender distinctions in Finnish pronominal reference (where *hän* stands for both ‘he’ and ‘she’). However, earlier results do not support this hypothesis: according to Teitto (2010: 52), at least in narrative texts the differences are rather due to differences between dialogue and narration – and in narration proper names are actually more common in F1 than in FT. What is more, the patterning is not uniform, as the positive mean scores are dominated by three subcorpora (LT-ru_academic: 11.5; L2-ru_narrative: 6.3; L2-de_narrative: 3.6) while the rest of the subcorpora have mean scores close to 0 or even negative. In other words, while constrained Finnish may favour explicitation by means of proper nouns, and while this is probably not due to a single L1/SL or register, the observation cannot be generalized to all data subsets.

#### *4.2.3 Dimension 3: Phrasal complexity*

The bigrams that load onto dimension 3 (see table 4) reflect two superficially diverging linguistic phenomena: non-prototypical clausal word order and noun phrase complexity. Pre-verbal non-core arguments (12) and direct objects (13) contribute to the former, whereas pre- (14) and post-nominal modification (15) – including non-finite verb constructions (16) – as well as phrasal nominal coordination (17) contribute to the latter. In terms the type of complexity seen here, examples (14)–(17) are all related to phrasal complexity, and in the case of (16) also to sentence complexity (Bulté & Housen 2012: 27). It is worth noting that many of the bigrams related to noun phrase complexity also load negatively on dimension 1 – which is in turn characterized by the interaction of morphological and syntactic complexity.

(12) ***Tulkki-na*** ***toimi*** *Nykopp.*
NOUNNODE_obl _VERBHEAD […]
interpreter-ESS act.PRET.3SG […]
'Nykopp **acted as the interpreter**.'

(13) *Autojen* ***pakokaasu-j-a*** ***vähenne-tään*** *katalysaattoreilla.*
[…] NOUNNODE_obj […] _VERBHEAD […]
[…] exhaust_gas-PL-PTV […] reduce-PASS.PRES […]
'Car **exhaust gases are reduced** by means of catalytic converters.

(14) ***Kysynnä-n*** ***puute*** *lisää yrityksien velkaa.*
NOUNNODE_nmod:poss _NOUNHEAD […]

demand-GEN lack […]

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

(15) *sellainen on* ***paluu-ta*** ***keskiaika-an***.
[…] NOUNHEAD_ nmod_NOUNNODE
[…] return-PTV middle_age-ILL

’that is a **return to the middle age**.’

(16) *retoriikan* ***käyttä-mä*** ***päättely*** *lähtee* […]
[…] VERBNODE_acl _NOUNHEAD […]
[…] use-PTCPAG reasoning […]

‘the **reasoning used** in the rhetoric begins…’

(17) *Lyydin* ***kieli*** ***ja*** *lyydiläisten* ***kulttuuri*** *vähitellen häviävät.*
[…] CCONJNODE_cc […] _NOUNHEAD
[…] NOUNHEAD_ […] conj_NOUNNODE
[…] language and […] culture

’The Ludic **language and** the Ludic **culture** slowly disappears.’

| **Positive features** | | **Negative features** | |
|---|---|---|---|
| NOUNNODE_obl_VERBHEAD | 0.537 | PRONNODE_nsubj_VERBHEAD | -0.477 |
| NOUNNODE_obj_VERBHEAD | 0.602 | | |
| NOUNHEAD_nmod_NOUNNODE | 0.533 | | |
| CCONJNODE_cc_NOUNHEAD | 0.514 | | |
| VERBNODE_acl_NOUNHEAD | 0.898 | | |
| NOUNNODE_nmod:poss_NOUNHEAD | 0.611 | | |
| NOUNHEAD_conj_NOUNNODE | (0.439) | | |

**Table 4**. POS bigrams loading onto dimension 3.

The mean score distribution for dimension 3 reveals a clear pattern. As seen in figures 6, 7, and 8, the upper and lower extremes of the y axis are dominated by L1- and LT-corpora. Positive values characterize the academic register and negative values characterize the narrative register, meaning that the above-described constructions are typical for academic texts in these varieties and atypical for narrative texts. As with dimension 1, the mean scores reflect a stronger tendency in L1 (L1-academic_pub: 7.5; L1-academic_nonpub: 7.4 vs. L1-narrative_pub: –4.4; L1-narrative_nonpub: –6.9) than in LT (LT-de_academic: 5.0; LT-ru_academic: 4.5 vs. LT-de_narrative: –5.7; LT-ru_narrative: –5.6). The L2-corpora occupy the middle ground, indicating that the features do not follow any register-related patterning in these subcorpora (L2-de_academic: –0.9; L2-ru_academic: 1.8 vs. L2-de_narrative: –2.0; L2-ru_narrative: –0.7).

The pre-verbal nominal arguments are dominated by two constructions: active voice clauses where the non-prototypical word order is used as a cohesive device, as in (12); and passive voice constructions, where the pre-verbal positioning is prototypical, as in (13). Earlier research has identified the lesser use of non-prototypical word order as one of the complexity features that distinguish even advanced F2 from F1 (Ivaska 2014b: 177–179), and the relative frequency of the passive has also been shown to correlate with proficiency (Seilonen 2013: 58–59). Noun phrase complexity, then, may be related to another tendency found earlier – that post-verbal noun phrases in F2 are often less complex than those in F1 (Ivaska 2014b: 174–176). In other words, the two superficially diverging phenomena are probably intertwined, and both may ultimately be related to noun phrase complexity.

The present results corroborate earlier findings, showing that differences are closely related to register typicalities. Interestingly, greater use of nouns and noun phrase complexity are among the features distinguishing involved vs. informational production in many MDA-based studies of variation across registers (Biber 2014), as well as a feature that differentiates constrained from non-constrained English (Kruger & van Rooy 2018: 229–231)

**Figure 6**. **Scatterplot of dimensions 1 and 3 in relation to register.**

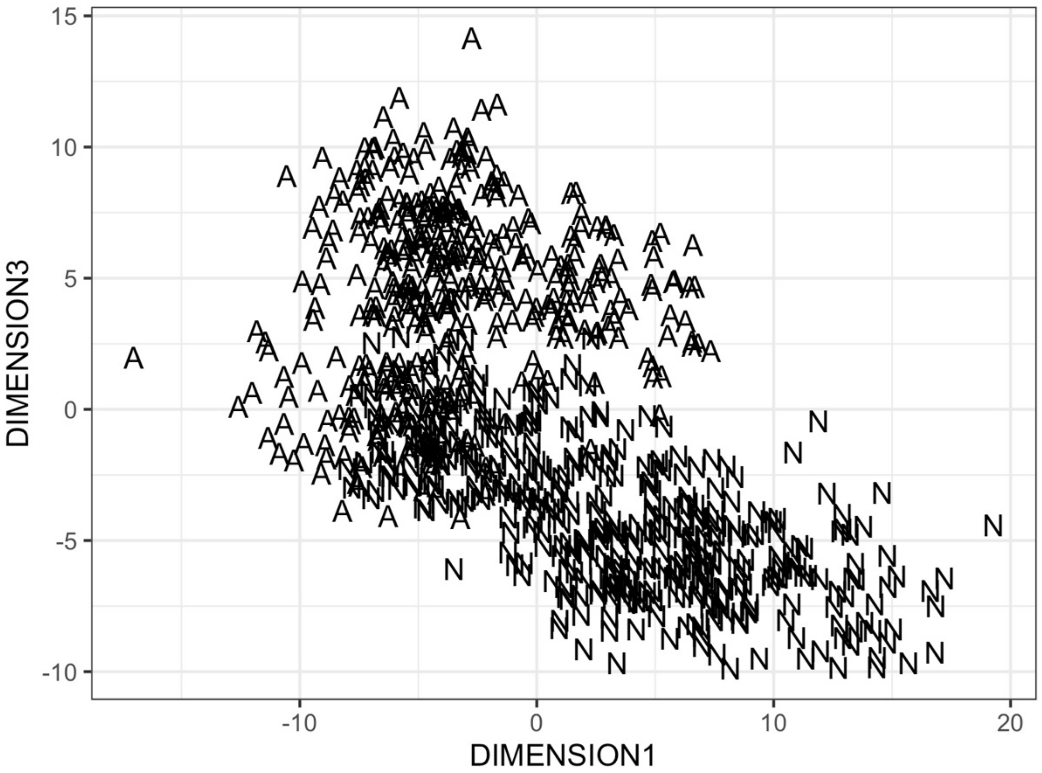


**Figure 7**. **Scatterplot of dimensions 1 and 3 in relation to variety.**

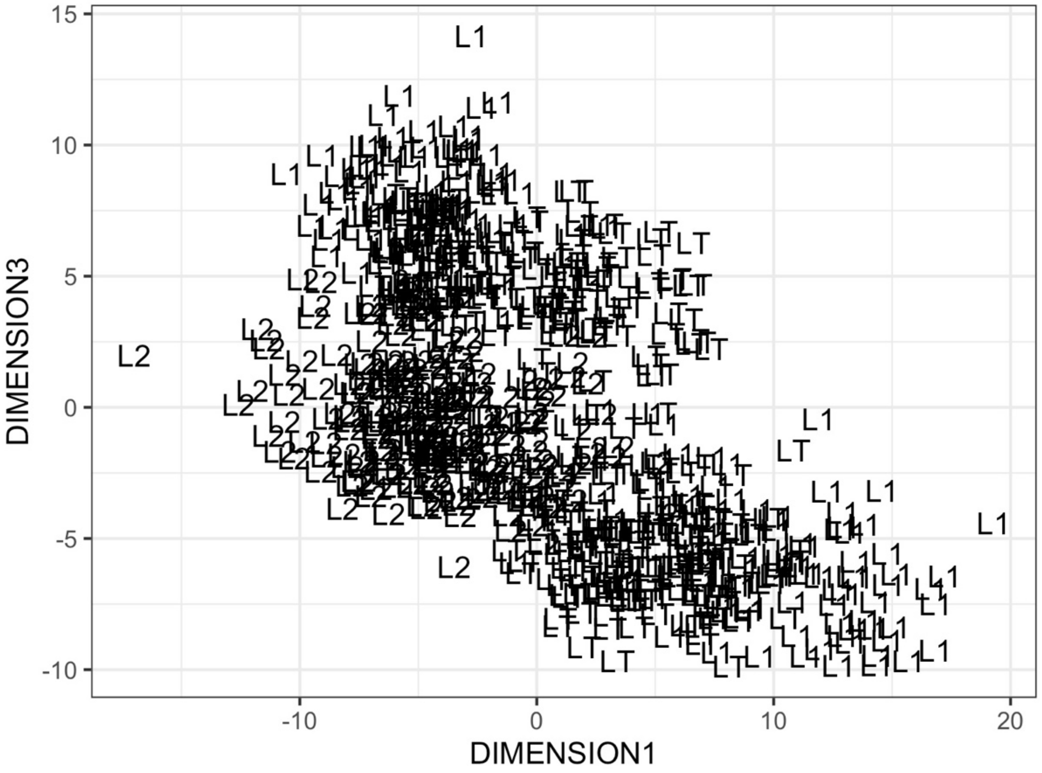
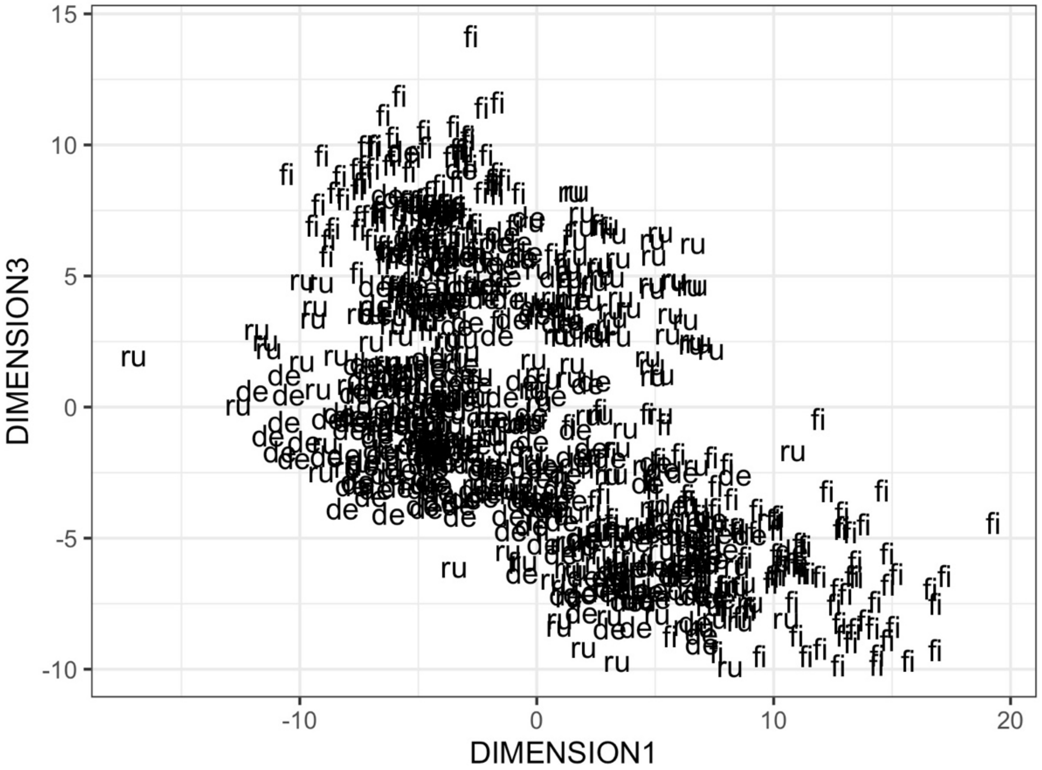


**Figure 8**. **Scatterplot of dimensions 1 and 3 in relation to first/source language.**



## 5. DISCUSSION AND CONCLUSION

We identified 32 syntactically defined POS bigrams as potentially relevant features in distinguishing constrained from non-constrained Finnish. In a multidimensional analysis, these bigrams formed three quantitatively distinct, functionally meaningful groups across datasets, indicating that verbal/clausal complexity, noun phrase complexity, and proper noun explicitation distinguish constrained from non-constrained Finnish across first/source languages and registers.

In the case of both verbal/clausal and noun phrase complexity, the direction of this relationship is not linear and uniform, but rather register-dependent. In non-constrained Finnish, the narrative register portrays higher verbal/clausal complexity than the academic register, whereas the academic register features more nominal complexity. The constrained varieties portray smaller differences: Finnish as a second language is generally characterized by lower verbal/clausal complexity, allowing one to group both registers together with non-constrained academic data. Translated Finnish, in turn, reflects the same distinction but to a lesser degree. As for noun phrase complexity, in non-constrained Finnish the academic register is more complex than the narrative register. Translated Finnish, again, reflects the same distinction but to a lesser degree, and Finnish as a second language positions itself in the middle of this continuum, with virtually no difference between the registers. Interestingly, while the distinction between these two dimensions is often related to the distinction between speaking and writing, some of the included non-finite constructions – while used for reporting speech – are actually less common in spoken than in written Finnish (VISK §538). Thus, the difference is due to the literary means for reporting spoken language rather than to spoken language as such, and the difference is not related to real-time production constraints but rather to the way they are syntactically mimicked in writing. The use of proper nouns does not show any clear register effects: it simply characterizes some of the constrained datasets where proper nouns are relatively more abundant than in the rest of the data.

Many of the constructions identified as verbally/clausally complex have been studied earlier in translated Finnish under the unique item hypothesis (Eskola 2004). Our results corroborate the hypothesis and suggest that it might apply to constrained language use in general. Similarly, noun phrase complexity has earlier been shown to distinguish the production of even advanced Finnish as a second language users from comparable first language users (Ivaska 2014b). Our results confirm this observation and lend partial support to extending it to other forms of constrained language use.

Looking beyond Finnish, it is noteworthy that similar features have been suggested to distinguish constrained from non-constrained language use in English, too (Kruger & van Rooy 2016), and shown to correspond to proficiency in written informational registers in English (Biber, Gray & Staples 2016). In both verbal/clausal and nominal complexity, the constrained varieties portray less inter-register variation than non-constrained data, which at the surface level supports the levelling out hypothesis (Baker 1996). However, as suggested by Szymor (2018), this may rather be a general usage-based mechanism, whereby constrained language use relies on models from registers with which the language users are more familiar. In terms of Bulté and Housen's (2012) SLA complexity taxonomy, our results reflect a register-related interaction between morphological and syntactic complexity. Our results could thus be used both as a point of departure for a more focused complexity study regarding constrained Finnish, and as a potential direction for interpretation of constrainedness effects in other languages. Finally, the use of proper nouns has been suggested to reflect explicitation in translation (Baker 1993). As in Volansky et al. (2013), our results lend partial support to the hypothesis, and extend it to constrained language in general: the difference is clear in some constrained datasets, but non-existent in others, suggesting that the use of proper nouns may be more sensitive to factors like topical variation rather than constrainedness.

Looking back to the three key theoretical constructs – varieties, CLI and registers – our results highlight their inherently intertwined nature. The most consistent quantitative differences between constrained and non-constrained varieties are indeed related to how they portray register typicalities. Both Finnish as a second language and translated Finnish show smaller differences between academic and narrative writing than non-translated first language Finnish. Still, the constrained varieties also diverge from each other, as the translated variety typically positions itself between the first and the second language varieties – corroborating the results of Kruger and van Rooy (2016). Overall, our results clearly underline the centrality of register sensitivity in both language teaching as well translator training and practice.

The present analysis only included two registers, and so the results are obviously shaped by their peculiarities. For example, switching between direct and indirect speech can be considered a particularity of narrative writing and, consequently, the inclusion of multiple registers could tease it apart from verbal complexity. However, with a topic as complex as constrained language use, one has to balance between what is desired and what is viable. Identifying and compiling the needed data even for the present design was highly challenging; including even one more register and keeping all the other variables unchanged would have required a minimum of five new data subsets.

The stepwise data-driven methodological procedure first contrasted each constrained data subset to a closely comparable non-constrained counterpart and then used these results to reveal the consistent differences. In other words, we could make full use of the existing resources, and narrow them down to a balanced subsample only for the second phase. This procedure was adopted to maximize data comparability without losing the generalizability of the results, and to control for the influence of each theoretically motivated construct, all the while doing justice to the inherent variation of naturally occurring linguistic data. We would suggest that studies conducted on a wider range of registers and typologically different languages, adopting sound quantitative methods supported by qualitative interpretations, are essential in the search for confirmation of the constrained language hypothesis.

## CORPORA USED

CTF = Corpus of Translated Finnish. Mauranen, Anna. 2000. Strange strings in translated language: A study on corpora. In Maeve Olohan (ed.), *Intercultural Faultlines: Research Models in Translation Studies*, 119–141. Manchester: St Jerome Publishing.

FinDe = Contrastive Corpus of Finnish and German. http://urn.fi/urn:nbn:fi:lb-20140730137

ICLFI = International Corpus of Learner Finnish. Jantunen, Jarmo. 2011. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. *Lähivõrdlusi. Lähivertailuja* 21, 86–105.

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

LAS2 = The Corpus of Advanced Learner Finnish. Ivaska, Ilmari. 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.

LAS1 = Corpus of Academic Finnish. https://www.utu.fi/en/university/faculty-of-humanities/finnish-and-finno-ugric-languages/syntax-archive

## REFERENCES

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

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, 175–187. Amsterdam: John Benjamins.

Baroni, Marco & Silvia Bernardini. 2006. A new approach to the study of translationese: machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3), 259–274.

Becher, Viktor. 2010. Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1), 1–28.

Berber Sardinha, Tony & Marcia Veirano Pinto (eds.). 2014. *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber*. Amsterdam: John Benjamins.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27(1), 3–43.

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14(1), 7–34.

Biber, Douglas & Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, Douglas, Bethany Gray & Shelley Staples. 2016. Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. *Applied Linguistics* 37(5), 639–668.

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter & Jan Hajič. 2013. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *Transactions of the Association for Computational Linguistics* 1, 415–428.

Breiman, Leo. 2001. Random Forests. *Machine Learning* 45(1), 5–32.

Bulté, Bram & Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, 21–46. Amsterdam: John Benjamins.

Eskola, Sari. 2004. Untypical frequencies in translated language: a corpus-based study on a literary corpus of translated and non-translated Finnish. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation Universals: Do they exist?*, 83–99. Amsterdam: John Benjamins.

Filipović, Luna & John A. Hawkins. 2013. Multiple factors in second language acquisition: The CASP model. *Linguistics* 51(1), 145–176.

Gabrielatos, Costas. 2018. Keyness analysis: nature, metrics and techniques. In C Taylor & A Marchi (eds.), *Corpus Approaches to Discourse: A critical review*, 225–258. Oxford: Routledge.

Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International*

*Journal of Learner Corpus Research* 1(1), 7–24.

Gries, Stefan Th. 2019. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* aop.

Grosjean, François. 2001. The bilingual's language modes. In J Nicol (ed.), *One Mind, Two Languages*, 1–22. Oxford: Blackwell Publishers.

House, Juliane. 2008. Beyond intervention: Universals in translation? *trans-kom* 1(1), 6–19.

Ivaska, Ilmari. 2014a. 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.

Ivaska, Ilmari. 2014b. Edistyneen oppijansuomen avainrakenteita. Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin. *Virittäjä* 118(2), 161–193.

Ivaska, Ilmari. 2014c. Mahdollisuuden ilmaiseminen S1-suomea ja edistynyttä S2-suomea erottavana piirteenä. *Lähivõrdlusi. Lähivertailuja* 24, 47–80.

Ivaska, Ilmari. 2015. Longitudinal changes in academic learner Finnish: A key structure analysis. *International Journal of Learner Corpus Research* 1(2), 210–241.

Ivaska, Ilmari, Elisa Reunanen & Kirsti Siitonen. 2016. Infinite Konstruktionen im fortgeschrittenen Finnisch als Fremdsprache. *Ural-Altaische Jahrbücher* 26, 46–76.

Ivaska, Ilmari & Kirsti Siitonen. 2017a. Learner language morphology as a window to crosslinguistic influences: A key structure analysis. *Nordic Journal of Linguistics* 40(2), 225–253.

Ivaska, Ilmari & Kirsti Siitonen. 2017b. tehdessä-konstruktio edistyneessä oppijansuomessa. Korpusanalyysin ja oikeakielisyysarviointien ristivalotus. *Sananjalka* 59, 154–180.

Ivaska, Laura. 2019. Distinguishing translations from non-translations and identifying (in-)direct translations' source languages. In Jarmo Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen & Katja Västi (eds.), *Proceedings of the Research Data and Humanities (RDHum) 2019 Conference: Data, Methods and Tools*, 125–138. Oulu.

Iwasaki, Shoichi. 2015. A multiple-grammar model of speakers' linguistic knowledge. *Cognitive Linguistics* 26(2), 161–210.

Jantunen, Jarmo. 2004. Untypical patterns in translations. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation Universals: Do they exist?*, 101–126. Amsterdam: John Benjamins.

Jantunen, Jarmo. 2008. Haasteita oppijankielen korpusanalyysille: oppijankielen universaalit. In Pille Eslon (ed.), *Õppijakeele analüüs: võimalused, probleemid, vajadused.*, 67–92. Tallinn: Tallinna Ülikool.

Jantunen, Jarmo. 2011a. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. *Lähivõrdlusi. Lähivertailuja* 21, 86–105.

Jantunen, Jarmo. 2011b. Avainsana-analyysi annotoidun oppijankieliaineiston tutkimuksessa: Alustavia havaintoja. (Ed.) Esa Lehtinen, Sirkku Aaltonen, Merja Koskela, Elina Nevasaari & Mariann Skog-Södersved. *AFinla-e* 3, 48–61.

Jantunen, Jarmo & Sari Eskola. 2002. Käännössuomi kielivarianttina: Syntaktisia ja leksikaalisia erityispiirteitä. *Virittäjä* 106(2), 184–207.

Jarvis, Scott. 2000. Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning* 50(2), 245–309.

Jarvis, Scott. 2010. Comparison-based and detection-based approaches to transfer research. *EUROSLA Yearbook* 10, 169–192.

Kaiser, Henry F. 1974. An index of factorial simplicity. *Psychometrika* 39(1), 31–36.

Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino & Tapio Salakoski. 2018. Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to*

*Universal Dependencies*. Brussels, Belgium: ACL.

Kolehmainen, Leena, Lea Meriläinen & Helka Riionheimo. 2014. Interlingual reduction: Evidence from language contacts, translation and second language acquisition. In Heli Paulasto, Lea Meriläinen, Helka Riionheimo & Maria Kok (eds.), *Language Contacts at the Crossroads of Disciplines*, 3–32. Cambridge: Cambridge Scholars Publishing.

Koppel, Moshe & Noam Ordan. 2011. Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1318–1326. Portland, Oregon: ACL.

Kruger, Haidee. 2017. The effects of editorial intervention. Implications for studies of the features of translated language. In Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical Translation Studies: New Methodological and Theoretical Traditions*, vol. 300.

Kruger, Haidee & Bertus 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, Haidee & Bertus van Rooy. 2018. Register variation in written contact varieties of English. *English World-Wide* 39(2), 214–242.

Kujamäki, Pekka. 2004. What happens to “unique items” in learners’ translations? In Anna Mauranen & Pekka Kujamäki (eds.), *Translation Universals: Do they exist?*, 187–204. Amsterdam: John Benjamins.

Kursa, Miron & Witold Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software, Articles* 36(11), 1–13.

Lanstyák, Istvan & Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1), 99–121.

Leech, Geoffrey. 2006. New resources, or just better old ones? The Holy Grail of representativeness. In Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus Linguistics and the Web*, 133–149. London: Brill.

Lefer, Marie-Aude & Svetlana Vogeleer. 2013. Interference and normalization in genre-controlled multilingual corpora: Introduction. *Belgian Journal of Linguistics* 27(1), 1–21.

Mauranen, Anna. 2000. Strange strings in translated language: A study on corpora. In Maeve Olohan (ed.), *Intercultural Faultlines: Research Models in Translation Studies*, 119–141. Manchester: St Jerome Publishing.

Mauranen, Anna. 2004. Corpora, universals and interference. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation Universals: Do they exist?*, 65–82. Amsterdam: John Benjamins.

Mauranen, Anna & Pekka Kujamäki. 2004. *Translation Universals: Do they exist?* Amsterdam: John Benjamins.

Mauranen, Anna & Liisa Tiittula. 2005. MINÄ käännössuomessa ja supisuomessa. In Anna Mauranen & Jarmo Jantunen (eds.), *Käännössuomeksi. Tutkimuksia suomennosten kielestä.*, 35–69. Tampere: Tampere University Press.

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, Ella, Sergu Nisioi, Noam Ordan & Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1870–1881. Berlin: ACL.

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

Rohdenburg, Günther. 1996. Cognitive Complexity and Increased Grammatical Explicitness in English. *Cognitive Linguistics* 7(2), 149–182.

Seilonen, Marja. 2013. Epäsuora henkilöön viittaminen oppijansuomessa. PhD Thesis, University of Jyväskylä.

Spoelman, Marianne. 2013. Prior linguistic knowledge matters: the use of the partitive case in Finnish learner language. PhD Thesis, University of Oulu.

Szmrecsanyi, Benedikt. 2017. Variationist sociolinguistics and corpus-based variationist linguistics: overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 62(4), 685–701.

Szymor, Nina. 2018. Translation: universals or cognition? A usage-based perspective. *Target* 30(1), 53–86.

Teitto, Heli. 2010. Human Referents in Subtitles: A Study on Personal Pronouns and Proper Nouns in Translated and Original Finnish. MA thesis, University of Eastern Finland.

Tirkkonen-Condit, Sonja. 2004. Unique items – over- or under-represented in translated language? In Anna Mauranen & Pekka Kujamäki (eds.), *Translation Universals: Do they exist?*, 177–184. Amsterdam: John Benjamins.

Tirkkonen-Condit, Sonja. 2005. Häviävätkö uniikkiainekset käännössuomesta? In Anna Mauranen & Jarmo Jantunen (eds.), *Käännössuomeksi. Tutkimuksia suomennosten kielestä.*, 12–137. Tampere: Tampere University Press.

Toury, Gideon. 2012. *Descriptive Translation Studies – and beyond : Revised edition*. Amsterdam: John Benjamins.

VISK = Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho 2004: Iso suomen kielioppi. Helsinki: Suomalaisen Kirjallisuuden Seura. November 24, 2019. http://scripta.kotus.fi/visk.

Volansky, Vered, Noam Ordan & Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities* 30(1), 98–118.

# NOTES

[1] Abbreviations used: 1, 2, 3 = first, second, third person; ABL = ablative; ACC = accusative; acl = clausal modifier of noun; ADJ = adjective; ADV = adverb; advcl = adverbial clause modifier; advmod = adverb modifier; amod = adjectival modifier; AUX = auxiliary verb; cc = coordinating conjunction; ccomp = clausal complement; CCONJ = coordinating conjunction; compound:nn = noun compound modifier; conj = conjunct; cop = copula; ELA = elative; ESS = essive; GEN = genitive; IMP = imperative; INE = inessive; INF2 = second infinitive; INF3 = third infinitive; ILL = illative; nmod = nominal modifier; nmod:poss = genitive modifier; NOM = nominative; NOUN = noun; nsubj = nominal subject; nsubj:cop = nominal copular subject; nummod = numeric modifier; obj = object; obl = oblique nominal; parataxis = parataxis; PASS = passive; PL = plural; PTCP1 = first participle; PTCP2 = second participle, PTCPAG = agent participle; PRES = present; PRET = preterite; PRON = pronoun; PROPN = proper noun; PTV = partitive; PX = possessive suffix; root = root of the sentence; SCONJ = subordinating conjunction; SG = singular; VERB = verb; xcomp = open clausal complement; xcomp:ds = clausal complement with different subject.

[2] In this paper, we use the term *variety* to refer to broad categorizations, such as acquisitional status (e.g. first vs. second language) and mode of production (e.g. translated vs. original).

[3] As far as TS is concerned, we mainly rely on the thorough meta analysis in Kolehmainen et al. (2014).

[4] All data cleaning was done using the Java programming language and all statistical analyses using the R programming environment (R Core Team 2018). The frequency data and the R scripts can be found here: https://osf.io/twg5u/?view_only=ca879e7d907d4d95987426d2b1dcb0ae.

[5] The features with loadings in parentheses in tables 2–4 have a higher loading on another dimension and are only included in that dimension score.

[6] Examples shows the relevant bigram in boldface, followed by the corresponding POS bigram, a morpheme-for-morpheme gloss, and an English translation.

**Appendix 1**. POS bigrams contributing the most to distinguishing constrained and non-constrained texts.

| Consistency | Bigrams |
|---|---|
| 7/8 | NOUNNODE_nsubj.cop_ADJHEAD (nominal subject of an adjectival copula clause) |
| | VERBHEAD_obl_PROPNNODE (post-verbal proper noun as a non-core argument) |
| | CCONJNODE_cc_PROPNHEAD (coordinating conjunction with a proper noun as a conjunct) |
| | NOUNNODE_compound.nn_PROPNHEAD (compound noun with an appellative and a proper noun) |
| | PROPNNODE_nsubj_VERBHEAD" (pre-verbal proper noun as a nominal subject) |
| 6/8 | PRONNODE_nsubj_VERBHEAD (pronoun as a nominal subject preceding the verb) |
| | ADVNODE_advmod_VERBHEAD (pre-verbal adverbial modifier) |
| | NOUNNODE_obl_VERBHEAD (pre-verbal noun as a non-core argument) |
| | NOUNNODE_obj_VERBHEAD (pre-verbal noun as a direct object) |
| | NOUNHEAD_nmod_NOUNNODE (post-nominal noun as a nominal modifier) |
| | VERBHEAD_obj_PRONNODE (post-verbal pronoun as a direct object) |
| | PROPNNODE_nmod.poss_NOUNHEAD (pre-nominal proper noun as a genitive modifier) |
| | PROPNNODE_obl_VERBHEAD (pre-verbal proper noun as a non-core argument) |
| | NOUNHEAD_nummod_NUMNODE (post-nominal numeral modifier) |
| | CCONJNODE_cc_NOUNHEAD (coordinating conjunction with a noun as a conjunct) |
| | VERBHEAD_advcl_VERBNODE (verbal predicate followed by a subordinate non-copular adverbial clause as a modifier) |
| | VERBHEAD_xcomp.ds_VERBNODE (verbal predicate followed by a non-finite clausal complement with the object of the governing clause as a subject) |
| | NOUNHEAD_nmod_PROPNNODE" (post-nominal proper noun as a nominal modifier) |
| 5/8 | PRONNODE_det_NOUNHEAD (pre-nominal pronoun as a determiner) |
| | VERBNODE_acl_NOUNHEAD (pre-nominal finite or non-finite clausal modifier of a noun) |
| | ADVNODE_advmod_ADVHEAD (pre-adverbial adverb as an adverbial modifier) |
| | NOUNNODE_nsubj.cop_NOUNHEAD (nominal subject of a nominal copula clause) |
| | NOUNNODE_compound.nn_NOUNHEAD (compound noun with two appellative nouns) |
| | ADJHEAD_xcomp.ds_VERBNODE (adjectival copula clause followed by a non-finite clausal complement with the governing clause as a subject) |
| | NOUNHEAD_ccomp_VERBNODE (post-nominal subordinate clausal complement of a noun) |
| | VERBHEAD_obl_NOUNNODE (post-verbal noun as a non-core argument) |
| | NOUNNODE_nmod.poss_NOUNHEAD (pre-nominal noun as a genitive modifier) |
| | NOUNHEAD_conj_NOUNNODE (two coordinated nouns) |
| | PROPNHEAD_flat.name_PROPNNODE (name consisting of two proper nouns( |
| | VERBHEAD_parataxis_VERBNODE (two coordinated non-copular clauses with no explicitly marked coordination) |
| | NOUNHEAD_parataxis_VERBNODE (nominal copula clause followed by a coordinated non-copular clause with no explicitly marked coordination) |
| | VERBHEAD_xcomp_VERBNODE (post-verbal non-finite clausal complement that shares the subject with the preceding governing non-copular clause) |