AUTHOR	Ilmari Ivaska and Laura Ivaska
TITLE	Looking under the hood: which linguistic features contribute to the source language classification of direct and indirect translations into Finnish, and why is that?
YEAR	2024
DOI	10.1556/084.2024.00912
VERSION	Accepted manuscript
CITATION	Ivaska, Ilmari & Laura Ivaska. 2024. Looking under the hood: which linguistic features contribute to the source language classification of direct and indirect translations into Finnish, and why is that? Across Languages and Cultures 25(2). 216–239. https://doi.org/10.1556/084.2024.00912

Looking under the hood: which linguistic features contribute to the source language classification of direct and indirect translations into Finnish, and why is that?[1]

Ilmari Ivaska* (ORCID: 0000-0001-7366-7111) and Laura Ivaska (ORCID0000-0002-6998-5077)

University of Turku, Finland

ORIGINAL RESEARCH PAPER

ABSTRACT

The study of features that affect the linguistic form of translated texts has been one of the central questions within the field of corpus-based translation studies. In the partially overlapping field of computational linguistics, previous studies have shown that source languages of individual texts can be detected automatically in direct translations and indirect translations (i.e., translations done from translations). However, computationally oriented approaches have paid limited attention to what specific linguistic features make successful classification possible. Consequently, the types of linguistic phenomena characterizing translations and the kinds of linguistic interference that can be detected in them remain underexplored. In this study, we study the linguistic features that contribute to the identification of the source language of direct translations from English, French, German, Greek, and Swedish, as well as indirect translations from Greek into Finnish, with English, French, German, and Swedish as mediating languages. Theoretically, this study builds on Halverson's (2017) gravitational pull model to explain the mechanisms behind our findings in a theoretically sound fashion and to generate theoretically motivated, specific hypotheses to be tested by future research. The analysis makes use of keyness analysis as a supervised machine learning technique, as well as exploratory factor analysis (EFA) as an unsupervised machine learning technique. The results indicate that sentence length, sentence-initial adverbs and sentence-final specification are the linguistic features that set the different types of translations apart from each other. Furthermore, the salient features of the ultimate source language outweigh those of the mediating languages in indirect translations or the entrenched parallels between specific language pairs.

KEYWORDS

indirect translation, crosslinguistic interference, keyness analysis, gravitational pull model, Finnish

---

*Corresponding author. E-mail: ilmari.ivaska@utu.fi

1. INTRODUCTION

[1] This is authors' accepted manuscript of a published article. Please refer to the final paper: Ivaska, Ilmari & Laura Ivaska. 2024. Looking under the hood: which linguistic features contribute to the source language classification of direct and indirect translations into Finnish, and why is that? *Across Languages and Cultures* 25(2). 216–239. https://doi.org/10.1556/084.2024.00912.

Translations are not always done directly from the original source text but can also be done from translations. In the latter instances, the process and product can be called *indirect translation* (Assis Rosa et al., 2017). Indirect translations (ITrs) are everywhere: in literary translation, news translation, audiovisual translation, interpreting – and even institutions like the UN and the EU make use of it (Pięta et al., 2022). Until recently, research on this topic has been scarce, possibly because of the negative attitudes towards the practice, as it is thought that the effect of the "message getting lost in translation" is exacerbated when translations are chained (St. André, 2020). However, studies focusing on the linguistic features suggest that something of the source text is relayed over even in cases when translations are chained. Namely, the effect of the source language can be detected in a translation also when done via a mediating language (Rabinovich et al., 2017; I. Ivaska & L. Ivaska, 2022; see also Ustaszewski, 2021). Less attention has been paid to the linguistic features that are carried over when translating indirectly, on the one hand, and what that tells us about the cross-linguistic influences (CLIs) that characterize translations, on the other hand; it is this shortcoming that the present study addresses.

Interestingly, while the theoretical point-of-departure for most studies on source language (SL) classification of translated texts is the concept of cross-linguistic influences between the languages involved – or interference, to use Toury's (2012) terms – they deal with this concept only in passing. In many ways, this corresponds to what Kotze (2020, p. 336) describes as "the ontological level at which explanations for empirical phenomena are sought", or in this case, only meager engagement in that discussion. Hence, the analysis of the linguistic phenomena that underlie any successful classification has so far been superficial at best, and the focus has been exclusively in generic system level interpretations, such as Koppel & Ordan (2011, p. 1325) concluding that "translations from similar source languages are different from non-translated texts in similar ways." The same applies even more categorically to any studies on SL classification of ITrs, ourselves included. The research mindset has been purely empirical and descriptive in nature, focusing mostly on how ITrs pattern in SL classification with regard to comparable direct translations in terms of the linguistic form they comprise (I. Ivaska & L. Ivaska, 2022; Rabinovich et al., 2017).

The overall goal of the paper is to contribute to Kotze's (2020) call for the field of empirical translation studies to "[c]onverg[e] *What* and *How* to find out *Why,*" by using product data and developing theoretically-grounded hypotheses and motivated lines of interpretation regarding the general nature of CLIs in translated language, and to see where ITrs land in those regards. First, we delve deeper into the nature of the linguistic phenomena that contribute to distinguishing translations from different SLs from each other, and the positioning of ITrs in that constellation. Second, we strive to develop theoretically plausible hypotheses, as to why it is those linguistic phenomena, and not some other, that contribute to the positioning of direct and indirect translations in a SL classification task. The research design is, hence, decidedly exploratory and not confirmatory: we are interested in generating theoretically sound hypotheses supported by the data that we make use of, so that future – possibly process-oriented – scholarship could build on our hypotheses, either corroborating, modifying, or rejecting them (for the importance of keeping the two types of research designs separate, see Winter, 2020, p. 275). Our data of direct translations are Finnish translations of literary texts translated from English, French, German, Greek, and Swedish, as well as ITrs from Greek into Finnish, with English, French, German, and Swedish as mediating languages. The research questions are:

1) Which linguistic features contribute the most to identifying the SL of direct translations from various languages into Finnish?

2) How are these linguistic features distributed in indirect translations into Finnish where the languages studied in research question 1 act as mediating languages, and the ultimate SL is kept constant?

3) In what ways do cognitively-oriented theories of translation lend themselves to explaining the mechanisms underlying such findings?

The paper is structured as follows: in the theoretical section, we discuss Halverson's (2017) gravitational pull model, linking it with the aims of the present paper. Then, we introduce the data and explain how we use in our study keyness analysis as a supervised machine learning technique, as well as exploratory factor analysis (EFA) as an unsupervised machine learning technique. We will then report on the results of the keyness analysis and the EFA, and conclude that sentence length as well as sentence-

initial adverbs and sentence-final specification are the linguistic features that set the different types of translations apart from each other.

2. THEORETICAL BACKGROUND

As a point of departure, we look for cognitively-oriented theorizing of translation, especially the revised version of the gravitational pull (GP) model (Halverson, 2017), and the usage-based models of language that the model in turn draws from (see e.g., Langacker, 2008 for an overview). The GP model is structured around the notion of salience, that is, the relative prominence of some items within the schematic networks that comprise language systems, whereby "some patterns of activation [---] will be more prominent than others, due to their higher frequency of use over time" (Halverson, 2017, p. 13). Salience, then, is expected to show in translated language according to three different mechanisms: 1) prominence in the target language (referred to as *magnetism*); 2) prominence in the source language (referred to as *gravity*); and 3) prominence in the connections between the two languages (referred to as *connectivity*).

Before discussing each of these three mechanisms of the GP model in greater detail from the point-of-view of the present study, we briefly turn to a conceptual issue underlying the CLIs in SL classification and their interpretation under GP. One the one hand, there are typologically motivated CLIs that relate to the structural differences between the languages involved. On the other hand, CLIs can also be motivated by differing genre conventions between the SL and the TL. As a recent example of typologically-oriented view of CLIs, Lefer & De Sutter (2022) point out in their study of French renditions of English concatenated nouns in the European Parliament discourse that the simple noun–noun pattern differs in several respects in the two languages: the constituent order is typically different but the pattern is also generally much more productive in English than in French, whereby the semantic relation between the components that is left implicit in English often needs to be explicitly expressed in French. Such differences are naturally reflected in differences of the relative frequencies of these items, and can consequently lead to cross-linguistic differences in their salience. Lefer and De Sutter show in their study that the effects of the TL magnetism and SL–TL connectivity overpower those of the SL gravity, but that this might be at least in part due to a categorical typological difference between the

languages involved, whereby SL-like choices would in fact often be considered ungrammatical in the TL.

CLIs may also be sensitive to the genre at hand – both in that genres may diverge in how a certain typologically motivated pattern ends up being attested (as mentioned by Halverson, 2017, p. 39) and in that genre conventions in and of themselves differ between cultures and languages (Woodstein, 2022, p. 58). In other words, CLIs might manifest, and be analyzed by means of GP, as differences of relative frequencies of linguistic features within texts of comparable genres also irrespective of systemic typological differences across the languages. Illustratively, Kruger and van Rooy (2018) and I. Ivaska et al. (forthcoming) show that genre variation is extremely important in accounting for variation in relative frequencies of linguistic forms in translated texts, irrespective of the language studied. Interestingly, as noted by Neumann (2014, p. 28), "register analysis has received scant treatment [within translation studies]". A notable exception is indeed Neumann (2014), where cross-linguistic register differences between English and German, and the relative positioning of translated texts, are also explored. Centrally to CLIs, Neumann (2014, p. 306) comes to the conclusion that "[o]n a general level, there seems to be a stronger tendency of the translations to adapt to [the register of] the target language". However, even in that study it is difficult to distinguish between differing genre-related cultural conventions and typological differences between the SLs. If the two types of CLIs could be teased apart, it could be argued that target texts in one TL from various SLs differ and can be distinguished because of genre-related differences between different lingua-cultures rather than typologically motivated systemic differences between the SLs.

From the viewpoint of GP, there is no fundamental difference between CLIs that stem from systemic typological differences and those that pertain to differing genre conventions, as they reflect a profoundly usage-based view on language and can both be explained in terms of differing salience of linguistic features. Yet, we think it is reasonable to identify them separately, as they lend themselves in part to different kinds of interpretations regarding CLIs. In other words, do translations of genre N from language X into language Y differ from translations from language Z into language Y because of the typological differences between languages X and Z, or because of genre differences of N in languages X and Z? The two are naturally not independent from each other, and the linguistic realization of a given genre quite concretely consists of the

structural elements of the respective language. However, cultural conventions do not necessarily correspond to the languages of those communities, and also changes in the cultural and linguistic planes might well take place independently from one another. Hence, in order to address the complex question of *why* translations are the way they are (as per Kotze, 2020), we think it is important to be able to tease the two apart when necessary.

Turning to the different components of GP, **gravity** posits that certain patterns of language use are so salient in the SL that their translational equivalents remain proportionally more frequent in the TL than in translations from other SLs where the pattern less salient. The differences may be due to typological differences or due to differing genre conventions. In SL classification, then, the differences of frequency in such patterns make it possible to tease apart texts with differing SLs. As a typologically motivated systemic example, if a grammatical subject precedes the verb in a source text and if such a constituent order is possible in TL, this is likely to relatively increase the likelihood of the constituent order to stay unchanged in the translation even if other constituent orders were possible in the TL. As for a hypothetical genre-induced difference, an example could be the use of direct and indirect speech in literary texts, whereby the preferred pattern of the SL literary conventions can be expected to increase its likelihood even in translations. In the context of SL classification of ITrs, the core question regarding gravity is whether the ultimate SL or the mediating language that serves as the immediate SL of the texts studied plays a greater role, as witnessed in the results of the classification.

As for the **connectivity** component in GP, it expects certain language-pair-specific translational choices between a given pattern in the SL, and another in the TL, to be so prominent that they lead to relatively more frequent patterns of use in translations. This entrenchment is independent from salience related exclusively to the SL (see gravity above) or the TL (see magnetism below). As a hypothetical example, if a given semantic notion has several near-synonymous equivalents both in the SL and the TL, but one of them is a cognate, this probably increases the relative likelihood of the cognate to be chosen, even if it was not particularly salient in either languages as such. Likewise, optional genre markers recognized in lingua-cultures of both the SL and the TL can be expected to occur relatively more frequently than in cases where they are only recognized by one of the lingua-cultures involved. From the point of view of SL

classification of ITrs, then, the question is whether such entrenched pairs are stronger between the ultimate SL and the mediating language or the mediating language and the ultimate TL. Connectivity between the ultimate SL and the mediating language would lead to increasing the relative likelihood of the ultimate SL in the classification results, whereas connectivity between the mediating language and the ultimate TL would give rise to a higher probability of the mediating languages in the classification results.

The magnetism component of GP gives rise to the salient features specific to the TL, and as such it is not as directly relevant to the study of CLIs as gravity and connectivity are. While magnetism has been shown central to studies explaining phenomena specific to translated language in general (e.g., Lefer & De Sutter, 2022), the focus of the present study is on the nature of the CLIs that do take place rather than to study how important they are in relation to the salient features of the TL. Note, however, that in the case of ITrs, the mediating language serves both as a target language and a source language, so magnetism may also play a part in that it amplifies the salient phenomena of the mediating language that are then preserved in the ultimate TL because of gravity. In other words, in the SL classification of ITrs, salient features of the mediating language could be expected to increase their relative likelihood in the classification results.

Based on the above, we would like to highlight what may already seem obvious by now: the linguistic form provided by the source text, and hence, the SL, are likely to affect the linguistic forms used in the target text, so long as they fit within the boundaries of the TL. Hence, our resulting general hypothesis is that CLIs are most likely to present themselves in translated texts in linguistic phenomena with several possible variants in both the SL and the TL but with clear differences in the typical distributions between those variants across the SL and the TL. While nothing new as such (for a discussion on the concept of "default translation" see Halverson, 2015, 2019), this general hypothesis does have some methodological consequences: CLIs stem from differences in the salience of linguistic items, and consequently, they are inherently distributional in nature. Hence, any attempt to prove them must be able to tease the CLIs apart from naturally occurring idiolectal variation between language users – be it authors of the STs or translators of the translated texts. One option is to generate explicit hypotheses based on a specific theory and relying on earlier research on that phenomenon (for a laudable example, see Lefer & De Sutter, 2022). The downside of such an approach is that the nature of such distributional characteristics –

differing relative frequencies of linguistic features can often be extremely difficult to identify and make explicit a priori (for a detailed discussion, see Tognini-Bonelli, 2001, Chapter 5). What is more, the less studied the language pair in question, the more difficult it is to specify the research gap in terms that are specific enough to allow for testing specific hypotheses. The other natural approach is to identify CLIs is by comparing distributional characteristics of a range of linguistic phenomena in a range of texts by different authors and with different SLs – the successful common practice in SL classification for the past 20 years, but without a proper theoretization of the underlying mechanisms. When these principles are mirrored against the concept of ITr, where the TL of a mediating text is also the SL of the final text, the whole constellation obviously becomes more complex, and exploring this further is one of the central goals of the present paper. To that end, we want to close the theoretical discussion by pointing out that gravity, connectivity and magnetism all have the logical potential to contribute to SL classification in favor of mediating languages over ultimate SLs, while only gravity and magnetism have the potential to contribute to SL classification in favor of ultimate SLs over mediating languages. The fact that earlier empirical work on SL classification of ITrs (Rabinovich et al., 2017; I. Ivaska & L. Ivaska, 2022) still suggests that the effect if ultimate SL is discernible, underlines the need to look under the hood of such classification results.

3. DATA AND METHOD

3.1 Data

The data consists in direct and indirect literary translations (see Table x for number of novels, authors and translators). All texts are in Finnish, and the direct translations are from English, French, German, Greek, and Swedish whereas the ITrs are from Greek via English, French, German and Swedish. All the novels used as data have been published (and translated) in the 20th and 21st centuries (the publication years of the translations range between 1952 and 2019). The majority of the data comes from the Corpus of Translated Finnish (Mauranen, 2004) and the InterCorp corpus (Čermák & Rosen, 2012) while a part of it – especially the translations from Greek – has been collected ad hoc. The indirectness and the mediating languages have been established in

a previous study (L. Ivaska, 2020). A full list of texts can be found in appendices 1 (direct translations) and 2 (ITrs)[2].

Table 1. Number of novels, as well as unique authors and translators across data

Type of data	Novels	Authors	Translators
Direct translations	69	65	44
Indirect translations	15	6	10

All the data were parsed using the Turku NLP dependency parser (Kanerva et al., 2018). The direct translations were then split into train and test sets, with roughly 70% of the texts in the train set and 30% in the test set. Care was taken to ensure that there were no texts from the same author or translator in both sets. Subsequently, the sentences were shuffled within the SL-specific subsets and reorganized into text chunks of 500 sentences each. Finally, a maximal subset of text chunks balanced across SLs was taken for each data subset, resulting in 555 text chunks of train data (111 of each SL) and 205 text chunks of test data (41 of each SL). This was done to ensure that the results reflect overall variance within the data while minimizing topical effects of individual texts. As a downside, they do not reflect variance between actual texts. As for the ITrs, the sentences were shuffled within each text and reconstructed into chunks of 500 sentences without breaking boundaries between individual texts. The same data has also been used in two previous studies (I. Ivaska & L. Ivaska, 2022; L. Ivaska, 2019), and more details on how the data has been compiled and prepared can be found in these studies. This is also the reason for the differing treatment between the ITrs and other data: one of the original goals was to develop a method to identify undocumented ITrs by predicting their likeliest SLs using a model trained on direct translations from various SLs.

3.2 Methodology

This study is methodologically built on two parallel pillars: first, we make use of keyness analysis as a supervised machine learning technique to classify directly translated texts according to their source language and reveal those linguistic features that contribute the most to a successful classification (see Gabrielatos, 2018 for the

[2] All appendices, the frequency data, and the scripts used in the statistical analyses are available on Open Science Framework: <https://osf.io/y3jtd/>

concept of keyness; see I. Ivaska & Bernardini, 2020 for an earlier example of the approach towards keyness adopted also in this paper). Second, we then employ exploratory factor analysis (EFA) as an unsupervised machine learning technique to group the contributing linguistic features into latent dimensions of variation that can, then, be interpreted qualitatively from the point-of-view of CLIs, in relation to the typological nature of the source languages involved (see Egbert & Staples, 2019 for an illustrative example on the practical implementation of EFA adopted also in this paper; see Fabrigar, 2012 for further detail regarding EFA).

Conceptually, keyness analysis is a way to reveal linguistic patterns that characterize, based on their frequencies of occurrence, certain data when contrasted with a different set of data. In the study at hand, we consider as key features those linguistic patterns whose frequency help us reliably differentiate between translations into Finnish based on the source language of the texts. We consider the keyness on two different levels of granularity: first, we want to identify those feature sets, i.e., the types of linguistic patterns, that are the most effective in this differentiation task. Second, we want to extract from the most effective feature sets those specific linguistic patterns that actually contribute this differentiation task. So far, most work on SL classification has focused solely on the first, arguably more coarse level of granularity (Islam & Hoenen, 2013; e.g., Koppel & Ordan, 2011; Lynch & Vogel, 2012). Adding the second phase will help zooming in to the more exact linguistic nature of the CLIs.

For the first phase of the keyness analysis (reported earlier in I. Ivaska & L. Ivaska, 2022), we considered 20 different sets of frequencies of linguistic features that describe the data on different levels of linguistic annotation (word, lemma, parts of speech, syntax, for a full list of feature sets considered, see Table 2). These sets portray the data as sequential N-grams of different lengths (1-gram, 2-gram, 3-gram). In addition, we also include 2-grams based on the syntactic relationship between the two words, 1-grams based on their position in a sentence, as well as character 3-grams. The sentence positions have been operationalized based on the dependency parses, as they are structured around sentences. We used the train data to train a separate statistical classifier for each feature set (e.g., all lemma 1-grams, all part-of-speech 2-grams, or all syntax 3-grams), and evaluated those classifiers based on how well they were able to distinguish between the SLs of the texts included in the test data of direct translations. Then, we took the three feature sets that provided the most reliable means to classify the

texts according to their SL and used them all to train a final model used in the subsequent analysis. This final model serves two purposes: first, as shown in I. Ivaska & L. Ivaska (2022), it can be used to predict the likeliest SLs of the ITrs, so as to see where they land with regard to the traceability of the different SLs involved – predictions that align with the ultimate SL indicate it has a relatively stronger effect, whereas predictions closer to the mediating language indicate that its effect overpowers that of the ultimate SL. Second, and centrally to the present paper, it is used in the second phase of the keyness analysis to establish which individual linguistic features included in the final model actually count as key features in CLIs, insofar that they actually contribute to a successful SL classification and should be included in the closer inspection using the EFA.

Table 2. Feature sets considered

Feature set	**Example**
word 1-gram	*pelaan* 'I play'
word sequential 2-gram	*pelaan_jalkapalloa* 'I play football'
word dependency 2-gram	*pelaan*HEAD_obj_*jalkapalloa*NODE
word sequential 3-gram	*Pelaan_jalkapalloa_huomenna* 'I play football tomorrow'
lemma 1-gram	PELATA
lemma sequential 2-gram	PELATA_JALKAPALLO
lemma dependency 2-gram	PELATAHEAD_obj_JALKAPALLONODE
lemma sequential 3-gram	mina_pelata_jalkapallo
part-of-speech 1-gram	VERB
part-of-speech sequential 2-gram	VERB_NOUN
part-of-speech dependency 2-gram	VERBHEAD_obj_NOUNNODE
part-of-speech sequential 3-gram	PRON_VERB_NOUN
syntax 1-gram	root
syntax sequential 2-gram	root_obj
syntax sequential 3-gram	nsubj_root_obj
word positional 1-gram	first_*pelaan*
lemma positional 1-gram	first_PELATA
part-of-speech positional 1-gram	first_VERB
syntax positional 1-gram	first_root
character 3-gram	*pel*

The second pillar of the methodological design is structured around unsupervised machine learning, whereby the key features identified by the keyness analysis are

grouped into bundles based on the inter-correlations of their frequencies. Crucially, neither the amount of these bundles nor the features that might be inter-correlated have been defined prior to the analysis. Rather, the groupings serve as a bottom-up categorization of the variation observed in the data. For this unsupervised component, we use EFA in a similar fashion to the multi-dimensional analyses of linguistic variation (Berber Sardinha & Pinto, 2019; e.g., Biber, 1988, 1989), although contrary to that tradition, the variables included in the analysis stem from the bottom up from the above-described keyness analysis. This approach includes two useful tools: first, it provides a measure of variance, the so-called eigen values, that can be used to determine the number of feature bundles to be included in the analysis. Second, once the optimal solution has been obtained, the final model also includes a measure called factor loadings that indicates to what the degree the individual linguistic features included are associated to the different factors. Together, these two can be used to model the dimensions of linguistic variation observed in the data, as well as to position each individual text on a continuum regarding each dimension (for details regarding operationalization, see below). Note, however, that as our goals profoundly differ from those of multi-dimensional analyses of linguistic variation, we deliberately refrain from calling our approach a multi-dimensional analysis, so as to avoid any confusion.

For a simplifying example of the EFA, let us assume that the frequency of nouns and adjectives is correlated so that texts that include many nouns, relative to all the texts in the data, also tend to comprise many adjectives. Furthermore, let us assume that this is inversely correlated to the frequency of verbs, meaning that texts that tend to have relatively many nouns also tend to have relatively few verbs. The EFA algorithm allows for revealing such typical correlation patterns of bundling linguistic features. What is more, EFA makes it possible to calculate for each individual text how sensitive that particular text is to each of the revealed correlation patterns, that is, to place each text on a variational continuum in relation to all the other texts. Thus, we can use EFA to identify which linguistic features contribute to grouping certain texts together and apart from other texts. We can then compare these groupings to the information we have on the SLs of the texts. In other words, we can investigate the degree to which the texts group according to their SLs, which linguistic features contribute to these groupings, and how these features correspond to the typological differences across the SLs. This, in turn, enables us to connect our observations to the hypothesized mechanisms of

salience-based variance related to CLIs in translation discussed above: predominantly gravity of the SL and connectivity between the SL and the T. Finally, analyzing the behavior of ITrs with regard to the SL-related variational dimensions make it possible to provide empirically founded, yet theoretically informed – and as such more readily generalizable – interpretations on the linguistic characteristics of such texts. What is more, the analysis of the ITrs allow for exploring magnetism of the TL inasmuch as it affects the linguistic make-up of the mediating text.

3.3 Operationalization of the study design

We conducted all the statistical analyses in the R programming environment (R Core Team, 2022). For the first phase of the keyness analysis, we used the ranger package implementation of the random forest algorithm and trained 20 parallel classifiers, one for each feature set to distinguish between the different SLs (for the algorithm, see Breiman, 2001; for the implementation, see Wright & Ziegler, 2017). Then, after obtaining the best performing feature sets and training the final classifier, we make use of the Boruta algorithm available in the Boruta package (Kursa & Rudnicki, 2010) that compares the actual features with duplicates whose frequency values have been randomly permuted. This allows us to target the subsequent analysis exclusively to those features that repeatedly outperform their randomized duplicates, irrespective of the exact data subsample. For the purposes of the present paper, these consistent features are considered the key features of SL classification in Finnish.

For the EFA, the procedure generally follows that described by Egbert and Staples (2019), and we use the functions available in the psych package. First, we ensure that the data are indeed factorable by means of the Kaiser, Meyer, Olkin (KMO) measure of sampling adequacy (for details regarding the measure, see Kaiser, 1974). Then, provided that the data prove sufficiently factorable, we decide on the optimal factor solution based on the visual exploration of the eigen values, using the screen plot to look for the so-called elbow, where the added contribution of an additional factor to capturing the overall variance is clearly smaller than that of the previous factor (e.g., Egbert & Staples, 2019, p. 102). After that, we run the final analysis and use the factor loadings from that model to calculate the factor scores for each text and, hence, to make it possible to position the texts in relation to one another according to the variation of their SL-related key features. In the analysis, we consider each feature to belong primarily to that factor where its loading is the strongest.

4. RESULTS

4.1. Keyness analysis

As mentioned above, the results of the first phase of the keyness analysis (regarding the type of feature sets that contribute to successful SL identification the most) have been reported and discussed in greater detail in I. Ivaska and L. Ivaska (2022). Fig. 1 reports the prediction accuracies of all the considered feature sets trained on train data and tested against separate test data. As a general trend, feature sets based on linguistic metadata (like syntactic functions or part-of-speech tags) seem to consistently outperform features defined based on lexical information or other formal surface structure, including word forms, lemmas, as well as strings of characters. The three best-performing feature sets include 2-grams of sequential parts-of-speech, as well as 1-grams of syntactic functions and parts-of-speech in relation to their positioning in the sentences. For the present paper, we targeted all the subsequent analyses to these tree feature sets (for a detailed discussion regarding the reliability of the classification, see I. Ivaska & L. Ivaska, 2022, pp. 386–387).

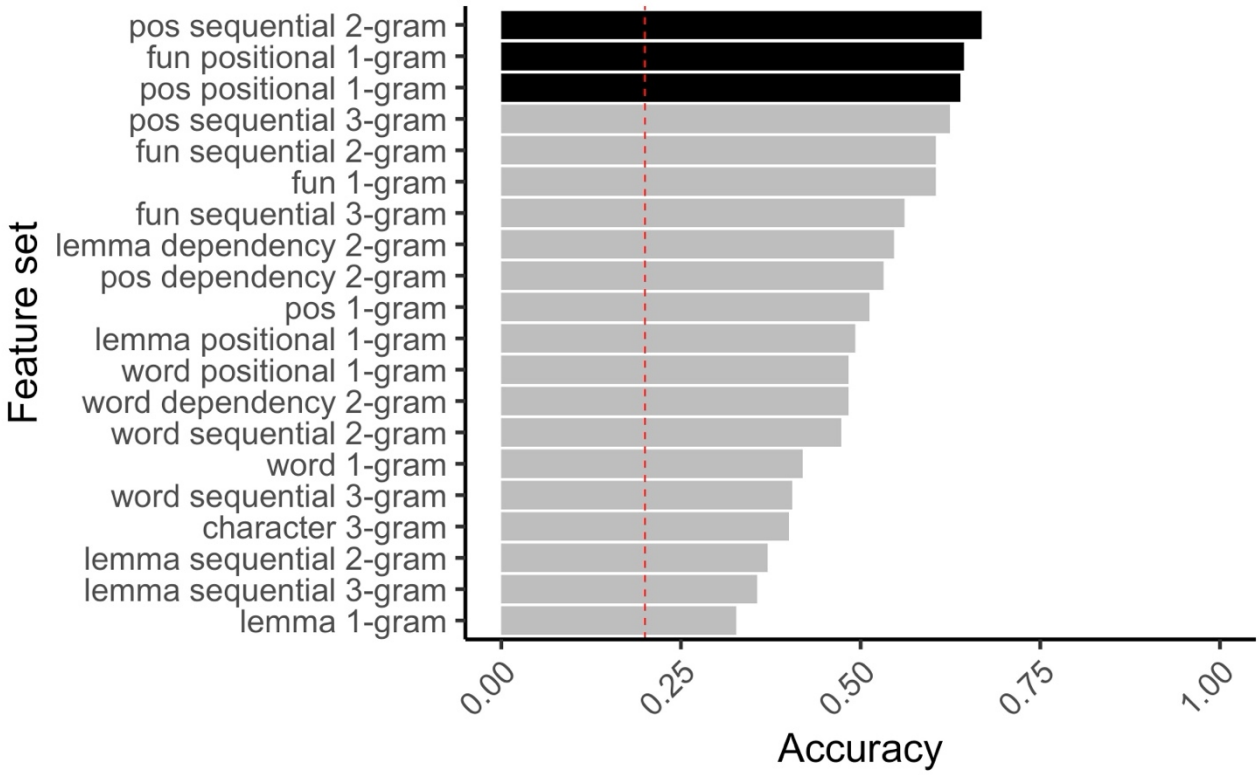

Fig. 1. Prediction accuracy of the feature sets included in the analysis. Three best performing feature sets in black and the baseline indicated by the dashed red line

The three best performing feature sets comprise all in all 363 grams that were considered as candidate key features in the keyness analysis. Out of these features, 118 were confirmed as actual key features in the keyness analysis using the Boruta feature selection algorithm. In other words, the frequencies of these 118 features were confirmed consistently useful in the task of grouping the texts translated into Finnish according to the SL from which they had been translated. Table 3 breaks down the distribution of these features across the feature sets, whereas a full list of all the key features can be found in Appendix 3. These 118 key features were used as input in the EFA, so as to reveal the latent variational dimensions that pertain to CLIs.

Table 3. Results of the keyness analysis

Feature set	Candidate key features (N)	Confirmed Key features (N)
Sequential POS 2-gram	162	50
Positional POS 1-gram	52	30
Positional syntactic 1-gram	149	38
All	363	118

4.2. Exploratory Factor Analysis: an overview

As the first step of the EFA, we validated the overall factorability of the data by means of the KMO measure of sampling adequacy. With the KMO value at 0.93, the data proved highly factorable, and we felt confident in moving on to choose a suitable amount of factors for the final model. To that end, as the eigen values visualized in Fig. 2 clearly suggest, an optimal factor model comprises four factors: up until there, each added factor clearly contributes to the explanatory power of the model, as indicated by the significant change in the eigen values. From the fifth factor and onwards, the relative proportion of additional overall variance of the data that the factor model would capture diminishes. The four-factor model obtained captures 56% of the overall variance of the data.

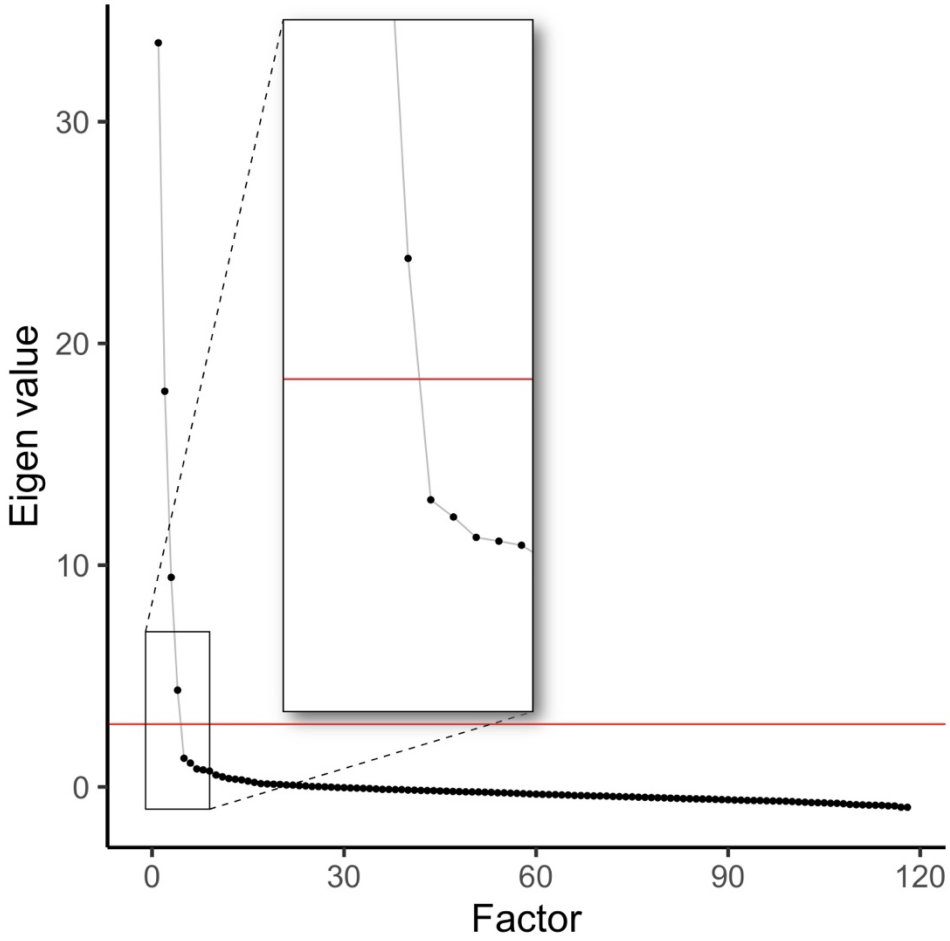

Fig. 2. Number of factors in relation to the explained variance

The final factor model provides, for each of the 118 confirmed key features, a factor loading that indicates how important that feature is for the respective factor: features with similar factor loadings – either high or low – on the same factor are likely to co-occur often in the text chunks. These feature groupings provide us a linguistically intelligible means to interpret the CLIs that take place in the translations. We will turn to those interpretations in a moment. Before that, however, we explore how the text chunks of direct translations with different SLs are positioned with regard to these factors – and how ITrs position themselves in relation to the direct translations. Hence, we calculated, for each text chunk, factor scores using the factor loadings of the final factor model. These scores make it possible to map each text chunk in terms of how it is positioned in relation to the features that characterize that respective factor. The factor scores for individual text chunks are visualized in Fig. 3. The upper panel on the left-hand side visualizes the positioning of the text chunks of direct translations with various SLs in relation to factors 1 and 2, and upper panel on the right-hand side factors 3 and 4,

respectively. The lower panels include the same data, but also the text chunks of ITrs from Greek into Finnish via various mediating languages.

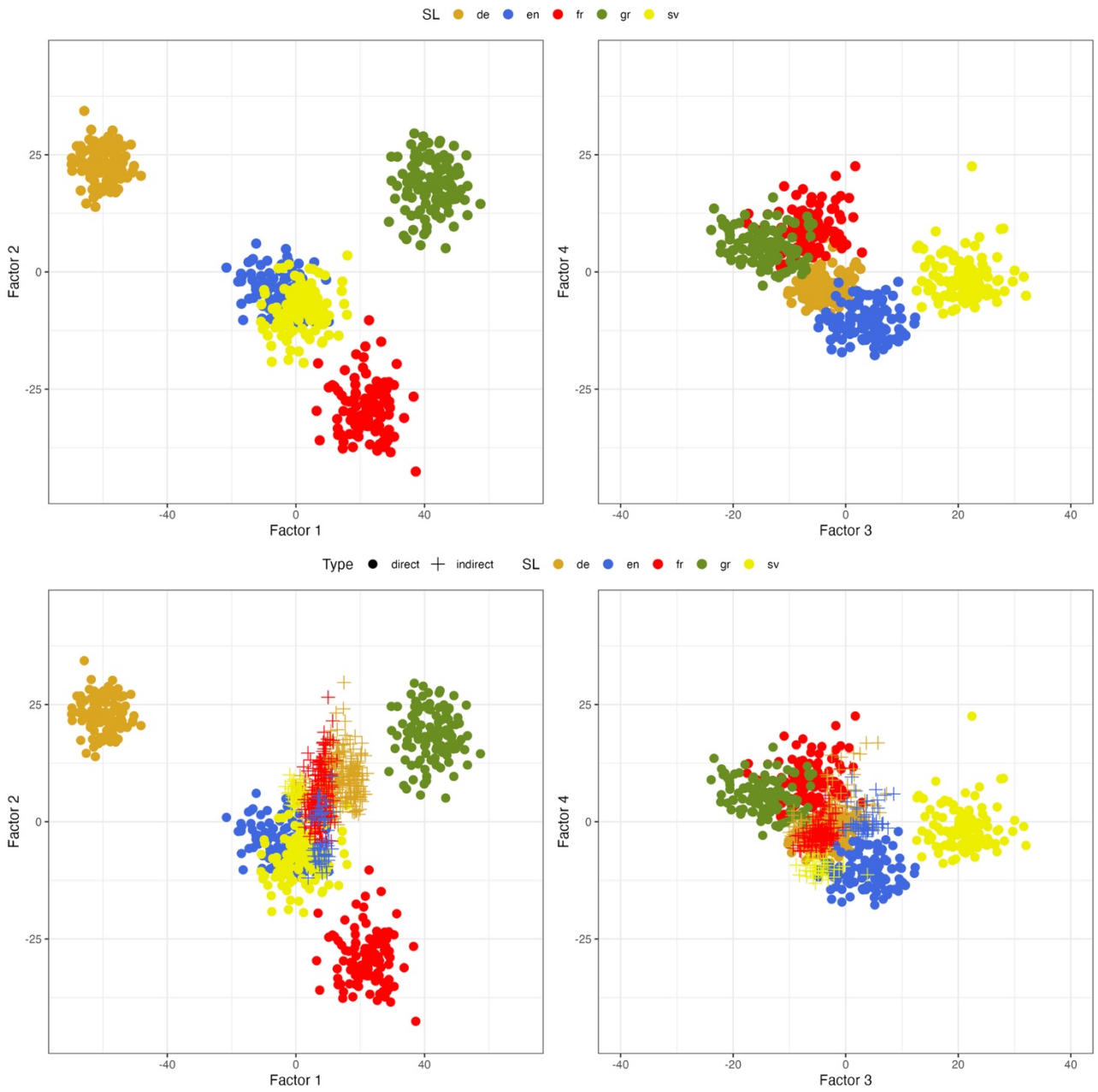

Fig. 3. Factor score mappings for each text chunk in relation to the SL they represent. Upper panels include only text chunks of direct translations, while lower panels also include indirect translations

As the mapping of factor scores of factors 1 and 2 indicate, by and large the text chunks with the same SL group together and away from the text chunks with other SLs. Most notably, factor 1 is very sensitive to the difference between translations from Greek and

translations from German, while translations from French are close to those from Greek, leaving translations from English and Swedish somewhat in the middle. What this means in practice is that those key features that have high positive loadings on factor 1 are relatively more common in translations from Greek and French than in translations From German. Similarly, features that have clearly negative loadings on factor 1 are relatively more common in translations from German than in translations from Greek and French. The same logic applies to all the factors: those key features that have high positive loadings on factor 2 are relatively more common in translations from Greek or German than those from French, with translations from English and Swedish falling again in the middle.

Before taking a closer look at the linguistic features that contribute to the factors, we also mapped the ITrs onto the above-described factor space. When comparing the text chunks of ITrs from Greek into Finnish via different mediating languages (indicated with + signs in the lower panels of Fig. 3) to the direct translations (indicated with dots in the lower panels of Fig. 3), two intertwining patterns emerge: first, when compared to the direct translations from various SLs, ITrs group relatively close to each other, rather than the respective mediating languages, irrespective of what the mediating language is. Second, ITrs group closer to the direct translations from Greek, the ultimate SL of all the ITrs, than the direct translations from various SLs do. This is the case particularly in factors 1 and 2 – which together cover 38% of the overall variance of the data of direct translations. In other words, factors 1 and 2 clearly capture the most important information for distinguishing between the different SLs of translations into Finnish, and in this respect the ITrs form a relatively uniform group that is closer to the translations from Greek than the direct translations from various SLs are. We take this to mean that features that contribute to factors 1 and 2 depict linguistic phenomena that carry over from the ultimate SL even across a mediating translation. Hence, we focus our subsequent analysis on linguistic features of factors 1 and 2 (All the factors and key features loadings are summarized in Appendix 4).

4.3. Sentence length

Out of the 118 confirmed key features, 34 contribute the most to the positive loading on factor 1 (for an exhaustive list of all features and their loadings, see Appendix 4), meaning that these features are relatively typical for direct translations from Greek and French – and to a slightly lesser degree also to ITrs from Greek irrespective of the

mediating language. Most of these key features represent clausal core constituents and their combinations: grammatical subjects and objects, as well as main verbs. What is more, several of the key features indicate that, in the translations from Greek, these features occur relatively more often in the sentence-initial or sentence-final positions. Keeping in mind that all feature frequencies have been normalized over 1,000 words, this result effectively means that in these data, a greater proportion of all words represent core constituents. In other words, sentences more often comprise only grammatical core constituents and are, thus, relatively shorter than in other data. Example (1) reflects several of those key features: it has a sentence-initial grammatical subject (*miehet* 'men') that is also a noun, a sentence-final grammatical object (*keskustelun* 'conversation'), a main verb (*virittivät* 'initiated') as the second – and the penultimate – word of the sentence, and a sequential 2-gram that consists of a verb and a noun. Similarly, example (2) also reflects some other features associated with factor 1: it has a sentence-initial pronoun (*mistä* 'about_what'), a sentence-final grammatical subject (*Fortis*), and a paratactic verb (*kysyi* 'asked') in the penultimate position of the sentence. Crucially, as the frequencies have been normalized over word count, shorter sentences result in relatively more numerous sentence-initial and sentence-final positions. What is more, in shorter sentences they are relatively more often occupied by the core constituents like grammatical subjects, objects and main verbs, that occur in almost all the sentences.

(1) Miehe-t viritt-i-vät keskustelu-n.
Men-PL.NOM initiate-PRET-3PL conversation-ACC
'The men initiated a conversation'
(Gr-Fi)

(2) Mi-stä asia-sta? kysy-i Fortis.
what-ELA thing-ELA ask-PRET.3SG Fortis.NOM
'About what? Fortis asked'
(Gr-Fi)

At the other end of the same continuum, there are 18 key features that contribute the most to the negative loading on factor 1, indicating that they are relatively more frequent in direct translations from German than in any other data subset. These key features comprise almost exclusively part-of-speech bigrams that consist of, or at least include, either adjectives or adverbs. Illustratively, example (3) includes a 2-gram of an

adverb followed by an adjective (*varmaan hassun* ‘probably funny’), and another one of two adjectives (*hassun näköisiä* ‘funny looking’). Example (4) includes a 2-gram of an adverb followed by zan adjective (*oikein hyvässä* ‘very good_in’), and example (5) two 2-grams of two adverbs (*jo kauan* ‘already long_time’ and *kauan sitten* ‘long_time ago’), respectively. Interestingly, and corroborating further the above-described interpretation of the sentence-length and minimalistic style as a phenomenon that carries over across languages, there are only two positionally defined features that load negatively on factor 1, and even they both represent adverbs in the second position of the sentence. Overall, as is the case in examples (3)–(5), too, features that are associated with factor 1 with negative loadings, depict phenomena related to describing qualitative or quantitative characteristics of people, objects or events. They occur exclusively in the middle of sentences and, hence, characterize sentences that are long in relation to sentences in the data in general.

(3) Ol-tiin me varmaan hassu-n näköis-i-ä [---].
be-PRT.PASS 1PL.NOM probably funny-GEN looking-PL-PART
‘We must have been funny looking’
(De-Fi)

(4) E-n minä itse=kään ole oikein hyvä-ssä vedo-ssa [---].
NEG-1SG 1SG self=also be.CONNEG very good-INE shape-INE
‘I’m not in a very good shape, either’.
(De-Fi)

(5) Alice kuol-i jo kauan sitten, Alice sano-i.
Alice.NOM die-PRET.3SG already long_time ago Alice.NOM say-PRET.3SG
‘Alice died already a long time ago, said Alice’
(De-Fi)

4.4. Sentence-initial adverbs and sentence-final specification

Contrary to factor 1, positively loading features on factor 2 highlight features that make texts translated from German and Greek resemble each other – and set them apart from translations from French. There are altogether 15 key features that contribute the most on the positive side of factor 2. Interestingly, many of these positively loading features reflect adverbs but, contrary to the negatively loading features of factor 1, these adverbs occur almost invariably in the sentence-initial positions. Examples (6)–(9) all depict a sentence where an adverb occupies either the first or the second position of the sentence, or them both, such as *ikään kuin* (‘as if’) in example (6). Often, they co-occur

with conjunctions, either when a conjunction links two adverb-like elements, as *Niin tai näin* (lit. ‘so or so’) in example (7), or *hitaasti ja lyhyin aironvedoin* (‘slowly and with short pulls on oars’) in example (8). At times, a conjunction also serves in an adverb-like function itself, like *ja* (‘and’) in the beginning of example (9).

(6) Ikään kuin yhä nyhjöttä-isi kiltisti sohvan_nurka-ssa-an [---]
as like still curl_up-COND.3SG obediently couch_corner-INE-POSS3SG
‘As if one were still obediently curling up in the corner of their sofa [---]
(De-Fi)

(7) Niin tai näin, hän astu-isi näyttämö-lle joka tapaukse-ssa.
so or so 3SG step-COND.3SG stage-ALL any case-INE
‘Be it as it may, (s)he would in any case enter the stage.’
(De-Fi)

(8) Hitaasti ja lyhy-i-n airon_vedo-i-n souta-e-n he peila-si-vat
slowly and short-PL-INSTR oar_pull-PL-INSTR row-INF2-INS 3PLmirror-PRET-3PL
ranto-j-a
shore-PL-PART
’Slowly and with short pulls on the oars, they mirrored the shores’
(Gr-Fi)

(9) Ja taas ikkunaverho-t heilahtel-i-vat.
and again window_curtain-PL.NOM swing-PRET-PL3
‘And again, the window curtains were swinging.’
(Gr-Fi)

In addition to the sentence-initial adverbs, there are also two sentence-final functions that set translations from German and Greek apart from the other SLs: vocative referring to people as well as appositional modification of a preceding noun phrase (where the modifier and the modified are interchangeable in order). *täti* (‘aunt’) in example (10) reflects a sentence-final vocative, as indicated by the preceding imperative verb form *kerro* (‘tell’). Example (11), in turn, exemplifies an appositional use where the noun phrases *Elben rannalta* (‘from the shore of Elbe’) and *Brühlin puiston terasseilta* (‘from the terraces of the Brühl park’) have an appositional relationship.

(10) Kerro minu-lle kaikki, täti.
tell.IMP.2SG 1SG-ALL all aunt.NOM

‘Tell me everything, aunt.’
(Gr-Fi)

(11) Kirjoittel-i-n häne-lle Elbe-n ranna-lta, Brühli-n
write-PRET-1SG 3SG-ALL Elbe-GEN shore-ABL Brühl-GEN
puisto-n terasse-i-lta
park-gen terrace-pl-abl
‘I wrote him/her from the shore of Elbe, the terraces of the Brühl park’
(De-Fi)

4.5. Auxiliary verbs

Finally, the key features that contribute to the negative loadings in factor 2 – and that are relatively more frequent in translations from French than in other SLs – portray different uses of verbs that have been annotated as auxiliaries. There are in total 16 such key features, and they typically represent either 2-grams with auxiliaries, or sentential positions they typically occupy. These include compound tenses as in example (12), where the verb *oli* (‘had’) is annotated as an auxiliary both in terms of the part-of-speech it reflects and in the syntactic function it serves, as well as in copula clauses like example (13). Much like in the positively loading features of factor 1, these mostly occur in short sentences. This is on the one hand indicated by their relatively frequent occurrence in the beginning of the sentence, especially as the second word, preceded by the grammatical subject, as in examples (12) and (13). On the other hand, they also occur at the end of the sentence as the penultimate word, followed by either the main verb of an intransitive clause, as *laihtunut* (‘lost_weight’) in example (12), or a predicative element of a copula clause, like *rasittavaa* (‘tiring’) in example (13). In other words, the distinguishing aspect from factor 1 is the auxiliary use. While systematic distributional analyses of individual key features go beyond the scope of this study, we also made an interesting qualitative observation that many of the auxiliaries seem to be in the past tense, indicating either pluperfect in intransitive clauses (as in example (12)), or simple past in copula clauses (as in example (13)).

(12) Hän oli laihtu-nut.
(S)he.NOM be.PRET.3SG loose_weight-PTCPL2.SG
‘(S)he had lost weight’
(Fr-Fi)

(13) Se oli rasittava-a.
it be.PRET.3SG tiring-PART
'It was tiring'
(Fr-Fi)

Next we will move on to discuss possible underlying reasons for these attested CLIs, addressing these issues especially from the point-of-view of ITrs and the possible theoretical implications that these results may have as regards the underlying processes related to such translations.

5. DISCUSSION

In this paper, we have studied the SL classification of translated texts, with two particular interests. First, contrary to most of the earlier research on the topic that has primarily focused on evaluating the overall success rate of such classification tasks, we wanted to explore in detail the linguistic nature of the phenomena that contribute to successful classification when using supervised machine learning techniques. Second, we were curious to understand better the linguistic reality of ITrs – translations made from translations – with regard to their positioning in relation to different SLs. The topic of ITrs can generally be considered highly under-researched, and the few earlier studies (I. Ivaska & L. Ivaska, 2022; Rabinovich et al., 2017) have mostly focused on the observation that the ultimate SL of the translation has so strong an impact on the translation process that it can be identified even via a mediating translation in a different language. Our ultimate goal here was to link both the SL classification and the status of ITrs to the theoretical models of translation as a cognitive activity, and to generate theoretically motivated targeted hypotheses for the future research to explore.

On a general level, our results corroborate the earlier research, in so far as translations from varying SLs can indeed be distinguished in a fairly reliable fashion even in the case of Finnish. The linguistic phenomena that contributed to the successful classification were generally structural rather than lexical in nature, and the sentence-level positioning of structural features was proven very important. In the more detailed analysis, we sorted out those individual features that actually contribute significantly to successful classification and grouped them into inter-correlated bunches that can be expected to reflect latent patterns. We identified two separate continuums, and hence,

four inter-correlated bunches of features that represent the four ends of these two continuums.

We interpret the first continuum to reflect differences in sentence length: translations from Greek and from French seem to generally comprise shorter sentences – especially when contrasted with translations from German – represented in our data by relatively higher frequencies of grammatical core constituents in the sentence-initial and sentence-final position. While this observation could be superficially related even to differences in the constituent order, our analysis revealed that all core constituents (most notably grammatical subjects and grammatical objects) occurred in translations from these languages relatively more frequently both in sentence-initial and sentence-final positions. This indicates that there are overall relatively more core constituents and sentences, rather than that their positioning would set the data subsets apart. This interpretation was further corroborated by the observation that translations from German were characterized by relatively more frequent use of non-core modifying arguments like adjectives and adverbs that occurred in non-initial and non-final positions – effectively indicating longer sentences and more detailed qualitative and quantitative description of people and events. Centrally to the goals of this article, these interpretations point to phenomena that are stylistic and primarily have to do with genre conventions and aesthetic ideals of the respective lingua-cultures – or at least we were unable to link them to any systemic typological differences across these languages.

As far as Halverson's (2017) gravitational pull model is concerned, we consider the results to reflect primarily gravity, where a feature of the SL (here, sentence structure) remains visible even in translations. Had we looked solely at direct translations, it would not be possible to tease apart the effects of the three forces of GP in such a bottom-up, corpus-driven research design. Crucially, the inclusion of ITrs made it possible to distinguish between the effects of gravity and connectivity as forces that set texts with different SLs apart from one another: if successful SL classification would be primarily due to connectivity, the ITrs with different mediating languages should be positioned further away from each other, as the entrenched features would be specific to language pairs. As this was not the case, we interpret this to mean that the classification is indeed primarily due to gravity. As for magnetism, it probably also played a role, indicated by the fact that ITrs did not group closer to the direct translations from Greek, the ultimate SL of all the ITrs. In sum, the gravity-oriented hypothesis we wish to

generate based on these exploratory observations is as follows: the proportional density of grammatical core constituents in texts is sensitive to the genre conventions and stylistic ideals of the lingua-culture of the original text, and these conventions and ideals are visible in translated texts via sentence length and the proportional frequency of core constituents within sentences. To test this hypothesis, a confirmatory study should contrast in these respects non-translated texts of a superficially comparable genre (such as literature written in the format of novel) from different lingua-cultures, as well as translated texts from the respective lingua-cultures into a common TL. Confirming this hypothesis would require that the translations diverge from each other and that the divergence is always towards the SL, otherwise the hypothesis can be rejected or at least tweaked according to the results. Looking at several TLs would obviously strengthen the argument even further, as would the use of ITrs as controlling devices.

Contrary to the first continuum, we interpret the **second continuum** to reflect two unconnected phenomena that may both be motivated by systemic typological differences across the SLs. The use of sentence-initial adverbs was relatively more common in translations from Greek and from German, which we interpret to potentially reflect differences in the characteristic constituent order of the respective languages. To this end, the positioning of adverbs is often described as a phenomenon where several options are possible but where the different options may be used to convey diverging meanings, or there may otherwise be probabilistic – but not categorical – preferences for one option over another (e.g., on Finnish, see Hakulinen et al., 2004, p. §870). Here, too the fact that the ITrs group closer to the direct translations from Greek than the translations from the other SLs corroborates an interpretation whereby the constituent order of the ultimate SL is retained in translations irrespective of the TL. This constitutes a clear argument in favor of a translational phenomenon related to the SL gravity, and the wide variation of adverb-related constituent order both within and across individual languages suggests that they constitute a good candidate for probabilistic, typologically motivated CLIs. Hence, we would like to put forth the following gravity-oriented hypothesis: the sentence-level positioning of adverbs in translated texts is sensitive to the patterns typical for the SL. To confirm this hypothesis, a confirmatory research design could make use of a corpus of translated texts of a given TL, and comparable corpora of non-translated texts of the respective SLs. Comparing the constituent orders of different types of adverbs allows for either confirming a

correlation between the preferred order of the SL to be visible in the TL, or alternatively rejecting it. Here, too, including ITrs in the design would provide a means to measure whether multiple rounds of translation reduce the effect of such a CLI.

Finally, we interpret the other end of the second continuum to also reflect typologically motivated CLIs, but unrelated to the constituent order discussed above. Translations from French included relatively more occurrences of verbs annotated as auxiliaries, in particular the verb *olla* 'to be' that is used in Finnish both as a copula verb and as the finite verb of the present perfect and pluperfect compound tense expressions. Interestingly, all the SLs included in the research design have both inflectional past tenses and compound past tenses, just like Finnish, but French is the only one where the formal properties of different past tense expressions differ typologically from Finnish. In Finnish, just like in English, German, Greek, and Swedish, the aspectually perfective past tense is expressed with verbal inflection (like English *I travelled*) and the aspectually imperfective past tense is expressed with a compound expression of a finite auxiliary verb together with the main verb (e.g., *I have travelled*). In French, on the contrary, it is the aspectually perfective past that is typically expressed with a compound expression (e.g., *J'ai voyage* 'I travelled') and the aspectually imperfective past that is expressed with a verbal inflection (e.g., *Je voyagais* 'I have travelled', 'I was travelling'). Hence, we interpret the observed difference to reflect a CLI that stems from the distributional difference between inflectional past tense expressions and compound past tense expressions between French and the other SLs. Here, too, the fact that the ITrs via French adhere to and group with the ultimate SL (Greek) rather than with direct translations from French could be seen to indicate that the effect is bi-directional: the French translations of Greek literature differ from French translations of, say, Italian literature, just like Finnish translations of French literature differ from Finnish translations of Greek literature. Then, when these originally Greek texts are translated further from French into Finnish, the outcome is, again, closer to the texts translated directly from Greek into Finnish.

Based on these observations and interpretations, we would like to propose the following hypothesis: in the event of formal constructional equivalents between SL and TL with fine-grained functional differences that are distributional rather than categorical in nature, the form–function distribution in translated texts diverges from non-translated texts towards the distribution witnessed in the SL. Then, as far as ITrs are concerned,

the distribution of the source text might actually be retained irrespective of the mediating language if the ultimate SL and the ultimate TL are alike. In the case of past tense expressions, two alternative research designs come to mind: one could either use a sentence-aligned parallel corpus of translations from French to Finnish and Finnish to French, and to compare the formal distribution of the translational equivalents of both inflectional and compound tense constructions. Alternatively, one could include comparable corpora of translated texts of varying SLs with several instances of both form–function pairings (e.g., translations from French and Italian into Finnish to be contrasted with translations from German and English into Finnish): to confirm the hypothesis, the translations from SLs with similar form–function pairings should resemble each other more in terms of the tense distribution than the translations from SLs with dissimilar form–function pairings. Finally, both these designs could be enriched with an indirect component: investigating the changes in the distributions of tenses in different phases of the translation chain can reveal the sensitivity of translation process to such distributional phenomena.

Note that both two phenomena that we hypothesized to reflect systemic typological differences are distributional and probabilistic in nature, not categorical. Provided that these hypotheses can be confirmed in the future confirmatory studies, we take this to support the idea discussed under the term ‘default translation’ (e.g., Halverson, 2019) in that the gravity of the SL is more likely to take place in linguistic phenomena that have multiple possible equivalents in both the SL and the TL, rather than in such phenomena where the SL and TL categorically diverge. While categorical differences or unique items could also lead to distributional differences in a given linguistic phenomenon (as shown e.g., in Hareide, 2016 regarding the presence/absence of a grammatical feature across SLs), at least in the context of the present study we could not link any of the most contributing differences to such phenomena. This may also relate to a limitation regarding the generalizability of the present study: the data of some SLs, and especially some language combinations in ITrs, were limited in size, and we cannot exclude that some of the results are data-specific in nature. We have tried our best to remedy this by careful research design and data partitioning in the model building (for details, see I. Ivaska & L. Ivaska, 2022, pp. 377–378), but acknowledge that data sparsity and limited generalizability is an ever-present issue within corpus linguistics (Leech, 2006), and even more so in an under-researched topic like indirect translation.

6. CONCLUSION

As far as the linguistic phenomena that contribute to the SL classification are concerned, we identified two distinct phenomena that reflect CLIs: on the one hand, there are stylistic typicalities that characterize language use in a certain lingua-culture. Such influences were visible via the typical sentence structures in translations from various SLs, and ITrs reflected the textual patterns of the ultimate SL rather than that of the mediating language. On the other hand, systemic typological differences across the languages involved also contributed to differentiating between the SLs involved. Here, it seems that typological differences that are distributional rather than categorical in nature are more prone to CLIs: all the differences that we hypothesized to be typologically motivated reflect phenomena in which various options are possible in the languages studied, and so it is rather the relative probability between these options that distinguishes translations from different SLs from each other.

In relation to the above-discussed results of this exploratory study and their linking with the theoretical nature of CLIs in translations, we want to point out the profound difference between CLIs that reflect genre conventions and aesthetical norms, on the one hand, and those that stem from typological differences between the different SLs, on the other. While both may well be important in the engineering-like task of automatically classifying translations according to their SLs, they differ very much in what they tell us about translating: both may stem from a general cognitive strategy that it is probably easier and less taxing for the translator to preserve the formal properties of the ST, yet they are theoretically separate constructs. The stylistic conservatism regarding the sentence boundaries reflects a holistic text-level phenomenon, and it can as such be linked to discussions related to layout and paratextual phenomena. What they have in common is that they are likely to reflect cultural affinity between the publishing contexts. The typological differences, on the contrary, stem from the linguistic diversity and similarity across languages – and while such typological groupings may align with cultural affinity, they do not necessarily do so.

REFERENCES

Assis Rosa, A., Pięta, H., & Bueno Maia, R. (2017). Theoretical, methodological and terminological issues regarding indirect translation: An overview. *Translation Studies*, *10*(2), 113–132. https://doi.org/10.1080/14781700.2017.1285247

Berber Sardinha, T., & Pinto, M. V. (Eds.). (2019). *Multi-DimensionalMultidimensional analysis: Research methods and current issues*. Bloomsbury Academic.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1989). A typology of English texts. A typology of English texts. *Linguistics*, *27*(1), 3–44. https://doi.org/10.1515/ling.1989.27.1.3

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Čermák, F., & Rosen, A. (2012). The case of InterCorp: A multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427. https://doi.org/10.1075/ijcl.17.3.05cer

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

Fabrigar, L. R. (2012). *Exploratory factor analysis*. Oxford University Press.

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.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., & Alho, I. (2004). Iso suomen kielioppi. *Suomalaisen Kirjallisuuden Seura*. http://scripta.kotus.fi/visk URN:ISBN:978-952-5446-35-7

Halverson, S. L. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical Translation Studies: New methodological and theoretical traditions* (pp. 9–46). De Gruyter.

Halverson, S. L. (2015). Cognitive Translation Studies and the merging of empirical paradigms: The case of ‘“literal translation.”’ *Translation Spaces*, *4*(2), 310–340. https://doi.org/10.1075/ts.4.2.07hal

Halverson, S. L. (2019). 'Default' translation: A construct for cognitive translation and interpreting studies. *Translation, Cognition & BehaviorTranslation, Cognition & Behavior*, *2*(2), 187–210. https://doi.org/10.1075/tcb.00023.hal

Hareide, L. (2016). Is there Gravitational Pull in translation? A corpus-based test of the Gravitational Pull Hypothesis on the language pairs Norwegian–Spanish and English–Spanish. In M. Ji, M. Oakes, L. Defeng, & L. Hareide (Eds.), *Corpus methodologies explained. An empirical approach to translation studies* (pp. 188–231). Routledge.

Islam, Z., & Hoenen, A. (2013). Source and translation classification using most frequent words. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, (pp. 1299–1305). https://www.aclweb.org/anthology/I13-1185

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., Bernardini, S., & Ferraresi, A. (forthcoming). 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 H. Kotze & B. van Rooy (Eds.), *Constraints on language variation and change in complex multilingual contact settings*. John Benjamins Publishing Company.

Ivaska, I., & Ivaska, L. (2022). Source language classification of indirect translations. *Target,* [Special Issue]*: What Can Indirect Translation Research Do for Translation Studies?*, *34*(3), 370–394. https://doi.org/10.1075/target.00006.iva

Ivaska, L. (2019). Distinguishing translations from non-translations and identifying (in)direct translations' source languages. In J. H. Jantunen, S. Brunni, N. Kunnas, S. Palviainen, & K. Västi (Eds.),. *Proceedings of the Research Data and Humanities (RDHum) 2019* Conference: Data, Methods and Tools, *2019* (pp. 125–138). University of Oulu. https://www.oulu.fi/sites/default/files/content/ProceedingsStudiaHumanioraOuluensia17.pdf

Ivaska, L. (2020). A Mixed-methods approach to indirect translation: A case study of the Finnish translations of modern Greek prose 1952–2004. [University of Turku]. https://www.utupub.fi/handle/10024/150755

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, (pp. 133–142).

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, (pp. 1318–1326). http://www.aclweb.org/anthology/P11-1132

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.

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

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, Articles*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Langacker, R. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

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). Brill.

Lefer, M.-A., & De Sutter, G. (2022). Using the Gravitational Pull Hypothesis to explain patterns in interpreting and translation: The case of concatenated nouns in mediated European Parliament discourse. In M. Kajzer-Wietrzny, A. Ferraresi, I. Ivaska, & S. Bernardini (Eds.), *Mediated discourse at the European Parliament: Empirical investigations* (pp. 133–159). Language Science Press. https://doi.org/10.5281/ZENODO.6977046

Lynch, G., & Vogel, C. (2012). Towards the automatic detection of the source language of a literary translation. *Proceedings of the COLING 2012*: [Posters] (pp. 775–784). https://www.aclweb.org/anthology/C12-2076

Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 65–82). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.07mau.

Neumann, S. (2014). Contrastive register variation: A quantitative approach to the comparison of English and German. *De Gruyter Mouton*.

Pięta, H., Ivaska, L., & Gambier, Y. (2022). What can research on indirect translation do for Translation Studies? *Target*, 34(3), 349–369. https://doi.org/10.1075/target.00012.pie

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.Rr-project.org/

Rabinovich, E., Ordan, N., & Wintner, S. (2017). Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 530–540). https://doi.org/10.18653/v1/P17-1049

St. André, J. (2020). Relay. In M. Baker & G. Saldanha (Eds.), *Routledge encyclopedia of translation studies* (3rd ed., pp. 470–473). Routledge.

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

Toury, G. (2012). *Descriptive translation studies – And beyond*: (rev. edition.). John Benjamins Publishing Company. http://ebookcentral.proquest.com/lib/kutu/detail.action?docID=1053083

Ustaszewski, M. (2021). Towards a machine learning approach to the analysis of indirect translation. *Translation Studies*, 14(3), 313–331. https://doi.org/10.1080/14781700.2021.1894226

Winter, B. (2020). *Statistics for Linguists: An introduction using R*. Routledge. https://doi.org/10.4324/9781315165547

Woodstein, B. J. (2022). Translation and genre. *Cambridge university press*.

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), 1–17. https://doi.org/10.18637/jss.v077.i01