Nor Jnl Ling 40.2, 225–253 © Nordic Association of Linguists 2017 doi:10.1017/S0332586517000129

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

Learner language morphology as a window to crosslinguistic influences: A key structure analysis

Ilmari Ivaska & Kirsti Siitonen

The study of crosslinguistic influences (CLI) has proven that morphosyntactic features exhibit CLI. Technical development and novel resources have enabled detection-based approaches, where potential CLI are revealed based on their observed frequencies and on differences between learners with different language backgrounds. The two research questions are as follows: (i) How construction-specific typological (dis)similarities between L1 and L2 affect the frequencies of linguistic features? (ii) Can such (dis)similarities be detected by comparing feature frequency data of L2? The data come from the International Corpus of Learner Finnish, and the methodology applied is the key structure analysis. The results support the applicability of the method: they show that constructional similarities may trigger CLI construction by construction, irrespective of the general similarities or genealogical categorizations. The results further imply the importance of controlling the genre-related and topical variation to account for skewed nature of the data when dealing with naturally occurring learner language data.

Keywords: construction grammar, corpus-driven approach, crosslinguistic influences, detection-based approach, Finnish as a second language, key structure analysis, learner corpus research

Ilmari Ivaska, Department of Interpretation and Translation, University of Bologna, Corso della Repubblica 136, Forli (FC), Italy. ilmari.ivaska@unibo.it
Kirsti Siitonen, Department of Finnish and Finno-Ugric Languages, FI-20014 University of Turku, Finland. kisiito@utu.fi

1. INTRODUCTION

Learners of second or foreign languages often end up making comparisons between the languages they know. Similarly, language teachers often find themselves reflecting upon learners' earlier language repertoire. The nature and the extent of influences from one language to another (e.g. from learner's first language L1 to the studied language L2) has interested second language acquisition (SLA) research throughout its existence (e.g. Odlin 1989, Jarvis & Pavlenko 2008). As Jarvis (2000) points out in his seminal work, the nature and even the definition of crosslinguistic influences (CLI) has remained somewhat unclear until recently. Jarvis' (2000, refined in Jarvis

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

2010) introduction of the systematic criteria for evidence of CLI has, however, led to an increased transparency and comparability of the results. The four cornerstones of this framework include statistically significant correlations between the speaker groups of certain L1s and the use of some features in the L2, so that the speakers of the same L1 behave similarly in the L2 whereas speakers of other L1s behave differently in the L2. Additionally, the L2 behavior of these two (or more) groups should align with the constructional nature of the studied linguistic feature in the different L1s in question, so that the different L1 systems corroborate the grouping. According to Jarvis (2010), meeting these conditions constitutes a comparison-based argument of CLI.

Technological advancements and the increased accessible computational resources together with greater availability of learner corpora have influenced also the study of CLI within the last ten years, and Finnish as L2 is not an exception (for an overview of some recent corpus approaches to CLI, see Helland Gujord et al. 2015, for an overview on Finnish learner corpora, see Jantunen & Pirkola 2015). Methodologically, many novel contributions center around the so-called corpus-driven approaches: a somewhat heterogenous array of different techniques which typically allow the researcher to refrain from choosing the features to be studied based on intuition and subjective evaluation, and instead support using computational techniques and the data at hand to lead the inquiry by identifying patterns of language use that are somehow characteristic or uncharacteristic to the data (e.g. Tognini-Bonelli 2001, Scott & Tribble 2006). The detection-based approach to CLI represents one such line of research. This approach, as described in Jarvis (2010) and Jarvis (2012), resembles in many ways the first part of the comparison-based argument, as it relies on the homogeneity of linguistic behavior in L2 among speakers of certain L1 together with the heterogeneity of linguistic behavior in L2 between the speakers of different L1s. As Jarvis (2010:183) points out, the premises of constituting a detection-based argument do, however, differ from those of the comparison-based argument – essentially in that the primary goal is to find the statistically best linguistic predictors of the L1 of any given text in the dataset at hand, which can then be used to interpreting the nature of the CLI involved (for examples of detection-based approaches, see e.g. Mayfield Tomokyio & Jones 2001, Koppel, Schler & Zigdon 2005, Wong & Dras 2009, Jarvis 2011, Jarvis & Crossley 2012, Pepper 2012, Ivaska 2015b).

In this paper we examine how form- or construction-specific typological (dis)similarities between L1 and L2 affect the frequencies of linguistic features in L2. This study is partially methodological, as we approach the question from a corpus-driven point of view. In other words, we are interested to see, whether (dis)similarities between L1 and L2 can be detected by comparing the feature frequency data of written Finnish as L2. Our research questions are following: (i) How form- or construction-specific typological (dis)similarities between L1 and L2 affect the frequencies of

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

linguistic features in L2? (ii) Can such (dis)similarities be detected by comparing feature frequency data of written L2 Finnish? Our data come from the International Corpus of Learner Finnish (Jantunen 2011), and we follow a step-wise corpus-driven methodological procedure called key structure analysis (for an overview of the procedure, see Ivaska 2015a and see Section 3.2 below). The results will shed new light on the kinds of constructional features that may be subject to CLI in Finnish as L2 in particular, as well as in any L2 in general. Additionally, we will also point out some data- and method-related issues that have been partially overlooked by earlier corpus studies on CLI.

This article is structured as follows: Section 2 will give a brief theoretical overview on the study of CLI and the standpoint chosen here. It also discusses some earlier results on CLI in L2 Finnish. Section 3 introduces the dataset of the present study and discusses the underlying rationale for the choices. It also gives an overview of the methodological framework and introduces the actual implementations of the methods chosen. We then report our findings in Section 4, and finally discuss them in Section 5 together with some comments regarding the applicability of the methodological choices in relation the study of CLI and crosslinguistic similarities in general.

2. THEORY AND LITERATURE REVIEW

2.1 Crosslinguistic influences as a phenomenon and an object of study

Language learners' earlier knowledge of languages has been recognized as a potential factor in SLA throughout the history of the field, since the early contrastive studies. The importance of CLI as a factor in SLA has varied considerably (see e.g. Ringbom 1987, 2007; Odlin 1989), as have the interpretations regarding the nature of these influences (for a detailed discussion, see e.g. Jarvis 2000:248–266) but, as Kaivapalu & Martin (2014:286–287) put it,

> [t]heories aside, language learners and teachers have always known that a closely related TL, with a lot of similarity with the L1, is easier and faster to learn than a more distant one. They have also noticed that some errors are typical of learners with a given L1, while those with another L1 will hardly ever commit them.

The CLI as a phenomenon is, however, of a multi-faceted nature, so that the systemic features of a given language are in interplay with language users' conceptualizations in general as well as their individual choices (Jarvis & Pavlenko 2008:13). This can be described as a varying degree of congruence between the actual similarity

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

(i.e. typological proximity) and the perceived and assumed similarity (i.e. the way the language users perceive the similarity and operationalize it in their own language production) (for discussion on the nature of these constructs, see e.g. Kaivapalu & Martin 2014:289–294). Kaivapalu & Martin note that '[t]he degree of actual similarity can be based on typological research while the fuzzy concepts of perceived and assumed similarity belong to the field of psycholinguistics' (ibid. 285). The comparison-based approach to CLI (Jarvis 2000, 2010) takes into account both these mechanisms, as typological similarities and differences between languages constitute two of its four premises, while the two other premises are related to L2 production, which is affected primarily by the perceived and assumed similarity (Ringbom & Jarvis 2009:106–107).

Stating a comparison-based argument for CLI requires that all its four conditions – similarities between a given L1 and the studied L2, differences in other L1s and the studied L2, similarities in L2 behavior among people with a shared L1 background, and differences in L2 behavior between people with different L1 backgrounds – are fulfilled simultaneously. This may sometimes lead to omitting such cases of CLI where the perceived and the assumed similarity are incongruous. The detection-based argument is concerned with a more data-driven approach to CLI (Jarvis 2010, 2012), and so it only takes into account the product-related premises – similarities in L2 behavior among learners with a shared L1 background and differences in L2 behavior between learners with different L1 backgrounds (e.g. Aarts & Granger 1998, Mayfield Tomokyio & Jones 2001, Jarvis 2011, Pepper 2012, Ivaska 2015b). In other words, the linguistic systems of the different L1s are not analyzed, which can, in turn, lead to detecting also more elusive forms of CLI of that are possibly related to the differences between the actual and perceived similarities. Very often the analysis then stops in simply stating the nature of the observed difference in frequency between the different L1 datasets (e.g. Pepper 2012) or making interpretations on the reasons without any more fine-grained analysis of the data or systematic comparison between the different L1s (e.g. Wiersma, Nerbonne & Lauttamus 2011). In this study, we follow and refine the model of Ivaska (2015b), so that a step-wise data analysis can link the assumed and the actual similarities to each other, so that findings based on a detection-based approach can be interpreted and then either rejected or validated by using a more detailed analysis of the detected features together with a strictly focused typological comparison, ultimately leading to a strong, comparison-based argument for CLI.

2.2 Construction-specific crosslinguistic influences

Studies on CLI have often focused on lexical and phonological phenomena, while morphosyntactic features have generally gained less attention, and sometimes even an almost categorical refusal of the possibility of such influences (e.g. Dulay, Burt

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

& Krashen 1982; for exceptions, see e.g. Selinker & Lakshamanan 1992, Jarvis & Odlin 2000). This is arguably due to the field's general focus on languages with more limited morphological inventories (Kaivapalu & Martin 2014:288). As Jarvis & Odlin (2000) point out, the morphological system of learners' L1 influences the choices made in the L2 production even though the nature of the effects may depend on the typological nature of the construction in question. In our opinion, this observation does support the applicability of the Construction Grammar (CG, see e.g. Goldberg 1995, 2006) as a way to define the objects of study. CG sees linguistic systems as consisting of constructions, repeated combinations of form, meaning and use. The constructions can vary in their level of abstraction, ranging from single free morphemes, such as Finnish word *lintu* 'bird', to bound morphemes, such as Finnish inessive *ssA* 'in', and to fully schematic argument structures, such as Finnish transitive construction $NP_{NOM} + V + NP_{PART/GEN/NOM}$. Such constructions are often thought of as being learned in a usage-based manner bottom–up (e.g. Tomasello 2003), and thus they are in the system-level language-specific (e.g. Croft & Cruse 2004:291–292). From the point of view of CLI, then, such constructions – the combinations of forms, meanings, and their typical uses in the languages involved – can be seen as the units that may influence linguistic choices across languages, construction by construction and irrespectively of the genealogical categorization and the amount of other constructional similarities between the languages observed.

Studies focusing on languages with rich morphological systems, such as Finnish, have in fact proven that overt morphological features and the constructions they represent or in which they occur do also exhibit CLI in general, and particularly the transferability of constructions that are perceived similar. A seminal work on this front is Kaivapalu's (2005; see also Kaivapalu & Martin 2007) study on learning plural inflection of Finnish nouns. Kaivapalu shows that L1-Estonian learners of L2-Finnish make use of the constructional similarity between the two languages when compared to L1-Russian learners of L2-Finnish. On a similar vein, Spoelman (2013) studies CLI in the use of partitive case in L2-Finnish. On the one hand, the results show that L1-Estonian learners generally have more target-like use of the partitive case across the different constructions in which it occurs, when compared with L1-Dutch and L1-German learners of L2-Finnish. On the other hand, Spoelman (ibid.) points out that situations where the constructions with partitive case have formally similar variants in Finnish and in Estonian but where their functional distribution diverges, CLI cause non-targetlike behavior in L1-Estonian learners' production which is not present in the production of the two other L1-groups. Finally, Ivaska (2015b) studies CLI in advanced L2-Finnish from a detection-based point of view and ends up with three constructions of structurally very different nature that exhibit potential CLI. Ivaska (ibid.) analyzes the use of these constructions – coordinating conjunctions, distributional differences between different tenses and special listing

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

constructions – and the differences between L1-Czech, L1-Hungarian, L1-Japanese, L1-Lithuanian, and L1-Russian learners of Finnish. Ivaska's study is more concerned with the applied methodology, and, thus, it relies solely on the detection-based argument without complementing it with typological analysis of the respective languages.

3. DATA AND METHODOLOGY

3.1 Data: International Corpus of Learner Finnish

Our data are part of the International Corpus of Learner Finnish (ICLFI; Jantunen 2011). For our purposes we have selected six L1-specific subcorpora: L1-Chinese (L1-zh),[1] L1-Estonian (L1-et), L1-German (L1-de), L1-Polish (L1-pl), L1-Russian (L1-ru) and L1-Swedish (L1-sv) to represent L1 backgrounds with varying morphological and syntactic typology.[2] While Finnish belongs to the Finnic genus of the Uralic language family, Chinese languages belong to the Sino-Tibetan language family, Estonian belongs to the Finnic genus of Uralic language family, German and Swedish belong to the Germanic genus of the Indo-European language family, and Polish and Russian belong to the Slavic genus of Indo-European language family. The present paper uses morphology as the point of departure, and the studied languages can all be considered generally strongly suffixing in terms of their inflectional morphology (Dryer 2013a). They do, however, differ from each other as well as from Finnish in numerous form- or construction-specific typological features, such as the use of articles (Dryer 2013b, 2013c), the use of pronominal subjects (Dryer 2013d), as well as the use of pronominal possessive affixes (Dryer 2013e).

ICLFI is a collection of texts written by students of Finnish as a foreign language who study Finnish outside Finland. The data consist of various kinds of text types and genres, and texts have been lemmatized and annotated semi-automatically in terms of morphological forms, parts of speech and syntactic dependencies using the Connexor fi-fdg parser (Järvinen et al. 2004) and then controlling and validating the outcome manually (for a more detailed description of the annotation, see Jantunen et al. 2014). To get a balanced and comparable dataset, we randomly selected 240 texts from each L1-specific subcorpus of the ICLFI, resulting in a dataset of a total of 342,656 tokens. The distribution between the subcorpora can be seen in the Table 1.

ICLFI consists of various written genres typical for foreign language studying context. The data have been produced for learning purposes in the Finnish language course in which the informant has participated. In other words, the texts have not been produced primarily for the purposes of ICLFI. On the other hand, the texts may have been used for assessing the students' linguistic skills. Furthermore, their overarching primary genre is a language course assignment, and as such they can be said to mimic

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

	Number of texts	Number of tokens
L1-zh (Chinese)	240	45,989
L1-et (Estonian)	240	46,606
L1-de (German)	240	30,176
L1-pl (Polish)	240	112,407
L1-ru (Russian)	240	68,347
L1-sv (Swedish)	240	39,131
Total	1440	342,656

Table 1. Dataset of the study.

Genre	Number of texts	
Story	592	(41.1%)
Essay	260	(18.1%)
Opinion	157	(10.9%)
Diary	118	(8.2%)
Review	104	(7.2%)
Letter	88	(6.1%)
Summary	76	(5.3%)
Not specified	16	(1.1%)
News	14	(1.0%)
Email	9	(0.6%)
Application	6	(0.4%)
Total	1440	(100.0%)

Table 2. Genre distribution of the data.

the genre they represent, rather than being actual instances of the respective genre (for discussion on genre in L2 writing and instruction, see e.g. Hyland 2004, 2007). The distribution between genres can be seen in Table 2. The distribution is somewhat unbalanced but, as the methodology description shows, the necessary precautions have been considered to account for the possible genre effects in the analysis.

Each text in our dataset has also been assessed and annotated in terms of the Common European Framework of Reference for Languages (CEFR) level they represent, as well as the amount of instruction received (CEFR 2006). The assessments have been done separately by two certified assessors, or three if the two first assessments diverged (Jantunen et al. 2014:67). Table 3 shows the data distribution in terms of the CEFR levels of the data. The amount of instruction has been divided into three categories: less than 200 hours of instruction (58% of data); 200–400 hours of instruction (15.6% of data); over 400 hours of instruction (26.4% of data).

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

CEFR level	Number of texts	
A1: Breakthrough	19	(1.3%)
A2: Waystage	121	(8.4%)
B1: Threshold	603	(41.9%)
B2: Vantage	487	(33.8%)
C1: Effective operational proficiency	160	(11.1%)
C2: Mastery	50	(3.5%)
Total	1440	(100.0%)

Table 3. CEFR proficiency level distribution of the data.

3.2 Methodology: Key structure analysis

In this study, we base our inquiry on the detection-based argument of CLI (Jarvis 2012). In other words, we look for repeated quantitative differences between the different L1-specific datasets, and focus the subsequent analysis of the thus revealed features. Our choice of methodology is the KEY STRUCTURE ANALYSIS, a step-wise methodological procedure that combines several well-established corpus linguistic methods (see Ivaska 2015a). In general, the procedure consists of three consecutive phases that together help create a detection-based argument for crosslinguistic influences: (i) to statistically detect linguistic features whose frequencies best distinguish the compared datasets; (ii) to analyze the inner and the contextual variation of the revealed features, so as to reveal the constructions in which the features typically occur; and (iii) to analyze the typical use of the found constructions, other possibly relevant constructions, and their relationship with extra-linguistic factors across the datasets. In other words, key structure analysis can be used to link observed quantitative differences between datasets with linguistically intelligible and qualitatively defined linguistic phenomena, and thus to access constructional differences or changes in a data-driven manner (for a more detailed discussion of the methodological procedure and its underlying mechanisms, see Ivaska 2015a). Key structure analysis has earlier been used successfully to reveal and analyze constructional differences between L1 and L2 (Ivaska 2014), differences between learners with various L2 backgrounds (Ivaska 2015b), longitudinal changes in L2 (Ivaska 2015c), as well as differences between different genres in L2 and L1 parallelly (Ivaska 2016).

As a point of departure, we extracted the frequencies of grams of all the morphological tags in each text unit, and normalized the frequencies over occurrences per 1000 tokens. Each morphological gram contains all the morphological information of one word, so that each gram may consist of several morphological tags, such as the grammatical number and case marking. Then, following the example of earlier similar studies (see Jarvis, Castañeda-Jimenez & Nielsen 2012, Pepper

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

2012, Ivaska 2015b), we focused solely on fairly frequent features, in this case on morphological forms that were among the 100 most common forms in any of the six L1-specific subcorpora.

Then, we used a statistical method called RANDOM FOREST (for the algorithm, see Breiman 2001; for possible applications in linguistics, see Tagliamonte & Baayen 2012). The method is originally designed for automatic data classification and regression, and it roughly works as follows: (i) it leaves aside a subset of data used in the evaluation; (ii) it randomly chooses a set of variables (in this study, the frequencies of morphological grams) and compares them to see which are the best predictors for the value of the response variable (in this study, the L1 background of the writer), effectively creating a classification tree; (iii) it repeats phase two a large number of times; and (iv) it uses the thus gained group of classification trees to classify data left aside. The success of this classification is measured with a permutation test. In other words, one by one, the values of each variable are randomly permuted to see, how much worse the model classifies the response variables after the permutation. In our statistical analysis, we used the *cforest* version of Random Forests to create the statistical model used in the classification and the method *varimp* to sort out the best predictors of the L1 background. Both methods are found in the R package *party* (Hothorn et al. 2006, Strobl et al. 2008, Strobl, Malley & Tutz 2009).

Because of the inbuilt cross-validation mechanism described above, the method successfully avoids typical problems of overfitting, i.e. explaining data at hand better than other data. To reduce the possible overfitting even further, we modelled five parallel Random Forests and used the average of the thus gained variable importance measures in our analysis. In this manner, we were able to find the morphological grams that do at the same time represent intra-L1-group homogeneity and inter-L1-group heterogeneity, and may thus indicate possible instances of CLI. We then chose ten best predictors for further analysis. Out of these ten, we ruled out other possible variables by means of another Random Forest modeling, so as to assure that the detected differences are actually more likely to be due to the L1 background and not to any other possible variable. These variables were writer's L1 background, CEFR level of the text, text genre, amount of instruction received and text length in words. All the morphological grams in which the L1 background was not the best predictor were ruled out of the closer analysis. We then sorted the nature of the differences between the different L1 background by means a Tukey HSD statistical test, which is designed for multiple comparison (see e.g. Pepper 2012:88).

We then proceeded to analyze each of these remaining grams and their typical environments (Francis 1993), so as to be able to reveal the constructions in which the found features typically occur, to see how the use diverges between the different L1-specific subsets, as well as to relate it to the nature of the respective and other relevant constructions in Finnish and, when needed, to the relevant constructions in the various L1s. In other words, we extracted from the corpus all the occurrences of these grams

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

together with their immediate linguistic context and ended up with a spreadsheet with all the lexemes occurring in each given morphological form together with every lexeme, every morphological form, every part of speech and every syntactic function occurring one or two words before and after the detected morphological form. We used Random Forests also here to reveal the best contextual predictors of the difference between the datasets, together with a classification tree method called *ctree* to further explore and to illustrate the nature of the observed difference. We then used this information to figure out the involved construction or constructions and to compare their use in the different datasets. We also used the obtained constructional information to narrow down the exact linguistic phenomena to focus on in the linguistic systems of the different L1s involved. This last phase helps us to formulate and evaluate a possible comparison-based argument for crosslinguistic influences.

We conducted all the corpus queries and data manipulation using automatized scripts written in Java programming language, whereas we conducted all statistical analyzes with R (R Core Team 2016).

4. RESULTS

4.1 Detecting the best predictors of writers' L1 background

There were all in all 2246 different morphological grams in the data. Of these grams, 150 were among the 100 most common grams in at least one of the six L1-specific subcorpora. The frequencies of use of these 150 grams were considered as possible predictors of the writers' L1 background in the five modelled Random Forests and in the variable importance measure tests. We used the thus obtained variable importance measures to list the different morphological grams in a decreasing order of importance regarding how well their observed frequency can predict the L1 background of the writer. Note that the values of the variable importance measure are not meaningful as such, since they describe each variable of the model in relation to the model as a whole and to the other considered variables. We then focused the subsequent analysis on the ten best predictors. The twenty best predictors can be seen in Figure 1.[3]

To make sure our inquiry focused on the possible CLI rather than other possible factors, we modelled another Random Forest for each of the ten best morphological grams and produced a variable importance measure about which of the L1 background, the text genre, the text length, the proficiency level and the amount of instruction was the best predictor of the observed frequency. We then ruled of the analysis all the cases where the L1 background was not the best predictor of the frequency. This left us with three grams: (i) exemplifies a singular nominative with a first person possessive suffix (SG NOM CLI POSS P1); (ii) exemplifies a first person personal pronoun in singular nominative (SG P1 NOM); and (iii) exemplifies

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

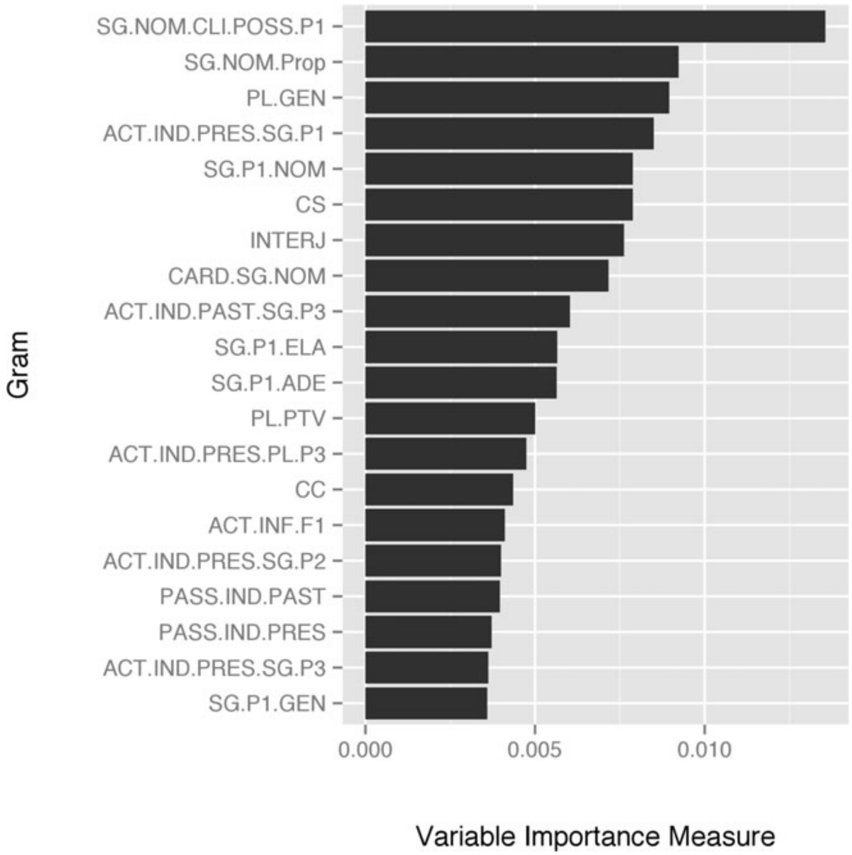

Figure 1. The mean of the variable importance measures of the twenty best predictors of L1 background.

a cardinal number in singular nominative (CARD SG NOM). The morphological forms in question are set in boldface in all the examples in (1)–(3).

(1) Ehkä yliopisto-n **elämä-ni** alko-i vähän normallinen.
maybe university-GEN life-PX1SG begin-PAST little normal
‘Perhaps my university life began pretty normal.’ (L1-zh – KI0001fA)

(2) Kene-n kanssa **minä** halua-n viettä-ä vapaa-ta aika-a.
who-GEN with I want-1SG spend-INF free-PTV time-PTV
‘With whom do I want to spend free time?’ (L1-ru – VE0078A)

(3) Asu-i-n **yksi** vuosi Brusseli-ssa lapse-na.
live-PAST-1SG one year Brussels-INE child-ESS
‘I lived one year in Brussels as a child.’ (L1-de – SA0145A)

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

4.2 Singular nominative with first person possessive suffix

The frequency of the singular nominative with a first person possessive suffix (SG NOM CLI POSS P1) is the best morphological gram to predict the L1 background of the author. In Finnish, the possessive suffixes are pronominal bound morphemes that express possessor by attaching the affix to the end of the possessed entity, which can be a noun or a non-finite verb and which always is the head of the phrase – as *isäni* in (4).

(4) **Isä-ni** on kuol-lut.
father-PX1SG *be.3SG.PRES* *die-PCPL2*
'My father is dead.' (L1-sv –RU0050A)

All three different grammatical persons have separate suffixes, and the grammatical number is also specified in the first and the second person. The possession is often double-marked, as possessive suffixes commonly co-occur with a genitive construction expressing the possessor, although spoken language tends to omit the possessive suffixes (VISK:§95–96). The mechanism is fairly common in the language of the world – 642 languages of the 902 languages recorded in the *World Atlas of Language Structures* (*WALS*) have either possessive prefixes, suffixes, or both (Dryer 2013e). Interestingly, none of the L1s covered in the present study have affixal system expressing the possession, and the possession is marked solely in the possessor with a genitive, which is the dependent of the phrase (Nichols & Bickel 2013; Erelt et al. 1993:120–121 for Estonian; Bielec 1998:106, 152 for Polish).

As Figure 2 shows, singular nominative with a first person possessive suffix is used more often in L1-Estonian subcorpus than in any of the other subcorpora (median: 11.14/1000 words). According to Tukey HSD, the difference is statistically highly significant between L1-Estonian and all the other subcorpora ($p < .001$) and not between any other subcorpora ($p > .05$). The most typical use of the gram is connected with family descriptions – the word *isä* 'father' is among the ten most common lemmas for the form in all the subcorpora, and other family words like *perhe* 'family', *sisko* 'sister', *veli* 'brother', as well as *tyttöystävä* 'girlfriend' and *poikaystävä* 'boyfriend', also commonly occur in the context.

We then analyzed the typical use of the morphological gram by means of another Random Forest, so as to reveal the differences in the context of use between the datasets. As can be seen in Figure 3, the clearly best contextual predictor is the lemma occurring in the first person possessive form. We then explored this difference using a classification tree, so as to find out the lemmas that account for the observed difference. As can be seen in Figure 4, words associated with text topics dealing with home or family (lemmas *huone* 'room', *isä* 'father', *koti* 'home', *nimi* 'name', *perhe* 'family', *päivä* 'day', *sisko* 'sister', and *veli* 'brother') all occur on one side

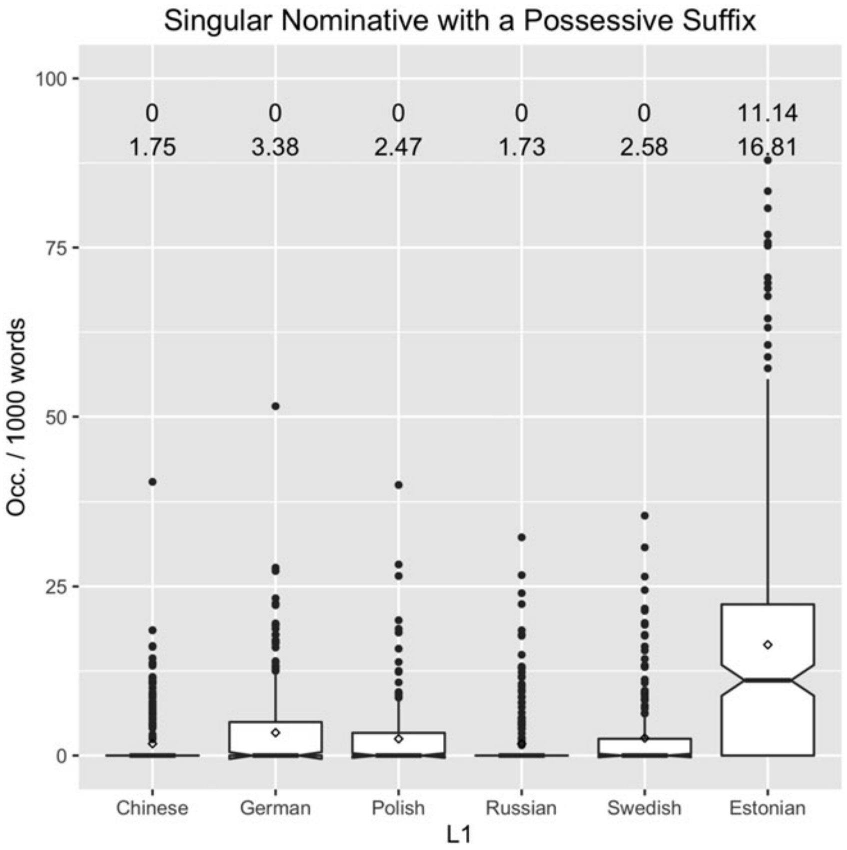

Figure 2. Boxplots and relative frequencies of singular nominative with a first person possessive suffix. The uppermost value tells the respective median value and the lower value the mean value.

of the classification tree that is also almost exclusively dominated by examples from the L1-et subcorpus, whereas words *päiväkirja* ‘diary’, *ystävä* ‘friend’, and all the other words occur on the other side, and the distribution is more even between the L1 backgrounds.

In other words, these words are good predictors whether the L1 of the writer is Estonian. The result suggests that, rather than the typological nature of the writers’ first languages, the difference is very likely due to a topical difference, so that the Estonian students have written texts on their families more than other students. Note that here, as in the other classification trees, the bars are based on the total number of occurrences, which is why the bar of the L1 with the overall highest frequency tends to be the highest in several branches of the tree, whereas the differences between the branches depict the differences between the different subcorpora.

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

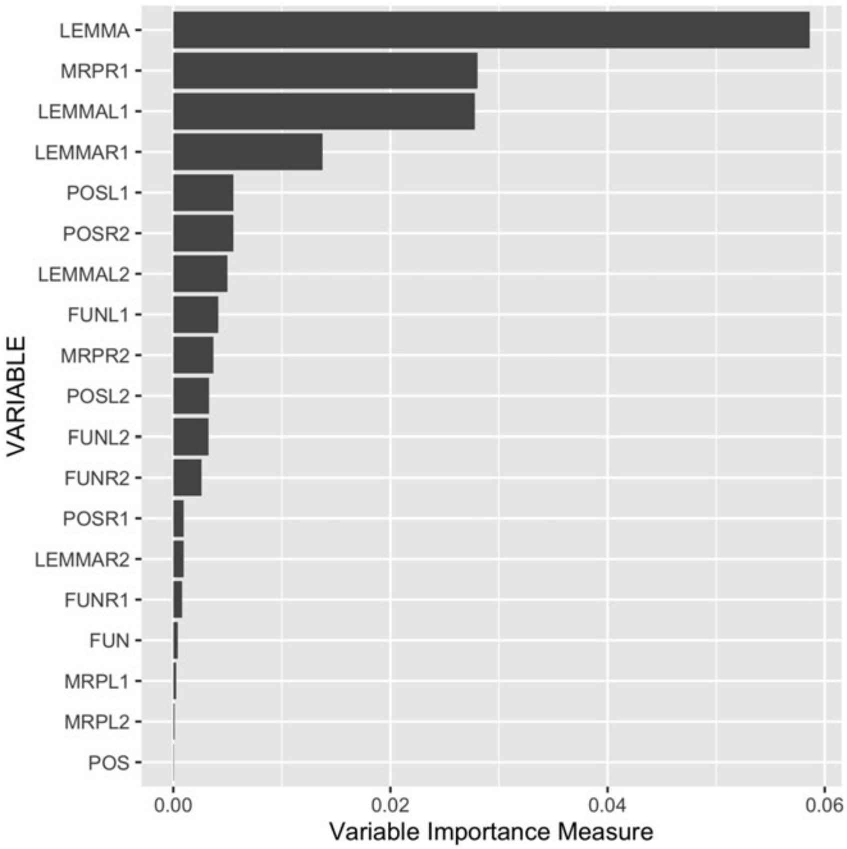

Figure 3. The variable importance measures for the linguistic context of the morphological gram SG NOM CLI POSS P1 for predicting whether the first language of the writer is Estonian or something else.

4.3 First person personal pronoun in singular nominative

The frequency of singular nominative in the first person, i.e. the personal pronoun *minä* ‘I’ (SG P1 NOM) is the second best morphological gram to predict the L1 background of the author. In Finnish, the word form typically occurs as a subject in normal clauses, as in (5).

(5) **Minä** **ole-n** Helsingi-ssä seitsemän päivä-ä.
I.SG.NOM *be-1SG* *Helsinki-INE* *seven* *day-PTV*
‘I am in Helsinki for seven days.’ (L1-sv – RU0043A)

The expression of a pronominal subject is optional in the first and the second person, as can be seen in (6) whereas it is obligatory in the third person (VISK:§914).

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

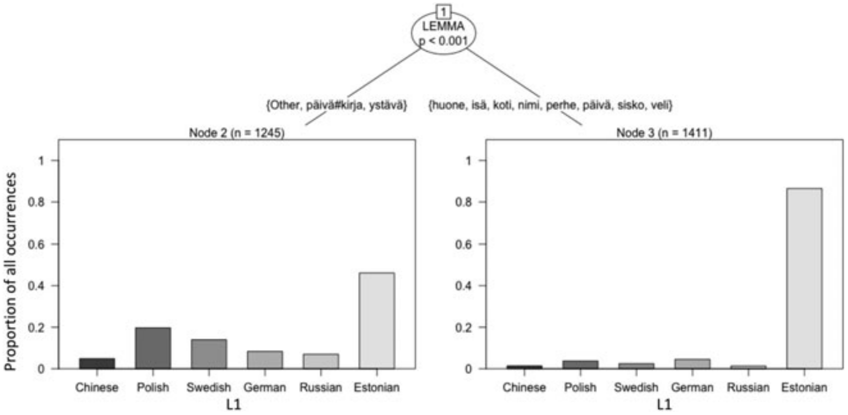

Figure 4. Classification tree for lemmas occurring in the morphological gram SG NOM CLI POSS P1 that best predict the first language of the writer.

(6) **Ole-n** hyvin iloinen.
*be-1*SG *very* *glad*
'I am very glad.' (L1-pl – PU003lA)

Finnish also marks the subject with a personal suffix attached to the verb, and in *WALS* Finnish is recorded as a mixed system with none of the types dominant (Dryer 2013d). As for the different L1s covered in the present study, in German, Russian, and Swedish, pronouns in the subject position are recorded as obligatory, whereas in Chinese they are recorded as optional and in Polish the subjects are recorded to behave as clitics that attach to variable hosts (ibid.). Estonian is recorded to represent the by far largest of the recorded groups, where pronominal subjects are expressed with affixes on verbs (ibid.), although subjects in Estonian are mainly omitted in first and second person and in unmarked clauses (Metslang 2013).

Personal pronouns in singular nominative are used more often in L1-Swedish subcorpus than in any other subcorpora (median: 9.5/1000 words). The distribution and some basic descriptive statistics are shown in Figure 5. According to Tukey HSD, the difference is statistically highly significant between L1-sv subcorpus and all the other subcorpora ($p < .001$) and not significant between any other subcorpora ($p > .05$). Interestingly, Finnish, Estonian, German, Polish, and Russian all have also verbal person marking for subjects, whereas that is not the case in Chinese or in Swedish (Siewierska 2013; except Metslang 2013:241–242 for Estonian and authors' personal knowledge for Swedish). This leaves Swedish as the only L1 with mandatory subject marking and pronoun as the only means to do it. In other words, given that the subject expression accounts for a majority of the observed difference, all the four premises of a comparison-based argument are fulfilled.

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

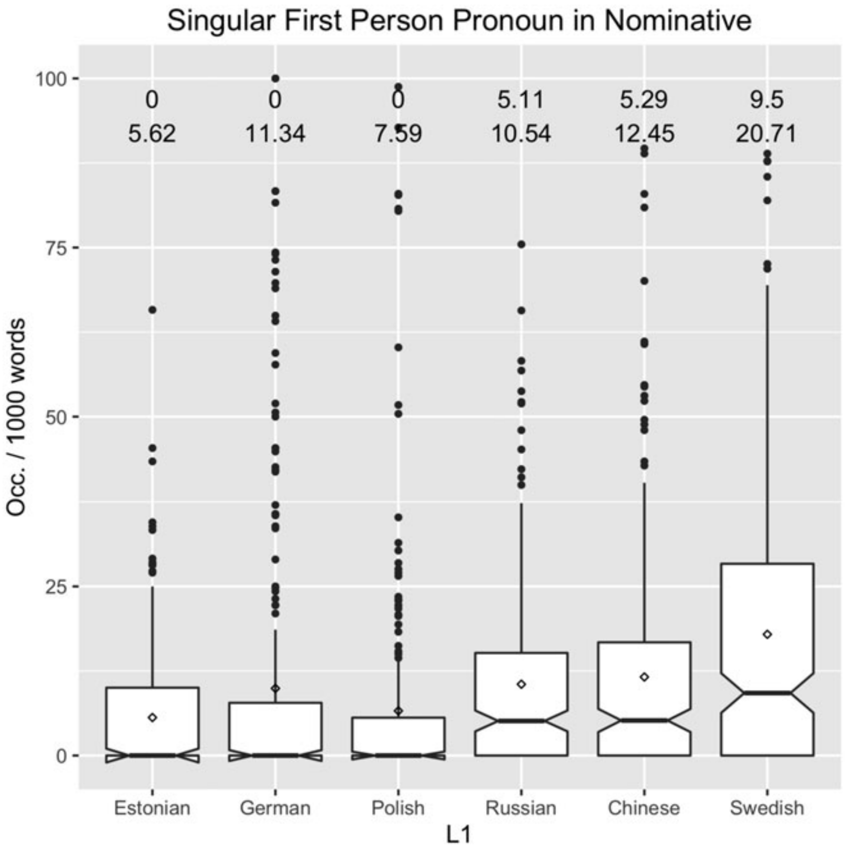

Figure 5. Boxplots and relative frequencies of singular first person pronoun in nominative. The uppermost value tells the respective median value and the lower value the mean value.

In the analysis of the best contextual predictors, lemmas occurring immediately after the first person personal pronouns are by far the best predictors for the L1 background of the writer. Variable importance measures can be seen in Figure 6. Closer look at the lemmas occurring after the pronoun does indeed prove that it is rather the variety of different verbs occurring after the personal pronoun *minä* that distinguishes L1-sv subcorpus from the other subcorpora than any specific word. The examples in the L1-sv subcorpus occur on the right branch of the classification tree in Figure 7, and the words occurring after the pronoun in this group include very frequent and generic verbs such as *haluta* 'to want', *mennä* 'to go', *olla* 'to be', *syödä* to eat', and *voida* 'to be able to'. This can be seen to indicate that it is the first person pronominal subject expressions in general that distinguish L1-Swedish learners from the other learner groups.

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

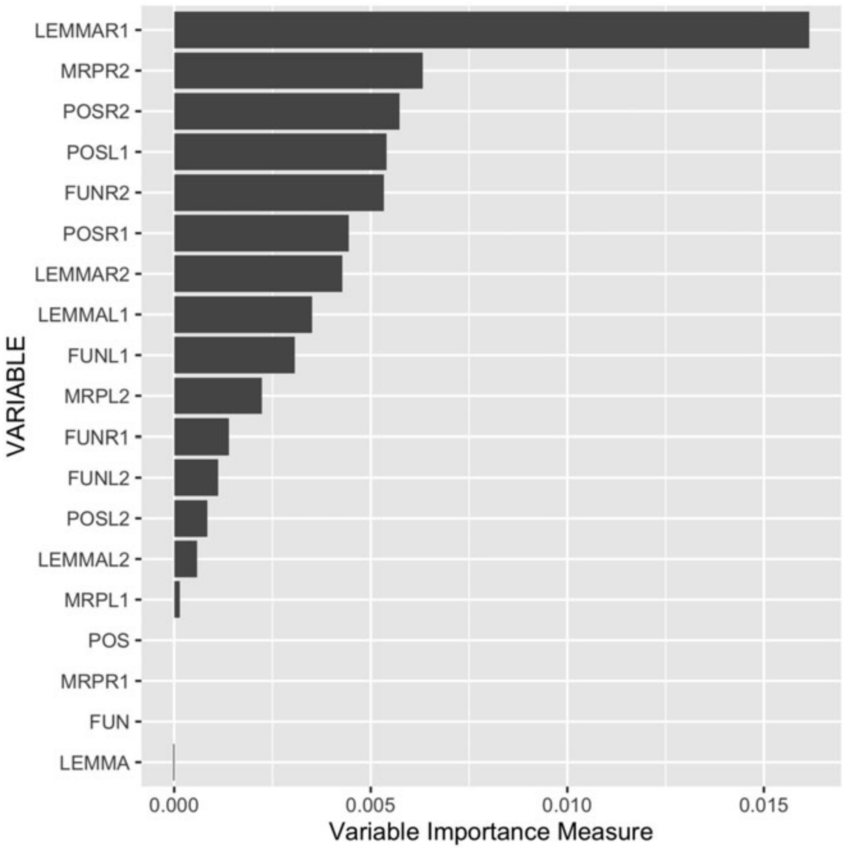

Figure 6. The variable importance measures for the linguistic context of the morphological gram SG P1 NOM for predicting whether the first language of the writer is Swedish or something else.

4.4 Cardinal number in singular nominative

The frequency of cardinal numbers in singular nominative (CARD SG NOM) is the second best morphological gram to predict the L1 background of the author. In Finnish, as in all the L1s represented in our data, numeral precedes the noun it occurs with (Dryer 2013f). The numeral phrase in Finnish has two parallel structures that differ formally from each other; (7) represents the unmarked version where the numeral *kaksi* 'two' is the head of the phrase in the nominative case, with the noun *vuotta* 'year' as the dependent in the partitive form.

(7)	Merili	on	**kaksi**	**vuot-ta**	nuore-mpi	kuin	minä.
	Merili	*is*	*two.*SG.NOM	*year-*PTV	*young-*CMP	*than*	*I*

'Merili is two years younger than me.' (L1-et – VI0001bA)

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

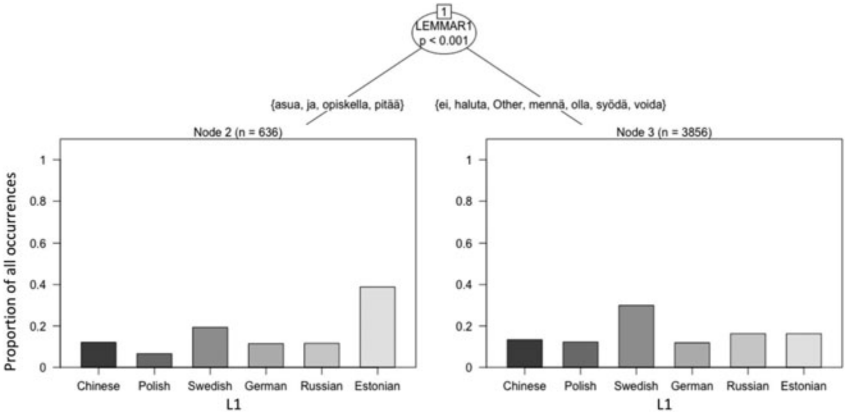

Figure 7. Classification tree for lemmas occurring immediately after the morphological gram SG P1 NOM that best predict the first language of the writer.

The example in (8) illustrates the marked option, where the noun *tavalla* 'manner' is the head of the phrase and the numeral *kolmella* 'three' agrees in the case marking.

(8) Nii-den täyty-y myös esiinty-ä **kolme-lla** **tava-lla**.
*it.*PL-GEN *must-3*SG *also* *occur-*INF *three-*ADE *manner-*ADE
'They also have to occur in three different ways.' (L1-et – VI0132a)

The numeral phrases are semantically plural but formally singular, except in the plurale tantum cases (*kahde-t housu-t* 'two [pairs of] pants') or when the phrase refers to several different representations of the referent (*kymmene-t kerra-t* 'tens of times') (VISK:§789).

In the different L1s covered in the present study the phrases with numerals behave in various ways; in Chinese separate measure words are obligatory when their associated nouns are quantified by numerals (Yip & Rimmington 2004:32). In Polish and in Russian numeral either acts as a dependent in a phrase agreeing in the case with the head noun, or it is the head reflecting the case of the phrase with the noun as the dependent occurring in genitive. The variation depends on the numeral involved and in the syntactic function of the phrase. (For Polish, Bielec 1998:241–247; for Russian, Timberlake 1993:876–878). In German and in Swedish, numerals act as dependents of the noun phrase, and generally only numeral 'one' – which is also the indefinite article in both languages – agrees with the noun (authors' personal knowledge). In Estonian, the system is in major parts similar with Finnish, and also the lexemes are close to their Finnish equivalents (Erelt et al. 1993:140).

As Figure 8 shows, cardinal numbers in singular nominative are used more often in L1-et subcorpus than in any of the other subcorpora and more often in L1-sv subcorpus than in the remaining subcorpora. According to the Tukey HSD, the

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

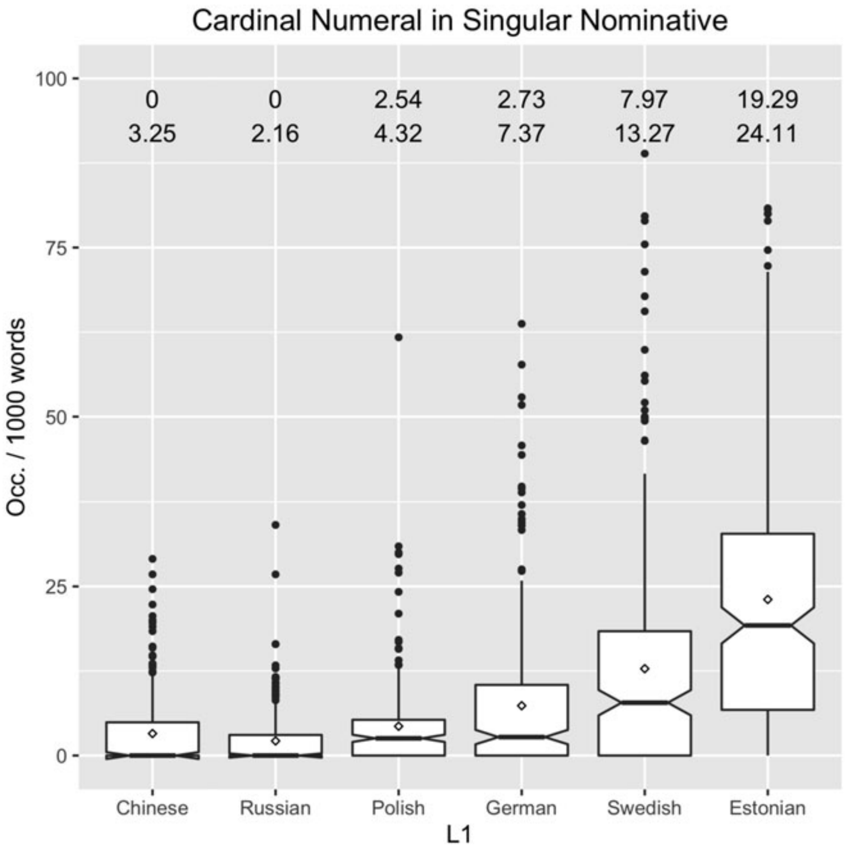

Figure 8. Boxplots and relative frequencies of cardinal numerals in singular nominative. The uppermost value tells the respective median value and the lower value the mean value.

difference is statistically highly significant between L1-et and all the other subcorpora ($p < .001$) and also highly significant between L1-sv and all the other subcorpora ($p < .001$), whereas there are no statistically significant differences between the rest of the subcorpora. We will therefore treat the L1-et and L1-sv separately in our subsequent analysis.

In our data in general, numerals in singular nominative typically occur either in predicative constructions expressing time or frequency like (9) or in possessive constructions like (10).

(9)	Kello	oli	**kaksitoista**	yö-llä.
	clock	*was*	*twelve.*SG.NOM	*night-*ADE

‘It was twelve at night.’ (L1-ru – RU0014fA)

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

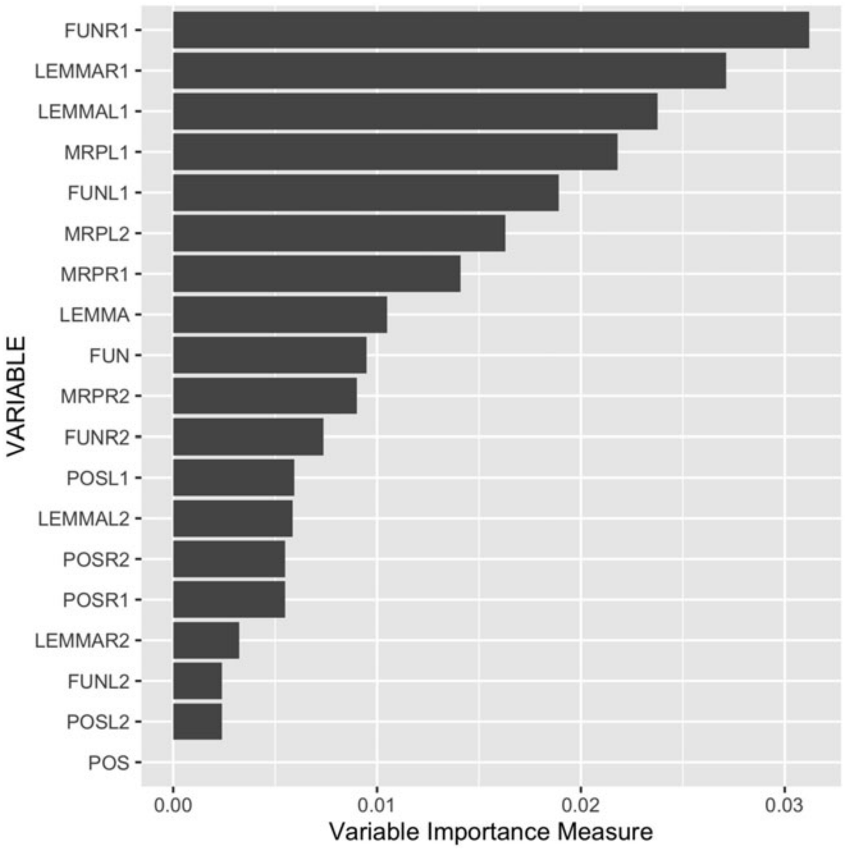

Figure 9. The variable importance measures for the linguistic context of the morphological gram CARD SG NOM for predicting whether the first language of the writer is Estonian or something else.

(10)	Minu-lla	on	**viisi**	kissa-a	ja	hevos-ta.
	I-ADE	*is*	*five.SG.NOM*	*cat-PTV*	*and*	*horse-PTV*

‘I have five cats and horses.’ (L1-pl – PU0008fA)

Functions occurring immediately after the numerals (FUNR1) turn out to be the best contextual predictors of difference between L1-et and the other subcorpora. Figure 9 shows the variable importance measures of all the contextual variables, none of which stands out as clearly better in accounting for the difference. This could indicate that the difference is not due to any one use of the numerals but rather to the similarity of the numeral system between Finnish and Estonian as a whole. As the classification tree about the syntactic functions following the numerals in Figure 10 shows, the clearest difference between L1-et and the other subcorpora is that the others are more

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

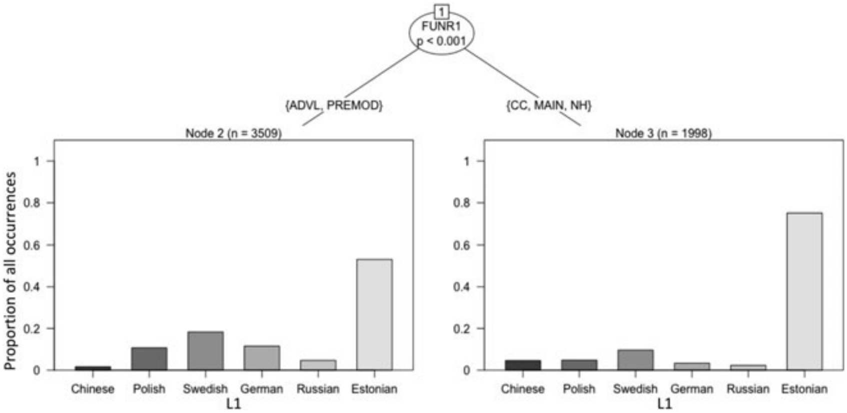

Figure 10. Classification tree for syntactic functions occurring immediately after the morphological gram CARD SG NOM that best predict the first language of the writer.

likely to represent the most common patterns of use with the numeral as the head of the phrase, followed by the modifier (PREMOD), whereas in L1-et the numerals are more often followed either by coordinate conjunctions (CC), main verbs (MAIN) or separate heads of nominal phrases (NH). This is not to say that the typical use does not occur in the L1-et but rather that there is likely to be more variation in the numeral use that is less common in other subcorpora. Given that the numeral systems of the other languages diverge from both Finnish and Estonian and knowing that the difference in use is not due to topical differences between the subcorpora, the use of numerals fulfills all the premises of a comparison-based argument for CLI.

As for the L1-sv subcorpus, functions occurring immediately before the numerals are the best contextual predictors of difference from the other subcorpora, followed by lemmas occurring in the same position. Figure 11 shows the variable importance measure for all the contextual variables. The closer analysis reveals that in the L1-sv subcorpus, the numerals are more often preceded by a word modifying the numeral (PREMOD) than in the the other subcorpora. This can be seen in the right branch of the classification tree in Figure 12. Furthermore, the lexical variation suggests that this is at least partially due to the expressions of time that involve either the word *kello* ‘clock, watch’, as in (11), or the word *puoli* ‘half’, as in (12).

(11)	**Kello**	**kahdeksan**	Timo	men-i	elokuv-i-in.
	clock	*eight*.SG.NOM	*Timo*.PROP	*go*-3SG.PAST	*movie*-PL-ILL

‘At eight o’clock, Timo went to movies.’ (L1-sv – RU0017cA)

(12)	Tul-i-mme	Luulaja-an	**puoli**	**kuusi**.
	come-PAST-1PL	*Luleå*.PROP-ILL	*half*	*six*.SG.NOM

‘We arrived to Luleå at half past five.’ (L1-sv – RU0017mA)

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

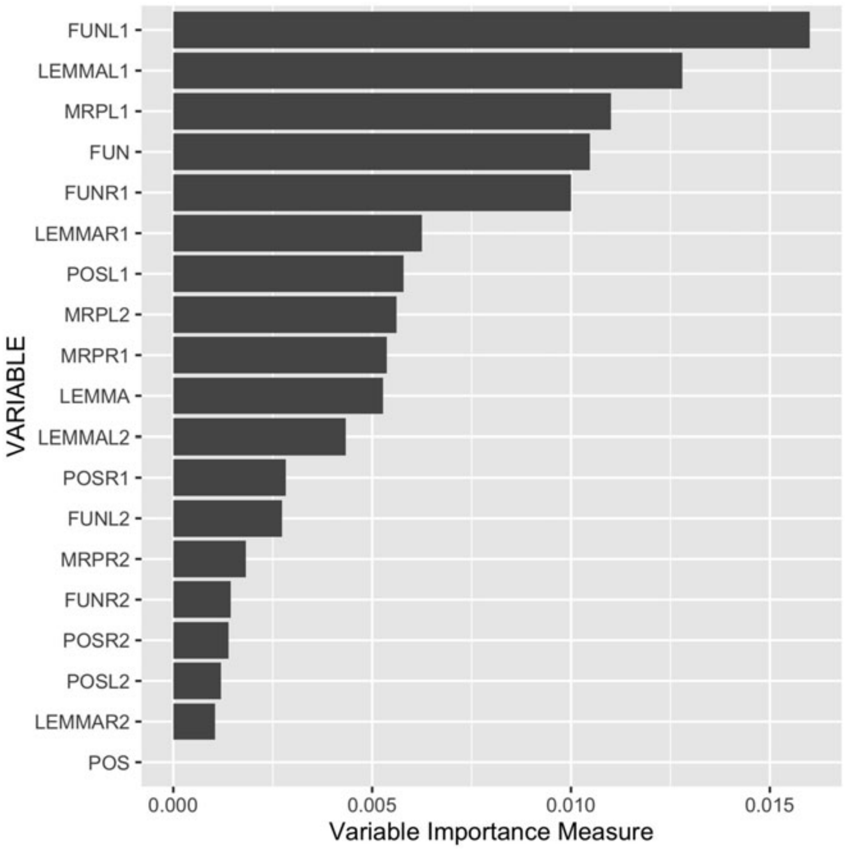

Figure 11. The variable importance measures for the linguistic context of the morphological gram CARD SG NOM for predicting whether the first language of the writer is Swedish or something else.

This can be seen in the classification tree in Figure 13, where the right branch and especially the node six contains the majority of the cases by L1-Swedish writers. In other words, L1-sv subcorpus has a lot of expressions of time, and that distinguishes it from the other subcorpora. This is very likely due to the topical difference between the subcorpora – L1-sv has more texts about the writers' daily routines which often call for the expressions of time.

5. DISCUSSION AND CONCLUSIONS

Our results suggest following answers to the research questions proposed: It seems that the existence of morpho-syntactically similar constructions in the L1 and L2 is

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

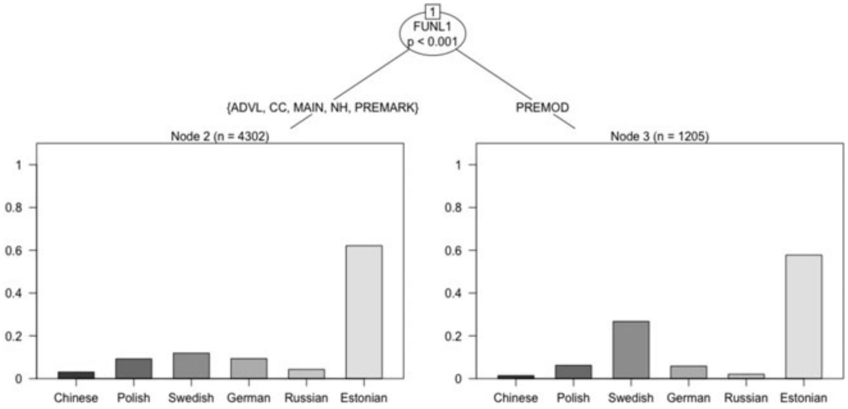

Figure 12. Classification tree for syntactic functions occurring immediately before the morphological gram CARD SG NOM that best predict the first language of the writer.

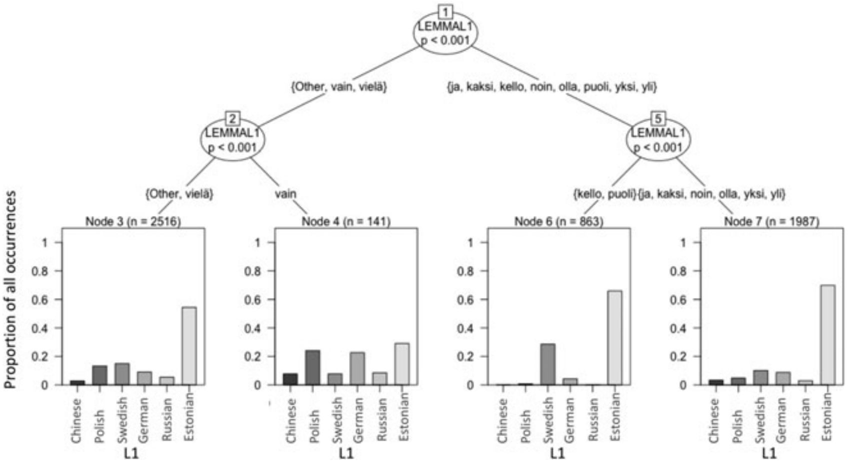

Figure 13. Classification tree for lemmas occurring immediately before the morphological gram CARD SG NOM that best predict the first language of the writer.

likely to increase the frequencies in the L2. We detected and analyzed in detail the use of three different morphological forms that did, based on their frequency, differentiate one or several L1-specific subcorpus from the other subcorpora. The forms analyzed are singular nominative with a first person possessive suffix, first person personal pronoun in singular nominative, and cardinal numeral in singular nominative. Our results suggest that two of these three indicators do in fact fulfill all the premises of the comparison-based argument and do, thus, reflect CLI, while one is in this case more likely indicating a topical difference between the subcorpora. First, it is likely that L1-Swedish learners of Finnish express the first person subject overtly more often

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

than the other learner groups, because Swedish is the only language that both has a mandatory pronominal subject-marking and lacks a conjugational subject marking. Both of these two constructional features of the L1 are also present in the Finnish, but CLI are likely to increase the likelyhood of such double-marked subject–verb constructions. Second, Estonian speakers are likely to use more numerals than the other L1 groups because the numerals as lexical items resemble each other in Finnish and in Estonian and because the numeral phrases are constructionally similar in the two languages and different in the other L1s. L1-Swedish learners also use numerals more often than the other groups but, based on the contextual analysis and the typical co-occurrence patterns, we suggest this difference is more likely due to a topical difference than to CLI. More specifically, the L1-sv subcorpus has a lot of texts in which the learners describe their normal days, which increases the frequency of time expressions. Similarly, we also suggest that the greater use of first person possessive suffixes by the L1-Estonian learners is at least in our data more likely due to a topical difference than to CLI. The possessive expressions in the L1-et subcorpus are very often related to family words, and they are responding to writing prompts that ask the learners to write about their own families.

As for the second research question, it can be said that the method applied was successful in detecting features in which CLI may possibly play a role. The key structure analysis can be used as a detection-based methodological procedure, and it can successfully provide pointers to look at right directions. It does not, however, give well-formulated answers, and it remains the duty of the researcher to interpret the constructions involved, and to dig deeper in the nature of the distributional differences and the typological comparisons between languages. The clearest strength of the method applied is that it can reveal constructions that seem to either function similarly in the different L1s or it can portray combinations of several simultaneous and intertwining CLI that together lead to assumed similarities. Thus, methodologically the results support a form- or construction-specific approach towards CLI, instead of relying solely or even primarily on genealogical relationships between languages.

The results raise a concern regarding the data used in this study but also in any other study that makes use of naturally occurring learner writing. In two cases in our analysis, the underlying reason for the observed difference between the different L1s is the topic of the texts. Due to the structure of the corpus used and especially to the multitude of ways the topics were documented ranging from verbatim prompts to vague task descriptions, we could not take the topic into account as a variable in our statistical models. While it can in some ways be seen as limiting the significance of the results, we see it as a more general phenomenon that is understated in the study of CLI in general. On the one hand, the fact that the method was able to reveal the topically skewed nature of the corpus does in our opinion further support the applicability of the method. On the other hand, we find it alarming that the topics have such a strong impact on the feature frequencies even in this big and diverse a

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

dataset. In other words, we do not see it is a problem to be solved but a variable to be considered. It seems to us that controlling the text type and the genre alone does not suffice, but one has to systematically take into account also the topical variation, optimally as a precondition when sampling the data, or alternatively in the post hoc analysis of the observed quantitative differences.

In conclusion, we think that approaching typological distance from the perspective of second language production can open new avenues for future research. We believe that crosslinguistic influences can indeed be accessed from bottom up in a quantitative corpus-driven fashion, provided that a suitable corpus is available. Also, we do believe that a form-based approach – together with the Construction Grammar as a holistic theoretical framework that takes into account both form, meaning, and the use – is well-equipped to reveal qualitatively understandable linguistic tendencies and describe them in a meaningful way. The adopted approach makes it possible to address crosslinguistic influences from a an item-based perspective – influences between constructions in the languages studied instead of more general and more generic assumptions on influences between languages as a whole.

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to all the participants of the REMU conference in Joensuu in 2015 for their insightful comments. We would also like to extend our thanks to our colleagues and students at the University of Washington and the University of Turku for numerous informal discussions and their thoughts regarding the nature of the various first languages covered in this paper of which we do not have any first-hand knowledge. Finally, thank you to the three anonymous referees for their valuable comments in making this paper more coherent. We are obviously fully responsible for any remaining shortcomings or inadequacies.

NOTES

1. ICLFI does not distinguish between different Chinese languages, but all L1-zh data have been collected in Beijing.
2. The language-name abbreviations follow the language codes provided in the ISO 639-1 standard.
3. Abbreviations: 1PL = first person plural; 1SG = first person singular; 3SG = third person singular; ACT = active voice; ADE, ADE = adessive case; ADVL = adverbial; CARD = cardinal number; CC = coordinate conjunction; CLI = clitic; CMP, CMP = comparative; CS = subordinate conjunction; ELA = elative case; ESS, ESS = essive case; GEN, F1 = Finnish as a first language; FUN = syntactic function; GEN = genitive case; ILL, ILL = illative case; INE, IND = indicative mood; INE = inessive case; INF, INF = infinitive; INTERJ = interjection; L1 = word occurring immediately before the gram in question; L2 = word occurring two words before the gram in question; LEMMA = lemma; MAIN = main

case; NH = head of a nominal phrase; NOM, NOM = nominative case; P1 = first person; P3 = third person; PASS = passive voice; MRP = morphological form; PAST = past tense; PCPL2 = past participle; PL, PL = plural; POS = part of speech; POSS = possessive suffix; PREMOD = modifier; PRES, PRES = present tense; PROP, Prop = proper noun; PTV, PTV = partitive case; PX1SG = first person singular possessive suffix; R1 = word occurring immediately after the gram in question; R2 = word occurring two words after the gram in question; SG, SG = singular.

REFERENCES

Aarts, Jaan & Sylviane Granger. 1998. Tag secuenqes in learner corpora: A key to interlanguage grammar and discourse. In Sylviane Granger (ed.), *Learner English on Computer*, 132–141. London: Longman.

Bielec, Dana. 1998. *Polish: An Essential Grammar*. London: Routledge.

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

CEFR 2006 = *Common European Framework for Languages: Learning, Teaching, Assessment*. 2006. Cambridge: Cambridge University Press.

Croft, William & D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Dryer, Matthew. 2013a. Prefixing vs. suffixing in inflectional morphology. In Dryer & Haspelmath (eds.), http://wals.info/chapter/26 (accessed 30 October 2016).

Dryer, Matthew. 2013b. Definite articles. In Dryer & Haspelmath (eds.), http://wals.info/chapter/37 (accessed 30 October 2016).

Dryer, Matthew. 2013c. Indefinite articles. In Dryer & Haspelmath (eds.), http://wals.info/chapter/38 (accessed 30 October 2016).

Dryer, Matthew. 2013d. Expression of pronominal subjects. In Dryer & Haspelmath (eds.), http://wals.info/chapter/101 (accessed 30 October 2016).

Dryer, Matthew. 2013e. Position of pronominal possessive affixes. In Dryer & Haspelmath (eds.), http://wals.info/chapter/57 (accessed 30 October 2016).

Dryer, Matthew. 2013f. Order of numeral and noun. In Dryer & Haspelmath (eds.), http://wals.info/chapter/89 (accessed 30 October 2016).

Dryer, Matthew & Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dulay, Heidi, Marina Burt & Stephen Krashen. 1982. *Language Two*. Oxford: Oxford University Press.

Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael & Silvi Vare. 1993. *Eesti keele grammatika II* [The grammar of Estonian language]. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.

Francis, Gill. 1993. A corpus-driven approach to grammar: Principles, methods and examples. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*, 137–156. Amsterdam: John Benjamins.

Goldberg, Adele. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: The University of Chicago Press.

Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Helland Gujord, Ann-Kristin, Susan Nacey & Silje Ragnhildstveit (eds.). 2015. *BeLLS 6: Learner Corpus Research: LCR2013 Conference Proceedings*. Bergen: University of Bergen. http://dx.doi.org/10.15845/bells.v6i0 (accessed 11 November 2016).

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro & Mark Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7(3), 355–373.

Hyland, Ken. 2004. *Genre and Second Language Writing*. Ann Arbor, MI: The University of Michigan Press.

Hyland, Ken. 2007. Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing* 16, 148–164.

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

Ivaska, Ilmari. 2015a. *Edistyneen oppijansuomen konstruktiopiirteitä korpusvetoisesti: avainrakenneanalyysi* [Corpus-driven approach towards constructional features of advanced learner Finnish: Key structure analysis]. Ph.D. dissertation, University of Turku.

Ivaska, Ilmari. 2015b. Tracing crosslinguistic influences in structural sequences: What does key structure analysis have to offer? In Helland Gujord et al. (eds.), 23–44.

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

Ivaska, Ilmari. 2016. Genre effects in academic L2 writing. Presented at The American Association for Corpus Linguistics (AACL) 2016 Conference, 16–18 September 2016, Ames, IO.

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

Jantunen, Jarmo, Sisko Brunni, Liisa-Maria Lehto & Valtteri Airaksinen. 2014. Oppijankieliaineistojen annotointi – esimerkkinä ICLFI:n annotoinnin prosessit, ongelmat ja ratkaisut [How to annotate learner language: Principles, problems and solutions of ICLI]. In Maarit Mutta, Pekka Lintunen, Ilmari Ivaska & Pauliina Peltonen (eds.), *AFinLA-e*: Special issue of *Soveltavan kielitieteen tutkimuksia* 2014(7), 60–80.

Jantunen, Jarmo & Silja Pirkola. 2015. Oppijansuomen sähköiset tutkimusaineistot. Nykytilanne [Electronic corpora of learner Finnish: Current situation]. *Virittäjä* 119, 88–103.

Järvinen, Timo, Mikko Laari, Timo Lahtinen, Sirkku Paajanen, Pirkko Paljakka, Mirkka Soininen & Pasi Tapanainen. 2004. Robust language analysis components for practical applications. In Björn Gambäck & Kristiina Jokinen (eds.), *Coling 2004: Proceedings of the Workshop Robust and Adaptive Information Processing for Mobile Speech Interfaces*, 53–56. Riga: The Baltic Perspectives.

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. In Leah Roberts, Martin Howard, Muiris Ó Laire & David Singleton (eds.), *EUROSLA Yearbook 10*, 169–192. Amsterdam: John Benjamins.

Jarvis, Scott. 2011. Data mining with learner corpora: Choosing classifiers for L1 detection. In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot (eds.), *A Taste for Corpora: In Honour of Sylviane Granger*, 131–158. Amsterdam: John Benjamins.

Jarvis, Scott. 2012. The detection-based approach: An overview. In Jarvis & Crossley (eds.), 1–33.

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

Jarvis, Scott, Gabriela Castañeda-Jimenez & Rasmus Nielsen. 2012. Detecting L2 writers' L1s on the basis of their lexical styles. In Jarvis & Crossley (eds.), 34–70.

Jarvis, Scott & Scott A. Crossley (eds.). 2012. *Approaching Language Transfer through Text Classification*. Bristol: Multilingual Matters.

Jarvis, Scott & Terence Odlin. 2000. Morphological type, spatial reference, and language transfer. *Studies in Second Language Acquisition* 22, 535–556.

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic Influence in Language and Cognition*. London: Routledge.

Kaivapalu, Annekatrin. 2005. *Lähdekieli kielenoppimisen apuna* [Contribution of L1 to foreign language acquisition]. Ph.D. dissertation, Jyväskylän yliopisto.

Kaivapalu, Annekatrin & Maisa Martin. 2007. Morphology in transition: The plural inflection of Finnish nouns by Estonian and Russian learners. *Acta Linguistica Hungarica* 54(2), 129–156.

Kaivapalu, Annekatrin & Maisa Martin. 2014. Measuring perceptions of cross-linguistic similarity between closely related languages: Finnish and Estonian noun morphology as a testing ground. In Heli Paulasto, Lea Meriläinen, Helka Riionheimo & Maria Kok (eds.), *Language Contacts at the Crossroads of Disciplines*, 283–318. Newcastle upon Tyne: Cambridge Scholars.

Koppel, Moshe, Jonathan Schler & Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 624–628. Chicago, IL: Association for Computing Machinery.

Mayfield Tomokiyo, Laura & Rosie Jones. 2001. You're not from 'round here, are you? Naive Bayes detection of non-native utterance text. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL '01), 239–246. Cambridge, MA: Association for Computational Linguistics.

Metslang, Helena. 2013. Coding and behavior of Estonian subjects. *Journal of Estonian and Finno-Ugric Linguistics* 4(2), 217–293.

Nichols, Johanna & Balthasar Bickel. 2013. Locus of marking in possessive noun phrases. In Dryer & Haspelmath (eds.), http://wals.info/chapter/24 (accessed 30 October 2016).

Odlin, Terence. 1989. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

Pepper, Steve. 2012. Lexical transfer in Norwegian interlanguage: A detection-based approach. Master's thesis, University of Oslo.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/ (accessed 11 November, 2016).

Ringbom, Håkan. 1987. *The Role of the First Language in Foreign Language Learning*. Clevedon: Multilingual Matters.

Ringbom, Håkan. 2007. *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual Matters.

Ringbom, Håkan & Scott Jarvis. 2009. The importance of crosslinguistic similarity in foreign language learning. In Michael H. Long & Catherine J. Doughty (eds.), *Handbook of Language Teaching*, 106–118. Oxford: Blackwell.

Scott, Mike & Cristopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

https://doi.org/10.1017/S0332586517000129 Published online by Cambridge University Press

Selinker, Larry & Usha Lakshamanan. 1992. Language transfer and fossilization: The multiple effects principle. In Susan Gass & Larry Selinker (eds.), *Language Transfer in Language Learning*, 197–216. Amsterdam: John Benjamins.

Siewierska, Anna. 2013. Verbal person marking. In Dryer & Haspelmath (eds.), http://wals.info/chapter/102 (accessed 30 October 2016).

Spoelman, Marianne. 2013. *Prior Linguistic Knowledge Matters: The Use of the Partitive Case in Finnish Learner Language*. Ph.D. disseratation, University of Oulu.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* 9(307). http://www.biomedcentral.com/1471-2105/9/307 (accessed 11 November 2016).

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and Random Forests. *Psychological Methods* 14(4), 323–348.

Tagliamonte, Sali & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2), 135–178.

Timberlake, Alan. 1993. Russian. In Bernard Comrie & Greville Corbett (eds.), *The Slavonic Languages*, 827–886. London: Routledge.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harward University Press.

VISK = Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho. 2004. *Iso suomen kielioppi* [The great grammar of Finnish]. Helsinki: Suomalaisen Kirjallisuuden Seura. http://scripta.kotus.fi/visk (accessed 11 November 2016).

Wiersma, Wybo, John Nerbonne & Timo Lauttamus. 2011. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing* 26(1), 107–124.

Wong, Sze-Meng Jojo & Mark Dras. 2009. Contrastive analysis and native language identification. *Proceedings of the Australasian Language Technology Association*, 53–61. Cambridge: MA: Association for Computational Linguistics.

Yip, Po-Ching & Don Rimmington. 2004. *Chinese: A Comprehensive Grammar*. London: Routledge.