Learner Corpus Research: LCR2013 Conference Proceedings

BeLLS 2015 Vol. 6, ISBN 978-82-998587-7-9, s. 23–44 BeLLS.uib.no

Tracing crosslinguistic influences in structural sequences: What does key structure analysis have to offer?

Ilmari Ivaska
University of Washington / University of Turku

Abstract

Following the detection-based approach, this article detects statistically significant frequency differences between the data of written Finnish by learners from various language backgrounds. It analyses crosslinguistic influences in a data-driven manner, as the analysis focuses on the morphological forms and their combinations (n-grams) that prove to be the best predictors of differing first languages. Following the methodology applied – key structure analysis – the article then goes on to analyse the found n-grams in terms of their inner and cotextual variation in order to find out which linguistic phenomenon actually distinguishes the subsets of data. The results show several quantitative differences that may be due to the crosslinguistic influences and they were all detected in a data-driven manner without hypotheses of potential differences. The method can be useful especially in finding and analysing elusive crosslinguistic influences that cannot be interpreted to be transferred directly from the respective first languages.

Key words: Finnish as a second language, crosslinguistic influence, detection-based approach, key word analysis, key structure analysis

* *Principle contact:*
Ilmari Ivaska, Visiting lecturer
University of Washington, USA/University of Turku, Finland
Tel.: 1-206-543-6883
E-mail: ilmari@uw.edu

Bergen Language and Linguistic Studies (BeLLS), May 30th 2015. © Ivaska
DOI: 10.15845/bells.v6i0.807

1. Introduction

Language transfer or crosslinguistic influence (CLI) is a phenomenon that has interested the studies of second language acquisition (SLA) to a varying degree throughout the existence of the scholarly study of learning second or foreign languages. The key interest is in how and to what extent does a person's knowledge of one language (often the first acquired language; L1) influence the same person's knowledge and behaviour in another language (a second, third, or n:th language; L2) (e.g. Odlin 1989; Jarvis and Pavlenko 2008). Despite the keen interest in CLI, there has until recently been, according to Jarvis, "a surprising level of confusion in the field concerning when, where, in what form, and to what extent L1 influence will manifest itself in learners' use or knowledge of a second language" (Jarvis 2000, 246).

In this article I will show how certain aspects in the study of CLI can be approached in a data-driven manner by means of a methodological procedure called key structure analysis (Ivaska and Siitonen 2011; Ivaska 2012, 2014a). My research questions are: 1) how can indicators of potential CLI in linguistic structures be detected without pre-existing hypotheses of the differences, and 2) how to decide which are the features and linguistic phenomena in which learners with different L1 backgrounds differ from each other. By answering these questions it is possible to focus on the contrastive analysis of specific features of the potentially affecting languages to reveal the most likely reasons of the difference. This can, in turn, facilitate the methodological procedure and help in constructing valid hypotheses to be tested. The methodology applied takes advantage of recurring frequency differences of linguistic features that correlate with the L1 backgrounds of the learners. The focus of this article is not to conduct any contrastive analysis of different languages or find typological or other crosslinguistic explanations for the found differences. Instead, the aim is to demonstrate how can these differences be approached from within, in a data-driven manner. In other words, the typical use of the detected differences correlating with the L1 backgrounds is analysed in greater detail. The underlying idea is that a thorough quantitative and qualitative analysis of the typical linguistic behaviour in the L2 makes it easier in the end to look for the reasons for this behaviour in the language background of language users – knowing what to know is of major importance. This becomes even more critical when there is a clear statistical difference between certain L1 groups, but the languages compared do not offer any straightforward contrastive explanation for it. What is more, narrowing down the type of the difference also makes it easier to take into account the influences even of languages where the researcher does not have competence. Thus, the methodology strives for connecting the findings of a detection-based approach to linguistic units of constructional or phraseological natures in a data-driven manner.

In this article, I include analysis of texts written in Finnish by native speakers from five different language backgrounds, comparing these texts with texts written by native speakers of Finnish. In doing so, I will demonstrate how key structure analysis may be used as a guiding procedure in conducting research. The aim is to detect L1-related differences quantitatively, but the analysis goes on to detect the linguistic phenomena that actually underlie the differences, or alternatively, rules out those that do not. The article is structured as follows: Section 2 links the present study to earlier research and Section 3 introduces the data and the methodological procedure. Findings are reported in Section 4, followed in Section 5 by a discussion of the results and the wider applications for the method.

2. Background

2.1. Detection-based approach towards crosslinguistic influence

According to Jarvis, it is of major importance to focus on the features that show "statistically significant correlation (or probability-based relation) […] between some feature of learners' IL [here, L2] performance and their L1 background" (Jarvis 2000, 252). To postulate a clear methodological framework for identifying CLI, Jarvis (ibid.; refined in Jarvis 2010) lists the effects that need to be taken into account to make any justifiable claims about CLI. These are:

1. Similarities in L2 behaviour among people with a shared L1 background
2. Differences in L2 behaviour between people with different L1 backgrounds
3. Similarities in a given L1 or given L1s and the L2
4. Differences in other L1(s) in question and the L2 (added in Jarvis 2010)

In search of methodological rigour for identifying CLI, Jarvis distinguishes two different types of arguments. While the 'comparison-based' argument requires and relies on all the four aforementioned types of evidence, the 'detection-based' argument only takes into account the first two: intragroup homogeneity and intergroup heterogeneity (Jarvis 2012, 5-7). Although the former argument can provide very strong evidence for CLI, it can easily overlook cases where CLI is present but where the L1 and L2 do not – at least superficially – act similarly. In other words, the precision of the comparison-based argument is very high but the recall may not be so (ibid.). Detection-based approaches have been applied to find such unclear cases of CLI. Having their roots in data mining, detection-based approaches are commonly used in SLA to automatically detect either nativeness or L1 background of a language user from a text sample which is compared with a variety of text samples covering all the alternatives considered possible (e.g. Mayfield Tomokyio and Jones 2001; Koppel, Schler og Zigdon, 2005 Wong and Dras 2009; Jarvis 2011; Jarvis and Crossley 2012; Pepper 2012). Detection-based approaches are, thus, examples of a 'supervised learning task'. This kind of unclear CLI could be detected relying solely on the statistical tools applied in the detection-based research design. It is, however, of major importance to determine what distinguishes the varieties under comparison, so as to discuss the potential sources of CLI. Before looking for differences or similarities in the respective L1s, one should thus analyse the typical behaviour of the linguistic features in question. For example, if the statistically compared units are word frequencies, possible questions include whether any differences uncovered are due to a specific lexical item, or to a group of semantically or formally similar items. It is also worth investigating whether differences result from the varying behaviour of a particular syntactic function, such as tense distribution. To summarise, the typical variation of the item and its typical environment should be analysed in order to decide what is the underlying nature of the detected difference (Francis 1993).

Many of the applications of the detection-based approach regarding CLI and SLA in general have focused on either solely reporting the statistically significant frequency differences (e.g. Pepper 2012) or making direct qualitative linguistic interpretations on the reasons for the observed frequency differences (e.g. Wiersma, Nerbonne and Lauttamus 2011). The item and the environment can, however, also be analysed in a data-driven manner. One possibility to analyse the typical inner variation of any given linguistic structure is collostructional analysis (Stefanowitch and Gries 2003), which aims at combining the strengths of corpus approaches and those of construction grammar, where one analyses the lexical variation of certain constructions and co-occurrence tendencies therein. Study of the typical environment can, in turn, be approached by analysing the typical cotextual features of

the detected items. Collocational and colligational preferences can shed light on the typical use and behaviour of the item in question, and the observed differences between subsets of data can help us better understand the nature of possible CLI (see Hunston 2001; Sinclair 1991, 115–119).

2.2. Linguistic annotation as a basis for analysing constructional features

One focus of learner focus research has been on the phraseological nature of language (cf. Granger 1998; Nesselhauf 2004; Meunier and Granger 2008), where co-occurring words are seen to constitute n-grams, i.e. multiword units (or bunches or clusters) of length 'n'. Granger and Paquot (2008) suggest that these should be further divided into 1) qualitatively defined phraseological units and 2) frequency-based n-grams characterised in terms of their distributional features and the relationship between their overall occurrence and co-occurrence of any given words. This idea of repeatedly co-occurring patterns of linguistic features as such is by no means unique to learner corpus research. It is essentially an example of analogy, which in turn is often considered a fundamental construct of all linguistic behaviour (e.g. Skousen 1989; Itkonen 2005). The co-occurrence patterns of linguistic features are also typical in linguistic descriptions following a construction grammar approach, where recurrence is considered the core requirement of a construction (e.g. Goldberg 2006). In this sense, the phraseological approach and construction grammar are in many respects similar despite the terminological differences, as has been pointed out by Gries (2008).

Many corpuslinguistic studies of learner language that focus on structural or compositional behaviour approach the topic with the lexical level as the basis. As Jantunen mentions, this often leaves more abstract structural features beyond the scope of analysis (Jantunen 2009, 361). Some applications aim at analysing the differences between native and non-native writers of English or distinguishing writers of English with different L1 backgrounds through a combination of part-of-speech (POS) annotation and an n-gram approach (e.g. Aarts and Granger 1998; Wiersma, Nerbonne and Lauttamus 2011). Jantunen (2011) and Ivaska (2014a) have used linguistic annotation in a similar manner in their studies of L2 Finnish. With the development and better availability of corpora (cf. Université catholique de Louvain. Centre for English Corpus Linguistics. n.d.) and natural language processing tools (cf. Stanford University. Natural Language Processing. n.d.), this type of combined approach may become more common, although the scope of the analysis depends mostly on the interests of individual researchers as well as on the language(s) being studied and the corpus at hand.

It is important to keep in mind that the majority of the aforementioned descriptions make at least the implicit assumption that the words of an n-gram follow each other consecutively and in the same order. As Sinclair points out, this may leave unnoticed typical variation or discontinuity of the structure under investigation (Sinclair 2001, 353). However, n-grams can also be analysed and counted as skip-grams, where the words do not have to follow each other consecutively, as long as they are in the same order and close to each other (Guthrie et al. 2006), or as concgrams, where the only criterion is the proximity of the words (Cheng, Greaves, and Warren, 2006). In addition to making it easier to take possible structural variation into account, typical patterns of language may thus be detected, even with smaller data sets (Guthrie et al. 2006, 1225).

3. Methodology

3.1. Data: Corpus of Advanced Learner Finnish

The data used in this study is a part of the Corpus of Advanced Learner Finnish (LAS2, Ivaska and Siitonen 2009; Ivaska 2014b), compiled at the University of Turku. All the data is written by L2 learners of Finnish studying in Finnish, at a Finnish university. The majority of the topics fall into linguistics or other fields in the humanities, and the texts were not written with the sole aim assessing the linguistic skills of the informant. Two texts from each informant have been assessed based on the Common European Framework of References (CEFR, Common European framework for languages: Learning, teaching, assessment 2006) proficiency levels, and the assessments are all divided between B2 and C2 (distribution of the data assessed so far: B2: 37%, C1: 61% and C2: 2%). Based on the production context and the results of the partial assessment of the data, the texts may be considered to represent advanced L2 Finnish.

All the data in the present study are exam essays. There are 20 texts from learners of five different L1 backgrounds and 20 texts from a reference corpus of L1 Finnish informants, a total of 120 texts. Details are presented in Table 1.

Table 1. The data set used in the present study

L1 background	Texts	Nr. of informants	Tokens
Finnish (fi)	20	20	10,283
Czech (cs)	20	4	8,745
Hungarian (hu)	20	4	10,885
Japanese (ja)	20	4	8,498
Lithuanian (lt)	20	4	11,397
Russian (ru)	20	4	9,721
In total:	120	40	59,529

The LAS2 has been lemmatised and annotated for parts of speech, morphological forms and syntactic functions. The annotation has been done semi-automatically as follows: each occurring word form type was annotated manually in terms of the most probable lemma, part of speech and morphological form; this information was added to the data; the data was then annotated syntactically with a probabilistic annotator created for the LAS2; finally the data was corrected manually by a single annotator. In addition, the corpus was annotated in terms of the paragraph structure and divided into sentences, clauses, and words. The corpus has been stored in a XML format that has been applied also in other Finnish corpora collected at the University of Turku (Lauseopin X-arkisto. n.d.; Inaba 2007). All data queries were carried out with query tools designed specifically for the LAS2 corpus (Ivaska 2014b) and all statistical analyses with conducted with the help of R (R Core Team 2013). Whenever the texts were compared with each other, frequency observations of each text were first normalised per 1,000 tokens so that each text unit is of equal importance in distributional descriptions as well as in statistical analyses.

3.2. Procedure: Key Structure Analysis and statistics applied

The procedure applied in this study follows the idea of key structure analysis (Ivaska 2012, 2014a), which in turn combines the detection-based concept of keyness in the keyword analysis (Scott and Tribble 2006) with the analysis of the item and its environment (Francis 1993). Roughly said, recurring quantitative differences between different L1-specific subsets of data are detected statistically to find potential occurrences of CLI. Then those linguistic

features found to show differences are analysed in greater detail in terms of the inner and cotextual variation[1] and possible L1-specific differences. Revealed tendencies of the given linguistic context together with the situational context in which they occur constitute contextual profiles that can be used to depict typical behaviour of the linguistic phenomenon in question as well as the possible differences in different subsets of data (Jantunen 2004; Ivaska 2014a). Items and their environment can be analysed both in the data as a whole and in the L1-specific subsets of data separately, and statistical tools for finding the distinguishing features can be applied. These steps together – the detection of statistically significant differences and the analysis of the typical behaviour of the features found to correlate with those statistical differences – constitute the core of the key structure analysis. From the point of view of the two aforementioned types of arguments, comparison-based and detection-based, it is relevant to point out that a thorough analysis of an item and its environment may be helpful when looking for the less transparent and more elusive forms of CLI. In this sense, then, key structure analysis can be used to link the observed behaviour with typological features of the languages in question. The method in itself is not tied to any specific theoretical framework, although a quantitative data-driven approach is in many ways in line with the usage-based nature of linguistic system (for summary of the usage-based models, cf. Barlow and Kemmer 2000). Further, much like in construction grammar, the linguistic phenomena revealed during the analysis can be defined as form–function pairs, which is why the units analysed in the scope of the present paper are called constructions.

I queried the data for the frequencies of all the morphological forms and their combinations occurring in the data by extracting the frequencies of single words (1-grams), two-word combinations (2-grams) and three-word combinations (3-grams). The skip-gram approach was applied in defining two- and three-word combinations, so the words do not have to occur consecutively in the data as long as they are in the same order and close to each other. In the present study I did not define any maximum distance for skip-gram members, although they had to occur in the same clause unit. As an example, the clause below consists of four words and of four morphological 1-grams, six morphological 2-grams and four morphological 3-grams.

	**Lapset*	*pitävät*	*nähdä*	*rahaa*.
lemma:	child	to like	to see	money
mrp:	<pl nom>	<fin ind pres pl3>	<inf1>	<sg part>
	‘Children like to see money’			

Skip-grams:

1-grams: <pl nom> (plural nominative)
<fin ind pres pl3> (present tense plural 3rd person finite verb in the indicative)
<inf1> (A-infinitive, in other words, the 1st infinitive)
<sg part> (singular partitive)

2-grams: <pl nom><fin ind pres pl3>
<pl nom><inf1>
<pl nom><sg part>

[1] Co-textual variation refers to the close linguistic context an item occurs in. In this article the span analysed ranges from two words left to two words right of the item in question, henceforth referred to as L2, L1, R1 and R2 positions.

<fin ind pres pl3><inf1>
<fin ind pres pl3><sg part>
<inf1><sg part>

3-grams: <pl nom><fin ind pres pl3><inf1>
<pl nom><fin ind pres pl3><sg part>
<pl nom><inf1><sg part>
<fin ind pres pl3><inf1><sg part>

Following this counting method, all the gram frequencies were extracted using the feature frequency tool of the LAS2. When there was no morphological information assigned, the tool extracted the part-of-speech tag instead. The aim was to focus on features that are fairly common in the data so that, in line with Jarvis, Castañeda-Jimenez, and Nielsen (2012) and Pepper (2012), the set of features considered contains only grams that are among the 30 most frequent grams in at least one of the five L1 groups' data. At this stage, the frequency counts meeting the threshold level were also normalised and saved in a machine-readable csv spreadsheet.

The different L1 backgrounds in the L2 part of the data were then compared with each other. I listed all the grams under investigation in decreasing order of importance in terms of how well they can predict different L1 backgrounds based on the gram frequencies, and focused all further analysis on the 30 best predictors. The statistical approach applied is called "random forest" (for the algorithm, cf. Breiman 2001). Random forests can be used in classification or regression to find the best predictor of any certain feature in the data (e.g. Strobl, Malley, and Tutz 2009; for examples in linguistics, e.g. Tagliamonte and Baayen 2012). There are several implementations of random forests in R, of which I used the version found in "cforest" method in R, as did Tagliamonte and Baayen (ibid.). The process goes roughly as follows: a sample of the data is taken aside (out-of-bag sample, roughly 33% of all the observations), and the rest (in-bag sample) is used to choose randomly a subset of variables to construct a conditional inference tree, i.e. it is used to look for the best predictor of the wanted classification. This random sampling of in-bag observations is repeated a great number of times to find the most reliable predictors. Finally the predictive power of the model is tested on the out-of-bag observations. The model can then be subjected to a method called "varimp" that can be used to obtain a permutation-based variable importance measure. It produces a ranked listing of the variables in the order of their accuracy in classification. All the aforementioned functions are found in the "party" package of R (Hothorn et al. 2006; Strobl et al. 2007, Strobl et al. 2008). The method is considered effective in choosing variables to be focused on. It is also highly practical, as it inherently does cross-validation with the out-of-bag sample. In order to reduce the effect of the random selection, the process was repeated five times with different seeding for the random selection, resulting in a variable importance listing based on the means of the five runs.[2]

As the same word can occur both as a 1-gram and as a part of a 2- or a 3-gram, gram frequencies can easily be collinear. It is also possible that two separate unigrams are significant separately and occur together in a 2-gram of 3-gram by chance. To ensure that the

[2] The seedings used for the random selection were 97531, 75319, 53197, 31975 and 19753. The response variable was the L1 of the informant and the predictors considered were frequencies of all the morphologically defined grams that met the threshold level. I used the default settings otherwise but due to the relatively high number of considered variables the sampling (the number of the inference trees) was repeated 10,000 times in each forest.

analysis focuses on the correct gram length, I took in such cases into consideration and analysed the shortest occurrence of the gram among the 30 best predictors.

The L2 data contains, on average, five texts from each informant and the observations therefore are not fully independent. Unfortunately the data at hand is too small to only have one text for each informant. To ensure that the differences analysed are due to inter-L1 differences rather than inter-learner differences I tested each potentially meaningful gram with another random forest.[3] In my analysis I took into consideration only those grams in which L1 background was the better predictor for gram frequency than informant ID. After sorting out the best predictors I then analysed how frequencies of different L1 backgrounds actually differ from each other predictor by predictor, and how they behave when compared to L1 Finnish. In this post-hoc test phase, I used Tukey HSD, which compares data set means in order to find out which L1 backgrounds show statistical separation and which are grouped together (for a discussion about the different statistics applied for similar tasks in earlier research, cf. Pepper 2012, 64).

After that, I analysed the grams found to show statistically significant frequency differences between different L1 groups in terms of their contextual profiles. The analysis is chiefly contrastive in nature, and it conceptually follows the idea of contrastive interlanguage analysis (CIA, Granger 1996, 2013). I focused mainly on the comparison between the different L1 groups that showed statistically significant differences in Tukey HSD. In order to find the typical use and the probable linguistic phenomenon – e.g. a phraseological unit or a construction – I also looked at the typical use of the elements in all the data. In other words, I descriptively analysed the typical use of the found grams and then looked at the most remarkable differences in the contextual profiles between the subsets found in the post-hoc test. I detected the best predictor of the difference using again random forests[4]. For computational reasons and for clarity, a maximum of five of the most frequent values of each variable (e.g. the five most common lemmas in L2 position, and so forth) were specified, and the less frequent values were considered as "other".

4. Results

4.1. Frequency differences indicating crosslinguistic influences in advanced Learner Finnish

There were all in all 126 different 1-, 2- or 3-grams of morphological forms that were in the pool of the 30 most frequent grams of any of the data sets. Table 2 presents the 30 best predictors of L1 background detected in the statistical analysis.

[3] The seeding used for the random selection was 7531. The feature to be predicted was the gram in question and the possible predictors considered were informant's L1 background and informant ID. The sampling was repeated 2,000 times.

[4] The seeding used for the random selection was 7531. The feature to be predicted was the grouping found in each post-hoc test and the possible predictors considered were inner lexical variation of the n-gram (and, when applicable, also variation in syntactic function and POS), and cotextual features, i.e. lexical variation, POS variation, morphological variation and variation in syntactic functions in positions L2, L1, R1 and R2 of each gram of the n-gram. Thus, there were 17 to 19 possible predictors for each gram of the n-gram, depending on the nature of the actual n-gram in question. The sampling was repeated 2,000 times.

Table 2. Grams that best predict the L1 background of the writer in descending order of importance. Comparison between the predictability of L1 background and informant ID has been conducted only for grams that do not contain other top 30-grams.

Variable importance rank	n-gram*	Contains other top 30-grams	L1 background better predictor than informant ID
1	<cnj>	no	yes
2	<fin ind pres sg3><adv>	yes	
3	<pl part><pl part>	no	no
4	<adv><adv>	yes	
5	<cnj><fin ind pres sg3>	yes	
6	<sg gen><sg gen><sg gen>	yes	
7	<sg gen>	no	no
8	<sg gen><sg gen>	yes	
9	<adv>	no	no
10	<sg ine><cnj>	yes	
11	<sg nom><pl part>	no	yes
12	<cnj><sg nom><fin ind pres sg3>	yes	
13	<sg nom><cnj>	yes	
14	<sg gen><pl part>	yes	
15	<cnj><sg nom><adv>	yes	
16	<sg ine>	no	yes
17	<cnj><sg gen>	yes	
18	<sg gen><sg nom>	yes	
19	<adv><sg nom><sg nom>	yes	
20	<cnj><cnj><sg gen>	yes	
21	<cnj><sg nom>	yes	
22	<sg nom><fin ind pres sg3><adv>	yes	
23	<adv><cnj><sg nom>	yes	
24	<sg ine><sg ine>	yes	
25	<cnj><sg ine>	yes	
26	<sg ine><sg nom><cnj>	yes	
27	<fin ind pres sg3>	no	yes
28	<fin ind pres sg3><sg ine>	yes	
29	<cnj><sg nom><sg gen>	yes	
30	<sg nom><adv>	yes	

* The used abbreviations are following (in the alphabetical order): adv 'adverb', cnj 'conjunction', fin 'finite', ind 'indicative', ine 'inessive', nom 'nominative', part 'partitive', pl 'plural', pres 'present tense', sg 'singular', sg3 'singular 3rd person'.

After sorting the grams in decreasing order of importance, taking into account only the shortest possible gram length and leaving out the grams in which inter-learner differences explain frequency variation better than inter-L1 differences, there were all in all three 1-grams and one 2-gram to focus on: 1-grams consisting of a conjunction (<cnj>), of a singular inessive (<sg ine>), and of a finite indicative present tense verb in singular 3rd form (<fin ind pres sg3>) and a 2-gram consisting of a singular nominative followed by a plural partitive (<sg nom><pl part>). I then carried out the Tukey HSD as a post-hoc test for each gram detected as possible indicator of L1 group differences. As the test on singular inessives showed no statistically significant differences, I did not analyse it in more detail. At this point

I also used the L1-fi data set to see how the different L1 groups compared to the reference data.

4.2. Use of conjunctions

As indicated in the variable importance listing, the use of conjunctions (<cnj>) is the best predictor of L1 group. Levene's test did not indicate any statistically significant difference in the homogeneity of variance between the L1 groups ($F(5) = 0.929$, $p = .465$) so I proceeded to the Tukey HSD. The results show indeed that conjunctions are less frequent in L1-cs, L1-lt and L1-ru than in L1-hu, L1-fi and L1-ja. The data groups split clearly into two subgroups, and there are no statistically significant frequency differences within the subgroups while the differences between the subgroups are all statistically significant. Details of the test results can be seen in Table 3, while the boxplots in Figure 1 depict the variance within L1 groups and the differences between the groups.

*Table 3. Tukey's HSD test results in the frequency of conjunctions. L1-groups are ordered based on the observed frequencies and statistically significant differences are marked with *.*

L1-group	**L1-group pair**	**Difference in mean / 1,000 w**	**p-value**
lt	cs (76.5)	2.1	p = .999
(74.4 / 1,000 w)	ru (76.5)	2.1	p = .999
	hu (94.5)	20.0	p =.018*
	fi (95.6)	21.1	p = .010*
	ja (97.3)	22.9	p = .004*
cs (76.5)	ru (76.5)	2.1	p = 1.000
	hu (94.5)	20.0	p = .046*
	fi (95.6)	21.1	p = .028*
	ja (97.3)	22.9	p = .012*
ru (76.5)	hu (94.5)	18.0	p = .046*
	fi (95.6)	19.1	p = .028*
	ja (97.3)	20.8	p = .012*
hu (94.5)	fi (95.6)	1.1	p = .999
	ja (97.3)	2.9	p = .997
fi (95.6)	ja (97.3)	1.8	p = .999

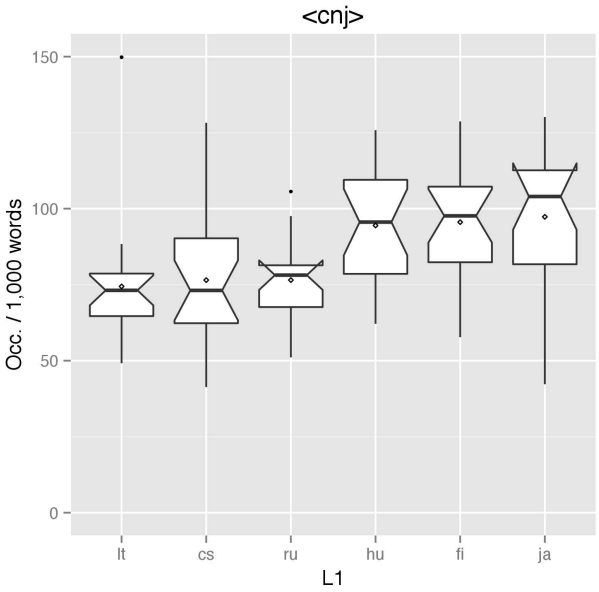

Figure 1. Frequency distribution of conjunctions in the different subsets of data

The most common conjunction is *ja* 'and', which accounts for 38% of all the conjunctions in the data. In general, conjunctions act typically either as clausal conjunctions (*ja* 'and' in example 1) or as phrasal conjunctions (*ja* 'and' in example 2).

1. […]	ja	mi-tä	ne	osta-vat	esim	TV:-ssa
	and	what-PTV	they.PL.NOM	buy.FIN.IND.PRS-1PL	e.g.	TV-INE

'[…] and what they buy for example in television'
(L1-cs: las2-25tt01te05)

2.	Vaikka	väli-ssä	on	i	ja	e […]
	though	between.SG-INE	be.FIN.IND.PRS.3SG	i	and	e

'Though there is an i and e in between […]'
(L1-fi: las2-vtt01vert128)

I then looked for the best contextual predictors of the observed difference in frequency by comparing the subgroups in terms of the inner and the cotextual variation with the help of another random forest. The best predictor for the difference is the inner lexical variation: lemmas acting as conjunctions. The distributions of the most common lemmas can be found in Table 4. The distributional variation does not reveal any clearly differing patterns. It shows that two most common conjunctions are more common in L1-lt, L1-cs and L1-ru data than in the rest of the data, but the distributional features are very similar. What is more, the five most common lemmas contain both clausal and phrasal conjunctions and both coordinating and subordinating conjunctions.

Table 4. The five most common conjunctions and their distributions in the two data subgroups

Lemma	L1-lt, L1-cs and L1-ru	L1-hu, L1-fi and L1-ja
ja 'and'	40%	36%
että 'that'	17%	14%
mutta 'but'	9%	12%
tai 'or'	6%	8%
kun 'when'	5%	6%

The second best predictor for the difference is syntactic functions that occur right after conjunctions (R1 position). The distributions of the most common functions can be found in Table 5. These functions do not reveal any clearly differing patterns, and despite the overall difference in frequency the contextual behaviour is very similar in L1-hu, L1-fi and L1-ja, and L1-lt, L1-cs and L1-ru, respectively.

Table 5. The five most common syntactic functions in R1 position of conjunctions and their distributions in the two data subgroups

Syntactic function in R1	L1-lt, L1-cs and L1-ru	L1-hu, L1-fi and L1-ja
modifier of a noun	27%	26%
nominal subject	22%	25%
adverbial	18%	17%
predicate	14%	10%
nominal object	5%	6%

What becomes clear is that the difference in the use of conjunctions is not due to any single recurring phraseological unit or construction. Possible explanations could include differences in the use of conjunctions in unambiguous contexts or other differences in academic writing traditions, such as clarity and complexity in phrase structure. This remains, however, to be explored in the future studies.

4.3. Use of finite present tense verb in the 3rd singular

As seem in the variable importance listing in Table 2, the use of finite present tense verbs in the 3rd singular (<fin ind pres sg3>) is the third best 1-gram predictor of the L1 group. Levene's test did not indicate any statistically significant difference in the homogeneity of variance between the L1 groups ($F(5) = 0.835$, $p = .528$) so I proceeded to the Tukey HSD. The results show that this 1-gram is clearly less frequent in L1-lt than in L1-fi and L1-hu and that the difference between these two subgroups is statistically significant. Details of the test results can be seen in Table 6 while the boxplots in Figure 2 depict the variance within all the L1 groups and differences between the groups.

*Table 6. Tukey's HSD test results in the frequency of finite present tense verb in the 3rd singular. L1-groups are ordered based on the means of observed frequencies and statistically significant differences are marked with **

L1-group	**L1-group pair**	**Difference in mean / 1,000 w**	**p-value**
lt	cs (77.3)	13.1	p = .714
(64.2 / 1,000 w)	ru (77.3)	13.1	p = .713
	ja (82.0)	17.7	p =.390
	fi (93.8)	29.5	p = .020*
	hu (94.9)	30.7	p = .014*
cs (77.3)	ru (77.3)	<0.1	p = 1
	ja (82.0)	4.7	p = .996
	fi (93.8)	16.5	p = .477
	hu (94.9)	17.6	p = .397
ru (77.3)	ja (82.0)	4.7	p = .996
	fi (93.8)	16.4	p = .478
	hu (94.9)	17.6	p = .398
ja (82.0)	fi (93.8)	11.8	p = .794
	hu (94.9)	13.0	p = .721
fi (93.8)	hu (94.9)	1.1	p = .999

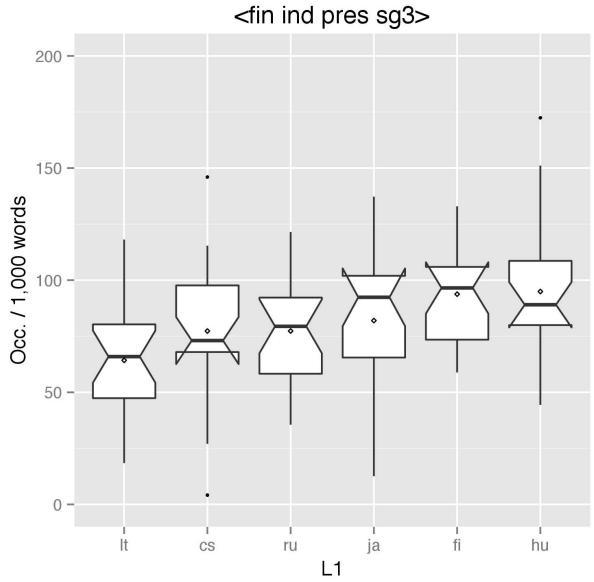

Figure 2. Frequency distribution of finite present tense verbs in 3rd singular in the different subsets of data

The 1-gram in question is either the main verb of a present tense clause (*vaikuttaa* 'affects' in example 3) or the auxiliary verb of a present perfect tense clause (*on* 'is' in example 4).

3. ammatti vaikutta-a kova-sti
profession.SG.NOM affect.FIN.IND.PRS-3SG hard.SG-ADV
'profession affects a lot'
(L1-ja: las2-18tt01te10)

4. konjunktiivi-sta on kehitty-nyt *-isi-*
conjunctive.SG-ELA be.FIN.IND.PRS.SG3 develop-PTCP2.SG.NOM *--isi-*
'[morpheme] *-isi-* has developed from the conjunction'
(L1-lt: las2-24tt01te08)

I then went on to look for the best contextual predictors of the difference between L1-lt data and L1-fi and L1-hu data. The best predictors for the difference are syntactic functions immediately after the 1-gram (R1 position). The distributions in the two subsets of data are found in Table 7.

Table 7. The five most common syntactic functions in R1 position of finite present tense verb in the 3rd singular and their distributions in the two data subgroups

Syntactic function in R1	**L1-lt**	**L1-fi and L1-hu**
modifier of a noun	23%	18%
second element of a predicate	22%	15%
adverbial	18%	25%
subject complement	10%	11%
nominal subject	4%	7%

The clearest difference is that the second element of a predicate is more common in L1-lt and that an adverbial is more common in L1-fi and L1-hu. Given that the semantic main verb of the present perfect is coded as second element of a predicate in the LAS2 it seems probable that the difference is due to the fact that present perfect cases account for more of the 1-grams in question in L1-lt than in L1-fi and L1-hu. This is supported by the observation that the second participle forms, the morphological form of the main verb of the present perfect, are more than twice as common in R1 position in L1-lt than in L1-fi and L1-hu. Obviously this does not mean L1-speakers of Lithuanian would use more present perfect tense but rather that they use less present tense than L1-speakers of Finnish and Hungarian. Reasons for this may be due to the different aspectual features of tenses or different conventions in academic writing. Interpreting this is, however, again beyond the scope of the present study.

4.4. Use of singular nominative followed by plural partitive

The third and last n-gram to be analysed is a 2-gram consisting of a singular nominative (<sg nom>) followed by a plural partitive (<pl part>). Levene's test did not indicate any statistically significant difference in the homogeneity of variance ($F(5) = 1.993$, $p = .085$) so I proceeded again to the Tukey HSD. The results show that there is a statistically significant difference between L1-ja, L1-cs and L1-hu data when compared to L1-fi data. L1-lt could also be included in the analysis, as the difference between it and L1-fi is almost significant ($p = .089$), but for the sake of uniformity the closer analysis focuses on the differences between the subsets showing statistically significant differences. Details of the test result are found in Table 8, and boxplots in Figure 3 depict the frequency variance in all the L1-groups.

*Table 8. Tukey's HSD test results in the frequency of 2-gram consisting of a singular nominative and a plural partitive. L1-groups are ordered based on the means of observed frequencies and statistically significant differences are marked with *.*

L1-group	L1-group pair	Difference in mean / 1,000 w	p-value
ja	cs (9.8)	0.4	p = .999
(9.4 / 1,000 w)	hu (9.8)	0.5	p = .999
	lt (11.9)	2.5	p = .995
	ru (17.1)	7.7	p = .095
	fi (24.7)	15.3	p = .022*
cs (9.8)	hu (9.8)	0.1	p = 1
	lt (11.9)	2.1	p = .997
	ru (17.1)	7.3	p = .116
	fi (24.7)	14.9	p = .027*
hu (9.8)	lt (11.9)	2.1	p = .998
	ru (17.1)	7.3	p = .119
	fi (24.7)	14.9	p = .028*
lt (11.90)	ru (17.1)	5.2	p = .287
	fi (24.7)	12.8	p = .089
ru (17.1)	fi (24.7)	7.6	p = .993

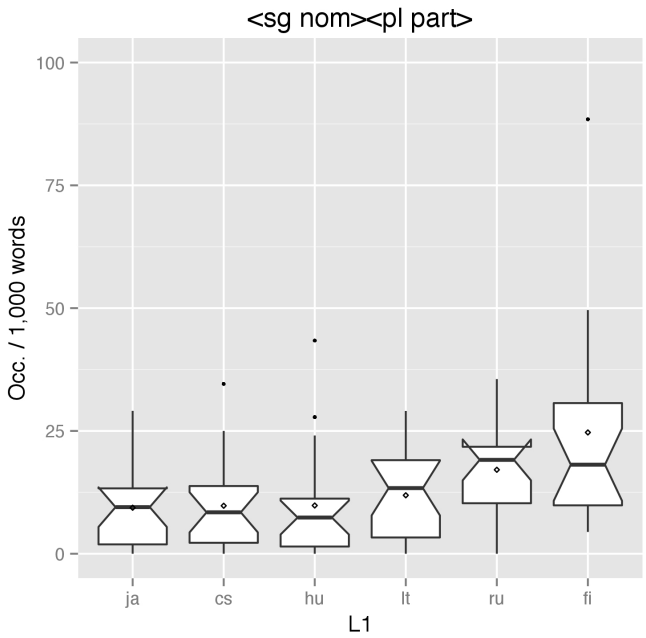

Figure 3. Frequency distribution of singular nominatives followed by plural partitives in the different subsets of data

The first word of the 2-gram (<sg nom>) is typically either a nominal subject (*hän* '(s)he' in example 5, *osa* 'part' in example 6) or a modifier of a noun (*suuri* 'big' in example 6) and the second word (<pl part>) is typically either a nominal object (*asioita* 'things' in example 5), a modifier of a noun (*muita* 'others' in example 5) or a subject complement (*suomenkielisiä* 'Finnish speaking' in example 6).

5. jos hän mu-i-ta asio-i-ta osaa […]
if (s)he.SG.NOM other-PL-PTV thing-PL-PTV can.FIN.IND.PRS.3SG
'if (s)he can do other things'
(L1-hu: las2-10tt01te01)

6. suuri osa kuulijo-i-sta o-vat suomenkielis-i-ä
big.SG.NOM part.SG.NOM listener-PL-ELA be.FIN.IND.PRS-3PL Finnish-PL-PTV
'major part of the listeners are speakers of Finnish […]'
(L1-lt: las2-30tt01te06)

I then went on to look for the best contextual predictors of the difference between L1-ja, L1-cs and L1-hu data and L1-fi data. The best predictor for the difference is the part of speech occurring immediately before the first word of the 2-gram (L1 position). The distributions are found in Table 9.

Table 9. The five most common parts-of-speech in L1 position of singular nominative and their distributions in the two data subgroups

POS in L1	**L1-ja, L1-cs and L1-hu**	**L1-fi**
clause boundary	27%	27%
conjunction	18%	14%
verb	16%	10%
adverb	14%	7%
noun	11%	36%

It is clear that nouns are much more common in the preceding cotext in L1-fi than in L1-ja, L1-cs or L1-hu. Closer qualitative analysis reveals that this is due in most cases to a certain kind of listing construction ('Xs X_1, X_2… and X_n), in which modifiers of a noun occur after the word they are modifying, contrary to the prototypical Finnish word order. Example 7 exemplifies this construction.

7. Sana-t järvi lampi meri ja joki
word-PL.NOM lake.SG.NOM pond.SG.NOM sea.SG.NOM and river.SG.NOM

kuvaa-vat kaikki erilais-i-a vesistö-j-ä
describe.FIN.IND.PRS-3PL all different-PL-PTV water_system-PL-PTV

'words lake, pond, sea and river all describe different kinds of water systems'
(L1-fi: las2-vtt01vert096)

The construction permits more than one modifier and they are all in the singular nominative in the construction, which is also why the second word of the 2-gram is in the plural. It is debatable whether the modifiers actually are nominative or only in an unmarked form, but this is the interpretation chosen in the LAS2. Further interpretations of possible CLI require a thorough analysis of the similar constructions in the respective languages. In any case, one might arrive at the conclusion that specific and probably infrequent constructions like the one

described here can easily be affected by prior linguistic knowledge of other structurally similar constructions.

5. Discussion and conclusions

In this study I have demonstrated how crosslinguistic influences can be approached by means of a methodological procedure called key structure analysis. Regarding the question of how potential crosslinguistic influences can be detected without pre-existing hypotheses, this study supports earlier studies in that statistical comparison of the frequencies of linguistic features is a reasonable point of departure. This is essentially also the first step of the key structure analysis. What is compared and how it is done depends largely on the language being studied, the data at hand and the interests of the researcher. In this study I compared frequencies of 1-grams, 2-grams and 3-grams of advanced L2 Finnish texts written by speakers of five different L1 backgrounds with each other. In order to also take into account structural variation, the frequencies were counted following the skip-gram approach, so the words did not have to follow each other consecutively to be counted as parts of the same n-gram. I then compared the frequencies statistically with the help of a supervised classification algorithm called random forest to find out the best predictors of differing L1 backgrounds. Due to the fact that Finnish is a morphologically rich language and word forms thus carry both semantic and syntactic information I decided to use morphological annotation as the basis of the analysis. Certain words in Finnish are not considered to carry any other morphological information than the word itself, and for these words I used part-of-speech annotation instead. Comparison of the observed frequencies revealed several n-grams that could possibly be used to distinguish different L1 backgrounds. I chose to more closely examine the 30 best predictors, but before analysing any of them in greater detail I shortened the list by removing all the 2- and 3-grams that contained any 1-grams included in the best predictors and after that all the 3-grams that contained any 2-grams. I did this in order to avoid pseudo-collinearity, as the same words are counted as 1-grams, 2-grams and 3-grams. What is more, it is likely that if two 1-grams included in the list occur in the same clause, a 2-gram containing the two words is high in the list by chance. After that I shortened the list even further by sorting out the n-grams in which a single informant ID is a better predictor of the difference than the L1-background. This left me with four n-grams of which one did not show any statistically significant differences in the post-hoc testing phase where also a reference data of L1 Finnish writers was used.

To be able to answer to the second research question on how to decide which are the actual features and actual linguistic phenomena behind the difference, I proceeded to next phase of the key structure analysis: the analysis of the typical inner variation of the feature and its typical cotextual variation. I analysed the three remaining n-grams in greater detail. It turned out that the L1 speakers of Czech, Lithuanian and Russian use conjunctions less frequently than those of Japanese, Hungarian and the reference group of L1 speakers of Finnish. Present tense finite verbs in the singular 3rd form are also clearly more frequent in the texts of L1 speakers of Hungarian and the Finnish reference group than they are in the texts of L1 speakers of Lithuanian. Finally, a 2-gram consisting of a singular nominative and a plural partitive is far more seldom in the texts of L1 speakers of Czech, Hungarian and Japanese than in the texts of the reference group. Closer examination revealed that the difference in the use of conjunctions probably is not due to any systematic phraseological difference, any one lexical item, any one syntactic function, or any one clausal or phrasal position. One possible explanation lies in the differences in writing conventions regarding when a conjunction is necessary. An earlier comparative study on the Finnish conjunction *ja* ('and') and its typical

equivalents in Lithuanian shows that the conjunction is more commonly used in Finnish and that the patterns of the way it is used in Lithuanian can also be transferred to L2 Finnish by L1 Lithuanian speakers (Penttilä 2008). Difference in the present tense 3rd person singular verbs was shown to be indeed due to the lesser use of present tense by Lithuanian speakers of Finnish and not due to other tenses containing present tense verbs. An earlier study has reported possible aspect-related CLI in the tense use of Czech and Russian learners of Finnish (Räisänen 2005), and while the texts of L1 Czech and Russian speakers did not show any statistically significant differences from the reference texts of L1 Finnish speakers, the difference is still worth noting. Closer study on the matter would require, besides the comparison of the tense-aspect systems of the respective languages, a thorough distributional analysis of all the tenses. Difference in the detected 2-gram was shown to be in most cases due to the use of a rather specific and non-prototypical listing construction in which modifiers of a noun are placed only after the noun, which is contrary to the typical word order of Finnish. I am not aware of any earlier study reporting similar results, but it seems reasonable that a construction like this is more likely to be used if a structurally similar construction is found also in the respective L1.

In general, the choices based on the procedure of the key structure analysis worked quite well, and the results support its applicability. None of the observed differences can be categorised as being due to CLI as such, and such interpretations require further contrastive analyses. What is more, different L1 populations should also be analysed separately, instead of only grouping together the ones behaving similarly. On the other hand, the method may detect single constructional or phrasal features that certain L1 populations use similarly, even if the actual L1s in question differ typologically and even if they do not belong in the same language families. This may, in turn, ease finding and analysing less clear and less uniform CLI. Further, in contrast to some previous detection-based studies (e.g. Wiersma, Nerbonne and Lauttamus 2011), I did not interpret the nature of the observed differences directly from the observed frequencies but analysed quantitatively also the typical use of the n-grams in question to form a data-driven interpretation.

The biggest shortfall undoubtedly is the small size of the data set. Unfortunately there are no bigger corpora available for learner Finnish that include data from advanced learners and that are as extensively annotated. Still, the statistical measures were chosen to fit the small data set and I was able to detect possible cases of CLI following the methodological procedure applied. In other words, the data did suffice to test the applicability of the method. It is likely that there are features that would have been found with a bigger data set but I doubt there were any false positive interpretations due to the size of the data. As previously discussed, using an annotation level as the starting point instead of raw text poses certain limitations to the data, but otherwise morphological annotation can probably better reveal possible CLI of a grammatical nature in a language like Finnish, where a good deal of grammatical information is presented as inflectional and conjugational elements. Theoretically, a morphology-based point of departure is also more natural from a usage-based perspective than parts-of-speech or syntactic functions, as it relies on concretely recurring items of language instead of any abstraction made for classification purposes. Defining the n-grams as skipgrams has two positive effects: first, a relatively small data set (as in this study) is enough to count frequencies that can be used statistically. Second, it makes it possible to also detect recurring multiword features that distinguish the data sets but that would not have been analysed using the traditional n-gram approach. For example, the analysed 2-gram would never have been detected without the skipgram approach. The chosen statistical

method, random forest, is a very handy procedure for linguistic applications in its robustness, as it does not have any distributional requirements and it can even deal with different kinds of variables (e.g. numeric and categorical) in the same model. What is more, it has an inherent cross-validation, or rather out-of-bag testing procedure, on which the reported results are based. The flexibility of the method is also very practical, as the same procedure could be used for detecting the best predictors, ruling out informant-specific features and sorting out the greatest contextual differences in typical behaviour of a linguistic item.

The downside of the method is that the results are not directly comparable with studies that apply other statistical methods (e.g. Jarvis and Crossley 2012). Further, different studies differ in what is considered to be the best classification statistics for language background recognition (cf. Estival et al. 2007; Jarvis 2011). In key structure analysis, this is not as big a problem as it may be in automatic classification tasks, since the analysis goes further than only stating detected differences, and the aim is not to guess the L1 of any text but to find the best predictors of differences between the groups. The analysis of the typical inner and cotextual variation makes it possible to find distinguishing features in a truly data-driven manner. The results of the present study also verify the point made by Scott with regard to key word analysis and keyness: the results of the statistical comparison are but indicators of where researchers should focus their interest (Scott 2010, 56-57). A good example of what this means can be seen in the analysis of the 2-gram with a singular nominative and a plural partitive. The likely difference is not due to the use of the forms themselves but rather to a construction related to those forms.

The present study raises questions for two main avenues of future research. First, it would be interesting (and necessary from the point-of-view of the comparison-based argument of CLI) to contrastively investigate the typical ways of expressing the analysed structures in the languages of different L1-groups of this study. The analysis presented here did narrow down the features that would need to be studied and, with this in mind, it could possibly be done with specified questionnaires answered by a few native speakers of each language. Thus it should not be an overwhelming task to take into account even languages the researcher is not formerly familiar with. Second, the same methodological procedure could be applied to the data of other languages, as well as to bigger data sets of different language backgrounds, different proficiency levels, different text types and different genres. These results could help in evaluating the generalizability of the method. Thus far the results seem promising, and I hope other researchers find them useful in their own research.

References

Aarts, J., and S. Granger. 1998. Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In *Learner English on computer*, ed. S. Granger, 132-141. London: Longman.

Barlow, M. and S. Kemmer (Eds.) 2000. *Usage based models of language*. Chicago: CSLI Publications.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1): 5-32.

Cheng, W., C. Greaves, and M. Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11 (4): 411-433.

Common European framework for languages: Learning, teaching, assessment 2006. Cambridge: Cambridge University Press.

Estival, D., T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th conference of the Pacific Association for Computational Linguistics (PACLING 2007), Melbourne, Australia*, 31-39.

Francis, G. 1993. A corpus-driven approach to grammar: Principles, methods and examples. In *Text and technology. In honour of John Sinclair*, eds. M. Baker, G. Francis, and E. Tognini-Bonelli, 137-156. Amsterdam: John Benjamins.

Goldberg, A. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In *Languages in contrast*, eds. K. Aijmer, B. Altenberg, and M. Johansson, 37-51. Lund: Lund University Press.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, analysis, and applications*, ed. A. P. Cowie, 145-160. Oxford: Clarendon Press.

Granger, S. 2013. Contrastive Interlanguage Analysis: A Reappraisal. Keynote speech. In *Learner Corpus Research Conference 2013*. Bergen/Os, Norway.

Granger, S., and M. Paquot. 2008. Disentangling the phraseological web. In *Phraseology: An interdisciplinary perspective*, eds. S. Granger and F. Meunier, 27-49. Amsterdam: John Benjamins.

Gries, S. Th. 2008. Phraseology and linguistic theory. In *Phraseology: An interdisciplinary perspective*, eds. S. Granger and F. Meunier, 3-25. Amsterdam: John Benjamins.

Guthrie, D., B. Allison, W. Liu, L. Guthrie, and Y. Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the fifth international conference on language resources and evaluation (LREC), Genoa, Italy*, 1222-1225. (http://www.lrec-conf.org/proceedings/lrec2006/)

Hothorn, T., P. Buehlmann, S. Dudoit, A. Molinaro, and M. Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7 (3): 355-373.

Hunston, S. 2001. Colligation, lexis pattern, and text. In *Patterns of Text: In honour of Michael Hoey*, eds. M. Scott, and G. Thompson, 13-34. Amsterdam: John Benjamins.

Inaba, N. 2007. Mikael Agricolan teokset tietokannan muodossa. In *Agricolan aika*, eds. K. Häkkinen, and T. Vaittinen, 147-161. Helsinki: BTJ.

Itkonen, E. 2005. *Analogy as structure and process*. Amsterdam: John Benjamins.

Ivaska, I. 2012. Key structure analysis of formally defined structures of learner Finnish. Paper presented at the conference *Learner Language, Learner Corpora*, University of Oulu, 2012.

Ivaska, I. 2014a. Edistyneen oppijansuomen avainrakenteita. Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin. *Virittäjä* 118:161-193.

Ivaska, I. 2014b. The corpus of advanced learner Finnish (LAS2). Database and toolkit to study academic learner Finnish. *Apples: Journal of Applied Language Studies* 8 (3): 21-38. (http://apples.jyu.fi/.)

Ivaska, I., and K. Siitonen 2009. Syntaktisesti koodattu oppijankielen korpus: mahdollisuuksia ja ongelmia. In *Korpusuuringute metodoloogia ja märgendamise probleemid*, eds. P. Eslon, and K. Õim, 54-71. Tallinn: Tallinna Ülikool.

Ivaska, I., and K. Siitonen 2011. Avainrakenneanalyysi. Tapa tutkia oppijankielen lauserakennetta korpusvetoisesti. *AFinLA-e* 3:35-47. (http://ojs.tsv.fi/index.php/afinla/issue/view/694)

Jantunen, J. H. 2004. *Synonymia ja käännössuomi. Korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin*. Joensuu: Joensuun yliopistopaino.

Jantunen, J. H. 2009. "Minulla on aivan paljon rahaa": Fraseologiset yksiköt suomen kielen opetuksessa. *Virittäjä* 113:356-381.

Jantunen, J. H. 2011. Avainsana-analyysi annotoidun oppijankieliaineiston tutkimisessa: Alustavia havaintoja. *AFinLA-e* 3:48-61. http://ojs.tsv.fi/index.php/afinla/issue/view/694.

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

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

Jarvis, S. 2011. Data mining with learner corpora: Choosing classifiers for L1 detection. In *A taste for corpora. In honour of Sylviane Granger*, eds. F. Meunier, S. De Cock, G. Gilquin, and M. Paquot, 131-158. Amsterdam: John Benjamins.

Jarvis, S. 2012. The detection-based approach: An overview. In *Approaching language transfer through text classification*, eds. S. Jarvis, and S. A. Crossley, 1-33. Bristol: Multilingual Matters.

Jarvis, S., and S. A. Crossley, eds. 2012. *Approaching language transfer through text classification*. Bristol: Multilingual Matters.

Jarvis, S., and A. Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. London: Routledge.

Jarvis, S., G. Castañeda-Jimenez, and R. Nielsen. 2012. Detecting L2 writers' L1s on the basis of their lexical styles. In *Approaching language transfer through text classification*, eds. S. Jarvis, and S. A. Crossley, 34-70. Bristol: Multilingual Matters

Koppel, M., J. Schler, and K. Zigdon. 2005. Automatically determining an anonymous author's native language. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining*, 624-628. Chicago: Association for Computing Machinery.

Lauseopin X-arkisto. n.d. School of Languages and Translation Studies of the University of Turku. Turku.(http://syntaxarchives.suo.utu.fi.)

Mayfield Tomokiyo, L., and R. Jones. 2001. You're not from 'round here, are you? Naive Bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics* (NAACL '01). Cambridge, MA: Association for Computational Linguistics.

Meunier, F. and S. Granger, eds. 2008. *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins.

Nesselhauf, N. 2004. *Collocations in a learner corpus*. Philadelphia: John Benjamins.

Odlin, T. 1989. *Language transfer: Cross-linguistic influence in language learning.* Cambridge: Cambridge University Press.

Penttilä, R. 2008. *Suomen* ja-*konjunktion vastineet* ir, o *ja* pilkku *liettuassa*. Master's thesis, University of Turku.

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

R Core Team 2013. *R: A language and environment for statistical computing*. Vienna, Austria. (http://www.R-project.org/.)

Räisänen, J. 2005. *Suomen tempusten semantiikka tšekin- ja venäjänkielisten suomenoppijoiden välikielissä*. Master's thesis, University of Turku.

Scott, M. 2010. Problems in investigating keyness, or clearing the undergrowth and marking out trails... In *Keyness in texts*, eds. M. Bondi and M. Scott, 43-57. Amsterdam: John Benjamins.

Scott, M. and C. Tribble. 2006. *Textual patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. 2001. Reviews of the Longman grammar of spoken and written English. I*nternational Journal of Corpus Linguistics* 6 (2): 339–359.

Skousen, R. 1989. *Analogical modeling of language*. Dodrecht: Kluwer Academic Publishers.

Stanford University. Natural Language Processing. n.d. Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources.(http://www-nlp.stanford.edu/links/statnlp.html.)

Stefanowitsch, A., and S. Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8 (2): 209-243.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8 (25). (http://www.biomedcentral.com/1471-2105/8/25.)

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* 9 (307). http://www.biomedcentral.com/1471-2105/9/307.

Strobl, C., J. Malley, and G. 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, S., and R. H. 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.

Université catholique de Louvain. Centre for English Corpus Linguistics. n.d. Learner corpora around the world. http://www.uclouvain.be/en-cecl-lcworld.html.

Wiersma, W., J. Nerbonne, and T. Lauttamus. 2011. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing* 26 (1): 107-124.

Wong, S.-M. J., and M. Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association*, 53-61. Cambridge: MA: Association for Computational Linguistics.