Laitinen, Mikko (forthcoming). Placing ELF among the varieties of English - Observations from typological profiling. In: Deshors, Sandra C. (ed.), Modeling World Englishes - Assessing the interplay of emancipation and globalization of ESL varieties. Amsterdam/ Philadelphia: John Benjamins, pp. 109-132.

Placing ELF among the varieties of English: Observations from typological profiling

Mikko Laitinen, University of Eastern Finland

Abstract

This study investigates how (dis)similar ELF is structurally from the core native varieties of English, indigenized L2 varieties and learner English. ELF is understood as second language use of English in settings where the interactants do not share a first language. The empirical part makes use of the method of typological profiling based on aggregate structural features. This method measures three indices (i.e. grammaticity, analyticity, and syntheticity), and it has previously been used to analyze a range of variety types but has not been applied to assess ELF. The results provide quantitative evidence that place ELF on the map and show that on purely structural grounds, ELF is a distinct variety type among English varieties. Moreover, the observations show that ELF is structurally different from second language acquisition, and there is a quantitative basis for drawing a distinction between ELF and traditional learner data.

1. Introduction

This article tackles the question of how different English as a lingua franca (ELF) is structurally from other varieties of English.[1] It makes use of the method of typological profiling based on aggregate structural features (Szmrecsanyi 2009). This method measures three indices and has previously been used to analyze various native Englishes, indigenized L2 varieties and learner English. In a seminal study, Szmrecsanyi & Kortmann (2011: 182) use the method to draw a distinction between English as foreign language (EFL) and English as a second language (ESL) on structural grounds. Up to today, the method has not been used to assess ELF, and my results provide quantitative evidence that place ELF on the map. ELF is here understood as second language use of English in settings where the interactants do not share a first language. As in Mauranen et al. (2015), my definition allows native speakers to be involved. Throughout the article, I will use the term 'ELF speaker' in a generic sense to include both speakers and writers.

[1] I want to thank the two anonymous reviewers for their valuable comments on an earlier version of this article. I have also benefitted greatly from the comments by Professor Anna Mauranen and from the discussions and comments at the Changing English conference in Helsinki in 2015, ISLE-4 in Poznan in 2016 and ICAME38 in Prague in 2017. The usual disclaimers apply.

ELF is a comparatively new object of study, and it is not included in the current models in the study of World English that are based on history and geo-political background (cf. Kachru 1985, Schneider 2007). Yet, the sociocultural and sociolinguistic importance of ELF among the English varieties today is substantial. In his article on the World System of Englishes, Mair (2013) classifies ELF as a super-central variety, i.e. a variety that is transnationally relevant, carrying demographic weight and sociocultural importance. Mair (2013) argues, that unlike some other super-central varieties, ELF is restricted to domain-specific uses such as academia and international business/law, but as will be shown in the material used here, this is a far too restricted view (cf. also Pietikäinen forthcoming on the uses of ELF in family setting). Sociolinguistically, ELF is not a focused variety, in the sense of Milroy (1987: 182–3). It has no native speakers nor do speakers share an idea of widely-held and recognizable set of norms in all levels of the language. Yet, recent empirical evidence suggests that spoken ELF is gradually emerging as norm-developing, rather than being simply norm-receiving (Low 2016).

This line of quantitative typological research is novel in ELF, where the research has predominantly been qualitative. Mauranen et al. (2015) point out that grammatical variability has been the least researched area, and ELF scholars have primarily only investigated individual grammatical features without in-depth quantitative information and systematic comparisons with other varieties of English. These include studies of the reorganization in the relative pronouns *who* and *which* (Cogo & Dewey 2012), and the regularization of the third person *-s* (Breiteneder 2009). Large-scale corpus studies are fewer. Ranta (2013) has focused on shared non-standard features in grammar, *viz*. vernacular universals, such as the inverted word order in indirect questions, the extended use of the progressive, the use of would in hypothetical *if*-clauses, and the use of singular agreement in existential there structures (Ranta 2013). In addition, Laitinen & Levin (2016) and Laitinen (2016, 2017) have looked into how ongoing grammatical change is adopted in ELF; These studies consist of investigations of the changes in non-aspectual uses of the progressive and investigations of a broader set of features, such as core and emergent modal auxiliaries (*can, should, have, need to*, and *be going to*, etc.) in ELF. Nevertheless, it remains fair to say that little is known about the typological status of ELF.

Drawing her evidence from the *English as a Lingua Franca in Academic Settings* corpus (ELFA), Mauranen (2012: 247) points out that spoken academic ELF is in many ways similar to native speech and "the overwhelming majority of lexis, phraseology, and structures are

indistinguishable from those found in a comparable corpus of educated ENL [English as a Native Language], including their frequency distributions". Her observations are based on n-grams and word lists. With regard to other non-native varieties, she argues that ELF is essentially dissimilar from learner English.

The research question is how different ELF is structurally from the core native varieties, indigenized L2 varieties and learner English. Since all ELF use has involved language learning at one point, my null hypothesis is that ELF is similar to learner data. The results answer to two questions: First, where on the unidimensional grammaticity index lingua franca evidence falls, and second, where on the two-dimensional analyticity–syntheticity plane can I place ELF? Both of these terms will be explained below.

Section 2 details the theoretical and the methodological basis of the profiling method adopted here. Section 3 discusses the ELF material analyzed, and the results are presented in Section 4. Lastly, Section 5 discusses the implication of the results for model building in World Englishes.

2. The method of typological profiling and ELF

The study employs the method presented in Szmrecsanyi (2009), who proposes that typological notions of analyticity and syntheticity by Greenberg (1960) could be amended by large-scale quantitative corpus data. This method makes use of three indices. Firstly, the synthetic index is based on the frequency of select bound grammatical markers, and the numeric value in this index is the number of lexical items that carry at least one bound marker. A prototypical case would be verbal third person *-s*, which marks two meaning, *viz*. nonpast and third-person singular. Secondly, the analytic index is calculated on the basis of a range of function words, which are "defined as being members of closed word classes" (2009: 320). Thirdly, the grammaticity index is the sum of synthetic and analytic markers per sample.

The indices are presented in detail in Szmrecsanyi (2009: 326–7). The analytic markers consist of the following:

(1) conjunctions, subjunctions, and prepositions
(2) determiners, articles and *wh*-words
(3) existential *there*
(4) pronouns

(5) analytic comparative and superlative markers
(6) *to*-infinitive marker
(7) modal auxiliaries
(8) negator *not*, or *n't*
(9) auxiliary *be*
(10) auxiliary *do*
(11) auxiliary *have*

The synthetic markers are:

(12) *s*-genitive
(13) synthetic comparative and superlative adjectives
(14) plural nouns
(15) plural reflexive pronouns
(16) inflected verbs

Some of the categories, *viz*. the analytic categories 9–11 also load the synthetic side. Similarly, the synthetic ones 15–16 load the analytic indices. This method results in frequencies that measure variability, but is not variationist as such, in which all the markers would be used to express one meaning with two forms. Szmrecsanyi (2009) points out that for some analytic markers there is a clear synthetic alternative (i.e. the forms of adjectival comparison, or the analytic and synthetic genitives). For some, such as the negator *not*, or the plural noun marking, this is not the case.

It should be noted that the method should also be viewed critically, especially since the grammatical components are not weighed relative to their basic frequencies. The component categories vary in token frequency, and the relative weight of the high-frequency elements is substantial. Therefore, tiny alternations in prepositions, determiners and pronouns (all on the analytic side) or plural nouns (synthetic) will lead to considerable alternations in the normalized frequencies. One way to improve the precision of the method could be to assign a relative weight to all the components, but since the aim is to compare ELF with previous results, this is left for future studies.

Since the method integrates various structural features, it offers a useful way of quantifying a variety that does not prioritize a single grammatical structure, thus limiting the bias inherent in single-feature studies. For instance, it has offered a more fine-grained picture of the interplay between synthetic and analytic tendencies in the history of English. As Szmrecsanyi (2009) points out, Standard English is often seen to be an analytic language *par excellence*, but the quantitative results obtained through the method have contested this monolithic myth, and his results show that both major varieties, *viz*. American English (AmE) and British English (BrE), have become more synthetic and less analytic during the second half of the 20th century. The same holds for earlier periods; Szmrecsanyi (2012) applies the frequency-based indices for post-Old English corpus-data. His results show that analyticity has been on the decline since the Early Modern English period, and syntheticity has increased.

The method has also offered a way of quantifying structural differences between varieties. Firstly, when it comes to grammaticity, varieties differ substantially with regard to how overtly redundant they are. Traditional L1 vernaculars (low-contact varieties) exhibit more grammatical marking than high-contact L1 vernaculars (e.g., AmE, New Zealand English, etc.), which in turn exhibit more grammaticity than indigenized L2 varieties (Singapore English, etc.). Secondly, in terms of syntheticity, low-contact varieties show higher frequencies of syntheticity than high-contact varieties, and L1 varieties in general exhibit more syntheticity than L2 varieties. Thirdly, among the L2 varieties, Southeast Asian Englishes (e.g. Singapore, Philippine, and Hong Kong) are less analytic and less synthetic than those outside (e.g. IndianE, JamaicanE, East AfricanE). Most importantly for my purposes, learner English data exhibit "less syntheticity and more analyticity than Standard British English" (Szmrecsanyi & Kortmann 2011: 182).

Many of the features covered in the indices are directly relevant to aspects of ELF and could therefore offer empirical insights of how similar or different ELF is when compared to the other English varieties. A case in point is one of the hallmark characteristics of ELF, i.e. negotiating meanings through online processing. According to Mauranen (2012: 244), one characteristic of spoken ELF is enhanced transparency through structural simplification. Since the typological indices are closely connected with language complexity, the method offers a way of quantifying such transparency in ELF. On the one hand, the more analytic a language is, the more it tends to contribute to transparency and explicit nature of communication. On the other hand, increasing syntheticity tends to create a more economical output, and grammaticity

contributes to explicit redundancy, meaning that the more grammatical markers there are, the less needs to be inferred from the contextual cues (Szmrecsanyi (2009).

While the article applies the method that is readily available, it still involves a considerable theoretical component. It deals with ensuring empirical validity and enlarging the scope of ELF corpora, as the existing datasets only cover a small set of genres. We need new corpora that offer a multi-genre view to ELF. These new corpora should ideally be such that they enable comparisons with other (native and non-native) corpora.

3. Material and methods

My material comprises of two sets of corpora, which for the first time make it possible to access a broad range of ELF genres. The term genre is understood as a concept that points to functions of communication, i.e. situation, audience, and the purpose (see Biber & Conrad 2009). The first set of data consists of the well-known first-generation ELF corpora, *viz*. the spoken *Vienna-Oxford International Corpus of English*, VOICE (Seildhofer 2011), and the newly-released *Written English as a Lingua Franca*, WrELFA, corpus (Mauranen et al. 2015).

VOICE is a 1-milion-word corpus of unscripted face-to-face spoken interaction from organizational settings. The informants come from a mixture of L1 backgrounds, and since the individual L1 collections result in small samples, it is used in its entirety. VOICE represents spoken communication in which the informants' objective is to inform and to maintain interpersonal relations.

WrELFA is a circa 1.5-million-word corpus of academic writing divided into three text types in the academic genre. The unedited research papers cover half of the material, the so-called SciELF corpus. The PhD examiner's report genre contains some 400,000 words, and the research blog component some 372,000 words. The collection process targeted the academic user of ELF, and, according to the compilers, the texts have not undergone professional proofreading or checking by an English native speaker (see http://www.helsinki.fi/englanti/elfa/wrelfa.html). It represents second-language use in written scientific communication, and 35 L1s are represented in it. Additionally, an undetermined number of blog commenters are included in the blog component, and according to the corpus compilers, their identities cannot be verified. Similarly to the other ELF corpora, native speakers of English are occasionally included in the blog and in

PhD examiners' subcorpora. Since the results in the following section are the first ELF results obtained using this method, I will only use the PhD examiner's statements of WrELFA.

To complement these first-generation corpora and to make up for the fact that "genuine ELF written text databases are still missing" (Mauranen et al. 2015: 402), the author and his associates are compiling second-generation written ELF corpora. They offer a larger sample for a smaller set of L1 backgrounds than the first-generation corpora and broaden the stock of ELF genres available. They concentrate on second language use of English is specific geographic settings.

Our pilot work focuses on two Nordic countries, Sweden and Finland, where the role of English has undergone considerable changes in the recent decades. The two countries are not undergoing a language shift, but the sociolinguistic situation is that of urban multilingualism in which English is used as an additional resource alongside the main languages primarily, but not exclusively, by younger generations who live in urban areas and work in white and pink collar professions (see Laitinen 2016).

The working titles for the corpora are SWE-CE, the *Corpus of English texts in Sweden*, and FIN-CE, the *Corpus of English texts in Finland*. They are systematically-collected and large enough sources of baseline data that fulfil the requirement for empirical validity. They contain texts from written mode of communication, and together with the already-existing spoken ELF corpora they make it possible to investigate a range of genres. The texts have been produced by non-native writers, who use English as a second language resource. The majority are taken from non-learner settings. The only exception is fiction, which we are collecting in collaboration with teachers organizing creative writing courses. The rationale is that (fan) fiction is an important arena of ELF writing (Leppänen 2012), but unfortunately such texts do not fulfil our need to identify the authors, and we have to collect material from educational settings.

We know to what extent the materials have been subjected to normative language checking by professional editors, translators and native speakers. Preference is given to texts that are not edited, but it is assumed that the more informationally oriented a text is, the more likely it is to have undergone some degree of language checking and collaborative effort. Furthermore, the informants' use of spell-checkers and other tools which nowadays are available in most web-browsers and mobile devices cannot be ruled out. To what extent such tools have an influence on our data is beyond our control, but it is clear that such tools are part of contemporary writing

practice and are equally used by native writers. Published materials edited by native speakers are excluded.

These second-generation ELF corpora cover a range of genres. We draw from Biber's (1988: 104–108) multidimensional analysis of textual variation, and more specifically from dimension 1, i.e. information density and exact content vs. interactional and generalized content, to place texts to the genre matrix. This dimension is used as a heuristic tool and is not yet empirically validated. Figure 1 visualizes the textual division covered in the study.

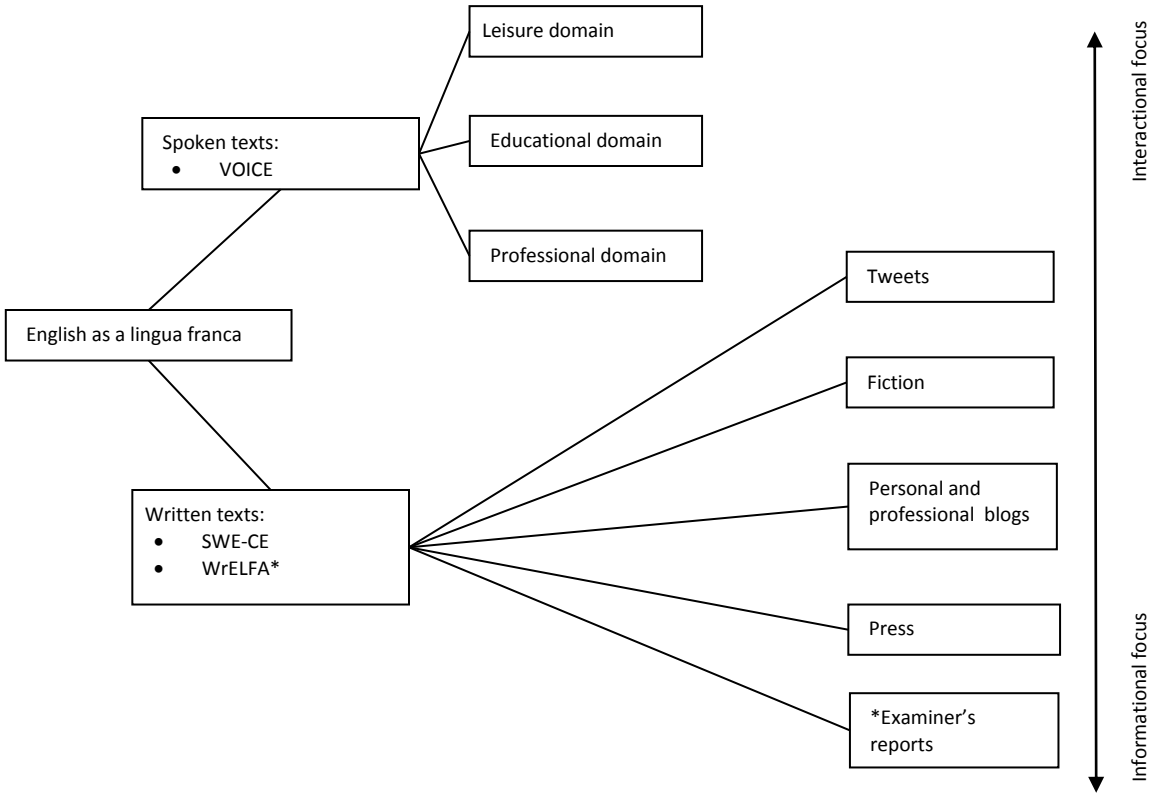

Figure 1. The genre distribution of the ELF corpora used in the study

When the written and spoken ELF materials are combined, the resulting corpora are nearly 2.3 million words in size. The results are based on 1,023,082 words of spoken VOICE. In the written side, I will use the Swedish component as the material. The written corpus consists of 332,290 words of tweets (short micro-blog messages). This material has been collected from 50 randomly-selected individuals at www.curatorsofsweden.com. The site collects tweets sent by people who are citizens of Sweden and who therefore manage the 'official' Twitter account

Sweden for one week at a time (for more on our Twitter data collection see Laitinen et al. forthcoming). The ELF fiction subcorpus is 193,755 words. Professional and personal blogs consist of 263,486 words, and they are considered as one component (note that we will divide them into professional and personal blogs, see Grieve et al. 2010 on the classification of blogs using linguistic criteria). The news subcorpus comprises of 196,232 words, and the examiner's statement part from which the native English writers have been excluded (276,712 words in 236 statements).

To ensure comparability with the previous observations, my study applies the method presented in Section 2 without major modifications. A minor modification is that the results for the VM category, (7) in the list above, also include the contracted forms *gonna*, *gotta*, *hafta*, and *wanna*, but their frequencies are low. The material that was untagged (WrELFA and SWE-CE) was parts-of-speech tagged using the CLAWS7 tagset. Tagging was also tested for blogs and tweets, but the error rate turned out to be high for these genres so it was determined best to run the material through a PERL script that attached 1,000 randomly selected items in the subcorpora with a tag. The script was kindly provided by Benedikt Szmrecsanyi. These items were manually analyzed for their POS. The results for the fiction, blogs, and press subcorpora are the frequencies generated through automatic POS-tagging. The VOICE results were obtained using the POS-tagged version (http://www.voice.univie.ac.at). Some of the results that required broader contextualization were checked using the XML-version of the corpus (https://www.univie.ac.at/voice/help). The tagging used in the VOICE corpus is based on a modified set of Hepple tags (Seidlhofer et al. 2014), and a scheme was created to convert the results comparable to the results using CLAWS7 (as illustrated in Table 1 below).

Table 1. The analytic (A) and synthetic (S) component categories as defined through the POS tags

Feature	**CLAWS7**	**VOICE tagset or search function**
A1: Conjunctions, subjunctions and prepositions	CC*, CS*, I*	CC, IN
A2: Articles, determiners and WH-words	APPGE, AT*, D*, RGQ*, RRQ*	DT, PDT, PRE, WRB, WDT, WP
A3: Existential THERE	EX	EX
A4: Pronouns	P*	PP*, indefinite, reflexive and reciprocal pronouns
A5: MORE/MOST	RGR, RGT	RBR, RBS

A6: Infinitive marker TO	TO	TO
A7: modals	VM*, gonna, gotta, hafta, wanna	MD, gonna, gotta, hafta, wanna
A8: negator NOT/N'T	XX	NOT, N'T
A9: auxiliary BE	VBD*|VBG|VBM|VBN|VBR|VBZ* + (*)? + V*	lemma: BE + (XX0)? |(*)? + (*)? + V*
A10: auxiliary DO	VD* + (*)? + V*, VD* + XX	lemma: DO + (XX0)? |(*)? + (*)? + V*
A11: auxiliary HAVE	VH*+ (*)? + V*, VD* + XX	lemma: HAVE + (XX0)? |(*)? + (*)? + V*

S12: Germanic genitive marker ('|'s)	GE, MCGE	POS
S13: Comparative and superlative adjectives	JJR, JJT	JJR, JJS
S14a: plural nouns	NN2, NNL2, NNO2, NNT2, NNU2, NP2, NPD2, NPM2	NNS
S15: plural reflexive pronouns	PPX2	*SELVES
S16: Inflected verbs	VBDR, VBDZ, VBG, VBM, VBN, VBR, VBZ, VDD, VDG, VDN, VDZ, VHD, VHG, VHN, VHZ, VVD, VVG, VVGK, VVN, VVNK, VVZ	VVD, VBD, VHD, VVG, VBG, VHG, VVN, VBN, VHN, VVZ, VBZ, VHZ, VHS, DOS, VBS, VBP

The numeric results for the indices have been provided by two trained research assistants, and their initial searches have been checked once. As in Szmrecsanyi (2009), the indices are ratios of the number of markers normalized per 1,000 words.

4. Results

4.1. Grammaticity

Table 2 illustrates the total number of both grammatical markers, i.e. grammaticity. It is the most robust category, showing how transparent a variety is. According to Szmrecsanyi (2009), the higher the score, the more efficient the output is in terms of pragmatic functions, since the relationship between overt marking and negotiating meanings is indirect. Similarly, the lower the index, the more needs to be negotiated using pragmatic means. To acquire understanding of where ELF is situated among the varieties of English, the ELF results (in bold) are compared with the figures drawn from Szmrecsanyi (2009: 329). To make the samples more comparable, the ELF observations at this stage exclude tweets. They will be included in the subsequent tables and figures, but as they constitute a highly distinct genre, they are presented separately (for more on characteristics of tweets and other e-genres, see Knight, Adolphs & Carter 2014).

Table 2. Grammaticity index (GI) of ELF compared with the other varieties (from Szmrecsanyi 2009)

Language variety/form	GI	z score
Hong Kong E	539	-1.93
Singapore E	549	-1.70
ELF	**574**	**-1.13**
Phillippine E	592	-0.72
Irish E	598	-0.58
New Zealand E	607	-0.38
St. AmE	607	-0.38
Somerset (southwest)	626	0.05
Jamaican E	627	0.07
Indian E	632	0.19
St. BrE	643	0.44
East African E	647	0.53
Kent (Southeast)	657	0.76
Lancashire (North)	667	0.98
Glamorgan (Wales)	669	1.03
Shropshire (Midlands)	680	1.28
Sutherland (Highlands)	689	1.49

The GI scores are the arithmetic means, and they indicate that ELF is roughly one standard deviation below the mean value of this index. It falls in between two outer circle varieties, just lower than Philippine English and just higher than the Southeast Asian L2 English varieties, Hong Kong and Singapore English. These last two in particular are highlighted by Szmrecsanyi (2009) as contact-induced varieties in which adult language learning plays a significant role.

The results in Table 2 are important on at least two levels. For the first time, we are not confined to a limited set of genres in ELF setting but can rely on evidence from various discourse situations on the spoken–written continuum. In addition, as opposed to much of the previous ELF evidence, Table 2 makes use of evidence based on aggregated linguistic structures rather than single grammatical, lexical, or phraseological features.

The results in Table 2 are quantitative evidence of structural simplification observed in previous ELF studies. They support some of the previous findings in the ELF literature based on spoken data (see Mauranen 2012: 244), namely that one characteristic of spoken ELF is negotiating meanings in interaction, which leads to enhanced transparency and structural

simplification. Table 2 offers a quantitative view of what this transparency means in corpus data. The results also add evidence that ELF speakers often avoid overt grammatical marking. According to previous studies by Breiteneder (2009) and Jenkins, Cogo & Dewey (2011: 289–290), one characteristic of ELF interaction is the omission of grammatical markers, such as third person *-s* or articles, both of which are included in the indices here.

While section 4.3 focuses on genre differences in the ELF corpora in more detail, I will next focus on how the spoken and written ELF modes differ from each other in terms of grammaticity. Specific attention is also paid to how one genre behaves relative to the spoken–written continuum. This genre, tweets, is written in form, but it tends to exhibit spoken characteristics. Since the standard corpora used in Szmrecsanyi (2009) do not contain material from this genre, it is kept separate. The results also include the arithmetic mean values of four main variety types of English, drawn from Szmrecsanyi (2009: 329–30). They serve for reference purpose to show how substantial the differences between the spoken and written ELF subcorpora are.

Table 3 illustrates that the tweet subcorpus has the lowest grammaticity index (GI: 536), and it is clearly a specific written genre in which more emphasis needs to be put on contextual cues and pragmatic inference than in spoken communication. VOICE corpus has the GI of 553. This result illustrates the emergent nature of spoken ELF, in which meanings are negotiated through enhanced explicitness (Mauranen 2012: 245). In the written side however, the result is markedly different, and the grammaticity score is substantially higher (the mean is 597), making it more similar with the outer circle L2 varieties than spoken ELF.

Table 3. Grammaticity indices of ELF compared with the data from Szmrecsanyi (2009)

Language variety/form	**GI**	**z score**
ELF tweets	**536**	**-1.26113**
Spoken ELF (VOICE corpus)	**553**	**-0.83599**
Southeast Asian Englishes (Singapore, Philippines, Hong Kong)	560	-0.66093
Written ELF (WrELFA and SWE-CE)	**597**	**0.264372**
Other L2 (outer circle) varieties	598	0.28938
Transplanted L1	607	0.514454
Low-contact L1 dialects	654	1.689839

The quantitative patterns observed are clear. With regard to the unidimensional grammaticity index, ELF falls in between the two L2 variety types of Southeast Asian Englishes and other outer circle varieties. It is clearly not on par with the transplanted L1 varieties and well below the average of the traditional low-contact L1 dialects.

The illustrations below show what these quantitative differences mean in texts. Note that for visualization purposes, only two of the analytic markers (determiners and modals) and synthetic ones (plural nouns and inflected verbs) have been included in the illustrations here. It goes without saying that any automatically-generated contextual information in tweets (i.e. the time of sending a tweet) and the material not keyed in by an individual author (URL-links, re-tweet mark-up, etc.) are separated by our text-level coding scheme used in the this subcorpus. They are not included in the results.

(1) <TIME>May 31, 2015, 1:18 p.m.</TIME> <AT>@47thANNA</AT> Haha.
<TIME>May 31, 2015, 1:18 p.m.</TIME> <AT>@MarissaTree</AT> <AT>@niannelynn</AT> When **will** the wedding be?
<TIME>May 31, 2015, 1:17 p.m.</TIME> <AT>@HarietaNoPotter</AT> Haha, sorry! True detective!
<TIME>May 31, 2015, 1:17 p.m.</TIME> <AT>@va_ellen</AT> Strangely enough I **haven't been** there.
<TIME>May 31, 2015, 1:15 p.m.</TIME> <AT>@Kyroenna</AT> **That's** my guess also ...
<TIME>May 31, 2015, 1:13 p.m.</TIME> The Vegetable Man **goes** to the beach wearing a zukini.
<TIME>May 31, 2015, 1:09 p.m.</TIME> What **could** be the favorite food and drink of <Q>True blood</Q> writer Nic Pizzolatto, I wonder?
<TIME>May 31, 2015, 1:05 p.m.</TIME> <AT>@niannelynn</AT> You **are** absolutely right.
<TIME>May 31, 2015, 12:59 p.m.</TIME> <AT>@niannelynn</AT> But we hardly know each other?!
<TIME>May 31, 2015, 12:56 p.m.</TIME> <AT>@dmacuk</AT> Well said.
<TIME>May 31, 2015, 12:56 p.m.</TIME> <AT>@niannelynn</AT> Do you propose?
<TIME>May 31, 2015, 12:50 p.m.</TIME> <AT>@dmacuk</AT> But the vote said no?
(SWE-CE, tweets, May 2015) (8 markers = c. 14%)

(2) yes to force to force to integr- to force **the** integration and that**'s the** that**'s the** main point of difference because i **went** abroad i **got my** education and i **used** it at home for **my** duke for **my** bishop fo- for **my** ho- hometown and today i think we get educa- er we get education and we **don't** know where we **gonna** use it and that**'s the** that**'s the** big difference (VOICE, EDsed251) (18 markers = c. 22%)

(3) As **mentioned** earlier, when **an** employee compile**s** **a** quote in **the** office, **the** customer **might** feel that he **is** not part of **the** process and worry that **the** price **is** manipulated by **the** company. Example**s** of this **can** be **found** in forum thread**s**, such as at byggahus.se [3], **discussing** **the** subject. Delayed price quote**s** and customer**s** **feeling** cheated **is** **a** problem for **the** company. This thesis **will** look at how **the** manual process **can** be **sped** up and **made** more transparent (SWE-CE, theses, 2014) (29 markers = c. 35%)

After the most robust category, i.e. the grammaticity index, the next section will look into analyticity and syntheticity in more detail and places ELF among the various world English varieties.

4.2. ELF on a two-dimensional plane

The backdrop to this section is the observation that the variety types differ substantially on a two-dimensional analyticity-by-syntheticity plane. Space permits me to illustrate some of the previous findings only briefly, but they are explicitly explained in the sources used in this section. Even with the risk of simplifying things, it is fair to say that the findings can be summarized as follows. On the one hand, Szmrecsanyi (2009) observes that the traditional regional dialects from the British Isles are more synthetic than the varieties labeled as high-contact varieties. This latter group forms a heterogeneous set of varieties. They exhibit considerable spread in which indigenized L2 varieties (East African English, Indian English, Jamaican English, Hong Kong English, Singapore English and Philippine English) form a clearly distinct group. This group is different not only from standard BrE and AmE, but also from language-shift Englishes (Irish and Welsh English) and transplanted L1 Englishes (i.e. New Zealand English and spoken AmE). The indigenized L2 varieties can be further divided into Southeast Asian L2 varieties, which are substantially less analytic and synthetic than "non-Southeast Asian L2 varieties" (Szmrecsanyi 2009: 328). Standard AmE is slightly less synthetic than BrE. On the other hand, Szmrecsanyi & Kortmann (2011: 182) observe that traditional learner essay data in the *International Learner Corpus of English* (ICLE) is less synthetic but clearly more analytic than Standard BrE. This observation forms the basis for them to draw a distinction between learner language and second language varieties on structural grounds.

Figure 2 visualizes the two-dimensional analyticity–syntheticity plane, setting the written and spoken ELF results side-by-side with some of the results presented in Szmrecsanyi (2009) and Szmrecsanyi & Kortmann (2011). Note that the learner English data consist of all the ICLE results combined and the same holds for the indigenized L2 Englishes. The spoken British

English data are from of the spoken genres in Szmrecsanyi (2009: 333) and are used as a point of comparison for VOICE. In the ELF side, the spoken data are from VOICE in entirety. The written ELF results exclude tweets and are based on 930,185 words in WrELFA and SWE-CE.

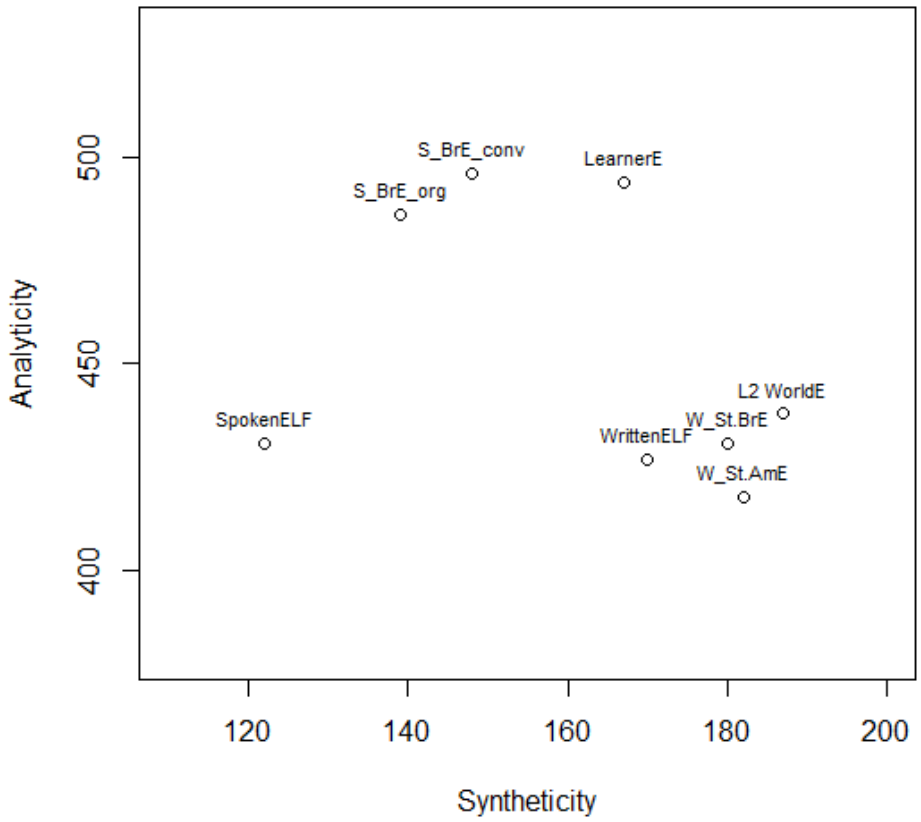

Figure 2. Written and spoken ELF compared with select variety types in Szmrecsanyi (2009) and Szmrecsanyi & Kortmann (2011)

The results visualize how spoken and written ELF could be positioned relative to a select set of English varieties. For spoken ELF, the synthetic value is 122, and the analytic one 431, and the respective values for written ELF are 170 and 427. The differences of the analytic values are not statistically significant (log-likelihood (LL) value 1.58, $p>0.05$), but they are highly significant for the synthetic values (LL 13.16, $p.<0.001$). This finding is slightly altered from our first preliminary observations (Laitinen, Levin & Lakaw in press), in which our written sample

only consisted of formal academic and news genres. However, they do not change the main observation indicating that considerable differences exist between spoken and written ELF use. Figure 2 illustrates how the differences, which were visible already in grammaticity (Table 3), are brought about by a smaller share of synthetic markers in spoken data. The Pearson residuals vary between -1.554 and 1.496, but the effect size in Cramer's phi for these nominal variables is minimal (0.073).[2]

The result regarding the differences between spoken and written modes in ELF is similar to the one observed in other varieties in Szmrecsanyi (2009). His result shows that all the major varieties of English exhibit similar decreases in analyticity and increases in syntheticity between spoken and written modes of production. The ELF evidence is not random but conforms to the general pattern relative to the mode of production. However, it needs to be pointed out that VOICE exhibits lower index values for both analyticity and syntheticity relative to spoken British English, so it clearly highlights the emergent characteristics of ELF in which meanings are negotiated in interaction.

More importantly for the ELF debate, the differences indicate increased transparency and output economy only in spoken ELF, but not necessarily in the written side. No such tendency is discernible in the written data, and more research needs to be carried out on the structural properties of written ELF.

Another important feature in Figure 2 is that the observations indicate substantial differences between traditional learner data and ELF. They confirm that, on purely structural grounds, language acquisition in foreign language settings should be viewed differently from second language use (cf. Mauranen 2011 on the notion that acquisition and use are connected, but dissimilar). The two forms of non-native English are different. Written learner data exhibit close similarities with spoken native varieties, and Granger & Rayson (1998) suggest that such tendencies are discernible in register interferences and the over-representation of speech-like features in learner data. However, the results here show that similar tendencies cannot be detected

[2] Pearson residuals are utilized to check whether the observed values in two-dimensional data are larger or smaller than the expected frequencies (cf. Levshina 2015: 120). This method makes it possible to observe the effect of the dependent variable. The values that are smaller than -3.841 or greater than 3.841 are considered to be particularly noteworthy, and their effect is more pronounced than those that fall in between. Cramer's phi is a post-test used to determine strengths of association between two variables and is a measure of association ranging between 0 and 1.

in my written ELF data. The latter are more synthetic and substantially less analytic than learner English data. The total figures for written ELF are 423 analytic markers and 170 synthetic ones, and for learner English, they are 494 and 167 (according to Szmrecsanyi & Kortmann 2011). The result is statistically significant for the analytic markers (LL 5.50, $p<0.05$), but not for the synthetic ones (0.03, $p>0.05$).

The total frequencies of written ELF in Figure 2 suggest that it crops up with the broad group of standard BrE and AmE (data from the Freiburg versions of the Brown corpora in Szmrecsanyi 2009) and the indigenized L2 varieties. On the whole, comparisons of the syntheticity indices between written ELF and written BrE (LL 0.06, $p>0.05$) and AmE (LL 0.02, $p>0.05$) show no statistical significant differences. As for the analytic indices, the same holds true, there are no statistically significant differences in the data.

So far, I have considered ELF only as spoken and written modes of communication and have compared these two with the other varieties. In the next section, I will focus on ELF, and explore to what extent the various written genres are structurally different from each other and from the spoken evidence in VOICE. Some comparative evidence from BrE is included.

4.3. Genre differences in ELF

The results in the previous sections establish that ELF (i.e. second language use) is structurally different from EFL (i.e. second language acquisition) and similar to other L2 uses of English in terms of both grammaticity and analyticity by syntheticity. These results, based on a large set of aggregate data, not only confirm a similar assumption in the ELF literature (Mauranen 2011), but they also illustrate correspondences between written ELF and the major standard varieties of English. There exist differences between ELF and Standard English, but these are more pronounced in the spoken side than in the written, as illustrated by the results in Table 3 and Figure 2. These results are important considering the status of ELF in general, since they show that it is a structurally distinct variety type. As pointed out in the introduction, I am only referring to its structural properties here, as we should be careful in assessing the sociolinguistic angle of ELF being a focused variety.

A key question in this section is the systematicity of ELF genres. If the genre differences are systematic, so that both spoken data and the spoken-like written genres and the various written genres (see Figure 1 above) exhibit similar tendencies as in the native varieties, the results

should indicate that ELF speakers show at least some degree of stability required from a focused variety and exhibit awareness of genre characteristics in terms of structural features.

These previous findings form the backdrop to the quantitative observations here. Figure 3 shows how the spoken VOICE and the five written ELF genres locate on a three-dimensional plane that integrates analyticity (y-axis), syntheticity (x-axis) and grammaticity (z-axis).

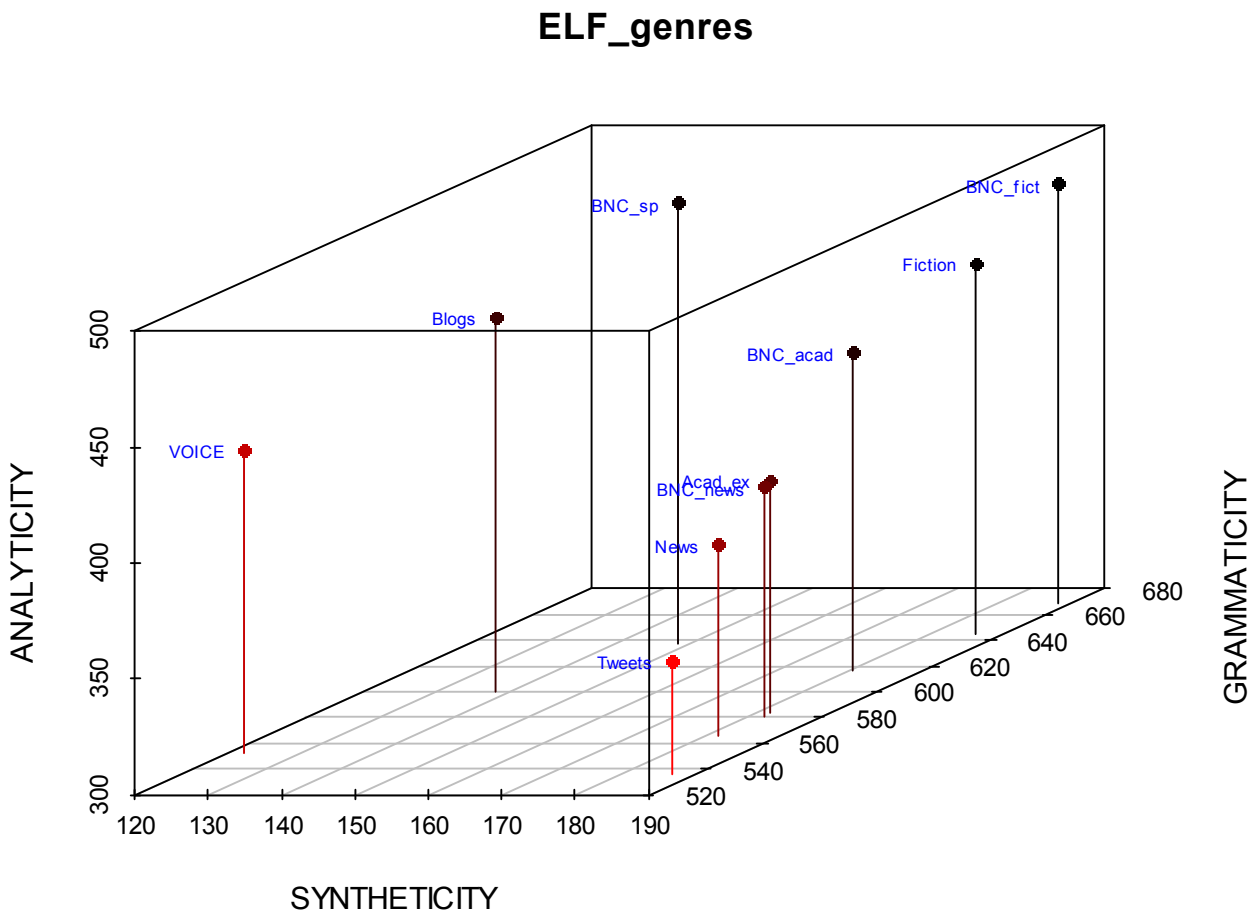

Figure 3. Spoken ELF data and written ELF genres in a three-dimensional genre space

I have include three written genres from Szmrecsanyi (2009), marked with "BNC_genre" in an abbreviated form (i.e. academic journal articles, news, and fiction). In addition, it includes the spoken British English (BNC_sp) results, which are the arithmetic mean figures of 16 spoken genres (2009: 333).

The results show first of all that spoken VOICE and the most interactive written ELF genre, tweets, are highly dissimilar. Recall from Table 3 above that both of them exhibit low grammaticity values, and Figure 3 illustrates that spoken material is characterized by higher frequencies of analytic markers, whereas tweets have more synthetic markers. This is also confirmed by the Pearson residuals, which indicate that there the observed frequency of the synthetic markers is higher than the expected (2.831). The Pearson residual value for the analytic markers in VOICE is (1.754). In all, Cramer's phi for nominal variables indicates that the effect size is small (0.142). These observations suggest that spoken ELF with its slightly increased analyticity highlights transparency and negotiating meaning through explicit analytic marking, but the same tendency is not true for interactive tweets. In tweets, economy and compressing information to 140 characters weighs more, but the correlation coefficient remains mild. The tweet component shows characteristics of more formal genres, such as news. This finding similar to the one observed in native tweets in Knight et al. (2014), whose evidence comes from the relative frequencies of broad syntactic categories.

The blog data visualized in Figure 3 stand out not only from tweets but also from the other written ELF genres. They are closest to spoken data and exhibit lower frequencies in syntheticity than the other written datasets. Pearson residuals are highest for these synthetic markers, but they are not outside the critical values 3.841 and -3.841. Similarly, Cramer's phi shows no significant effect (-0.052). The finding goes against the observations in Knight et al. (2014), whose results in the native side show that blogs show characteristics of formal genres.

When comparing the two spoken corpora against each other, the differences in the data are small. There is a slightly higher share of analytic markers in the BNC than in VOICE, and the same holds for the synthetic markers. However, Pearson residuals of this variable (analytic vs. synthetic in two datasets) show that the differences between the observed and the expected frequencies show no noteworthy differences. Cramer's phi, measuring the effect size, falls close to zero (0.014).

The analytic and synthetic indices for the ELF news genre are 383 (AI) and 182 (SI), and the differences are not statistically significant when we compare them to the BNC_news frequencies (the Pearson residuals are vary between 0.189 and -0.187, Cramer's phi shows no effect (0.009). In this genre, ELF and standard native English are therefore very close to each other, which is not surprising. According to Hundt & Mair (1999: 236), news as a genre tends to

be agile, and authors (journalists) in native settings tend to be "receptive to" change and innovation. It would be unusual to assume that they would not be in ELF. In addition, it is likely that as language professionals, native and non-native journalists alike, must be aware of their language production. The only marginal difference between ELF and native Standard English is the higher analytic score in the latter, but the correlation coefficient remains low.

In the academic genre, the datasets represent highly formal and informative academic writing, but the difference is that ELF evidence comes from examiners' statements, whereas the BNC material is taken from academic articles. Despite these differences, the quantitative patterns observed in the material are highly similar, with no statistically significant differences (the Pearson residuals vary between 0.462 and -0.449, and Cramer's phi shows no effect of correlation 0.022). The only difference is the slightly higher analytic index in the native dataset when compared with ELF, but not at any statistically significant levels.

Finally, the same tendency of remarkable genre similarities continues in fiction. In the ELF side, the analyticity index is 460 and that of syntheticity 186, and in the BNC data, the corresponding values are 481 and 188. The differences are not statistically significant (the Pearson residuals vary between 0.167 and -0.164, and Cramer's phi shows no effect of correlation 0.007), and ELF is highly similar to native data. The finding is noteworthy since our fiction component is closest to learner writing. The texts were collected from creative writing courses in Sweden, but they exhibit only little of the traditional learner language characteristics in the indices. One reason for this close similarity could be tied to our informants: people learning to write fiction in their L2 could be aware of their language production capabilities, which results in these close similarities.

5. Discussion and implications

In this last section I will first provide a brief overview of my observations. After that, I will explore some of the implications of these observations and aim at connecting them to the theoretical models of World Englishes in general and to the issues related to the study of ELF in particular.

The results presented are the first ELF results obtained using the typological profiling method, and they enable assessing the variety status of ELF. They show that on purely structural grounds, ELF is another distinct variety type among the English varieties. The quantitative

patterns observable in the data are clear: Second language use is structurally different from second language acquisition, and there is a quantitative basis for drawing a distinction between ELF and traditional learner data (EFL) using purely structural criteria as is done here.

The results in Section 4.1 show that with regard to grammaticity, ELF is similar to the many indigenized L2 Englishes. There exist substantial differences between spoken and written modes, and new genres, like tweets, which are characteristic of the globalization of English, behave quite distinctly from the more traditional genres. Section 4.2 shows that when it comes to ELF and native evidence, spoken ELF is structurally different from spoken native data. This observation can be compared to Mauranen et al.'s (2015: 402) observations. She points out that "even a short fragment of ELF talk heard or seen in transcription is usually enough to tell it is not ENL" [i.e. English as a native language]. My results offer quantitative confirmation to this. However, they also contrast some of the previous findings in ELF and add another angle to them. Mauranen et al. (2015) continue that when it comes to "word lists of individual word and n-grams" there is "notable overall similarity" between ELF and native uses in academic settings. My results show that spoken ELF is lower on both syntheticity and analyticity when compared with native spoken data. It is important to note that Section 4.3 shows that no such differences can be discerned in the written side. When we use analyticity and syntheticity as an index of structural similarities, written ELF is not distinguishable from native data.

The empirical results enable refuting the null hypothesis, and they therefore force us to rethink the traditional tripartite division of Englishes. For instance, studies that have explored closing the paradigm gap between learner varieties and post-colonial indigenized L2 uses have suggested that the EFL–ESL should be seen as a continuum (cf. Mukherjee & Hundt, eds. 2011). However, the results here show that such a discussion excludes a crucial component, *viz*. a structurally distinct variety type of ELF. One possible explanation for such an exclusion is the fact that ELF is still seen to be limited to certain domains. However, as the corpus materials used here show, it is clear that the underlying determinants of the expansion of English, such as digitization and technologization, have led to a situation in which the ELF serves a much broader set of communicative functions than just certain specific purposes. This is clearly seen in the digital genres covered in WrELFA, such as scientific blogs, which were not included in the results here, and in personal and thematic blogs as well as in tweets. The tweet component serves as a prime example of technologization and digitization since the material for the second-

generation ELF corpus used here comes from a government-funded site that recruits ordinary citizens to manage the official Twitter account of Sweden. The great majority of these messages are in English; indeed the entire site is in English. Despite that fact that is just one specific case, it leads to a question of the extent to which the current corpora used in world Englishes can indeed capture the diversification of English and whether the many of the corpora used actually reflect the real world. At the same time, it needs to be acknowledged that we also need more comprehensive geographic coverage of written ELF corpora, since the observations here are based on the Nordic context.

Lastly, one of the key points of ELF in Mauranen's (2012) study is that the emergence of second language use and the global spread of English add extra uncertainty to what we know of language change. She argues that the ELF may lead to a situation whereby "we do not know in which respects the processes observed in earlier research on language change are valid" (2012: 243). One example of an area where rethinking is needed is dissecting the observation that the spread of English through contact and adult language learning lead to simplification. The results presented here illustrate that some simplification takes place, especially in the most robust category of grammaticity, but there are also areas in which structural simplification is not present. The ELF corpora here display remarkable similarities with native Englishes and post-colonial L2 varieties, and new theoretical approaches are need to understand such observations. In a separate study, I have together with my colleagues (Laitinen et al. 2017) explored the idea of applying variationist sociolinguistics as the theoretical toolbox to understand ELF. We have used the social network model of the diffusion of innovations as the starting point (cf. Milroy 1987). We have suggested that ELF speakers, who are multilingual by definition, might have more weak ties in general than those who do not use English as a second language resource. These multilingual individuals could act as agents of linguistic change. Our results come from a 'big data' network of nearly 200,000 Twitter accounts in the Nordic region, where English is often used as ELF. We made use of two parameters that are automatically generated and available for third-party users in Twitter stream. Our main finding is that those who tweet in English have substantially larger number of network ties than those who primarily use the main L1s of the region (Danish, Finnish, Icelandic, Norwegian and Swedish) in the communication. This result supports the idea that the ELF settings and multilingual speakers in general favor innovation and change, and such settings

and speakers might therefore offset some of the impact of simplification that normally takes place in language acquisition.

References:

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

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

Breiteneder, Angelika. 2009. English as a lingua franca in Europe: an empirical perspective. *World Englishes* 28: 2, 256–269.

Cogo, Alessia & Martin Dewey. 2012. *Analysing English as a Lingua Franca: a Corpus-driven Investigation*. London: Continuum.

Granger, Sylviane & Paul Rayson. 1998. Automatic Profiling of Learner Texts. In Sylviane Granger (ed.), *Learner English on Computer,* 119–131. London: Longman.

Greenberg, Joseph. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26:3, 178–194.

Grieve Jack, Douglas Biber, Eric Friginal & Tatiana Nekrasova. 2010. Variation among blog text types: A multi-dimensional analysis. In Alexander Mehler, Serge Sharoff & Marina Santini (eds.), *Genres on the Web: Corpus Studies and Computational Models*. New York: Springer-Verlag.

Hundt, Marianne & Christian Mair. 1999. "Agile" and "Uptight" Genres: The Corpus-based Approach to Language Change in Progress. *International Journal of Corpus Linguistics* 4:2, 221–242. DOI 10.1075/ijcl.4.2.02hun.

Jenkins, Jennifer, Cogo, Alessia & Dewey, Martin. 2011. Review of developments in research into English as a lingua franca. *Language Teaching* 44(3): 281–315.

Kachru, Braj B. 1985. 'Standards, codification and sociolinguistic realism: the English language in the outer circle'. In Randolph Quirk & Henry G. Widdowson (eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press for The British Council.

Knight, Dawn, Svenja Adolphs & Ronald Carter. 2014. CANELC: Constructing an e-language corpus. *Corpora* 9:1, 29–56.

Leppänen, Sirpa, 2012. Linguistic and generic hybridity in web writing: the case of fan fiction. In Mark Sebba, Shahrzad Mahootian & Carla Jonsson (eds.), *Language Mixing and Code-Switching in Writing,* 233–254. London: Routledge.

Laitinen, Mikko. 2016. Ongoing changes in English modals: On the developments in ELF. In Olga Timofeeva, Sarah Chevalier, Anne-Christine Gardner & Alpo Honkapohja (eds.), *New Approaches in English Linguistics: Building Bridges*, 175–196. Amsterdam: John Benjamins. DOI: 10.1075/slcs.177.07lai.

Laitinen, Mikko & Magnus Levin. 2016. On the globalization of English: Observations of subjective progressives in present-day Englishes. In Elena Seoane & Cristina Suárez-Gómez (eds.) *World Englishes: New Theoretical and Methodological Considerations*, 229–252. Amsterdam: John Benjamins.

Laitinen, Mikko, Magnus Levin & Alexander Lakaw. In press. Charting new sources of lingua franca data: a multi-genre corpus approach. In Terttu Nevalainen, Irma Taavitsainen & Carla Suhr (eds.). Amsterdam & New York: Brill/Rodopi.

Laitinen Mikko, Jonas Lundberg, Magnus Levin & Alexander Lakaw. Forthcoming. On the globalization of English: Making use of complex Twitter data as a diagnostic tool in language choice. *Journal of Universal Computing Science*.

Laitinen, Mikko, Jonas Lundberg, Magnus Levin & Alexander Lakaw. 2017. Revisiting weak ties: using present-day social media data in variationist studies. In Tanja Säily, Minna Palander-Collin, Arja Nurmi, & Anita Auer (eds.), *Exploring Future Paths for Historical Sociolinguistics*, 303–325. Amsterdam: John Benjamins. DOI 10.1075/ahs.7.12lai.

Levshina, Natalia. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.

Low, Ee-Ling. 2016. A features-based description of phonological patterns in English as a lingua franca in Asia: Implications for standards and norms. *Journal of English as a Lingua Franca***,** 5:2, 309–332.

Mair, Christian. 2013. The World System of Englishes: Accounting for the Transnational Importance of Mobile and Mediated Vernaculars. *English World-Wide* 34, 253–278.

Mauranen, Anna. 2012. *Exploring ELF: Academic English Shaped by Non-Native Speakers*. Cambridge: Cambridge University Press.

Mauranen, Anna, Ray Carey & Elina Ranta. 2015. New answers to familiar questions: English as a lingua franca. In Douglas Biber & Randi Reppen (eds.), *Cambridge Handbook of English Corpus Linguistics*, 401–417. Cambridge: Cambridge University Press.

Milroy, Lesley. 1987. *Language and Social Networks.* Second edition. Oxford: Blackwell.

Mukherjee, Joybrato & Marianne Hundt (eds). 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap.* Amsterdam: John Benjamins.

Pietikäinen, Kaisa. Forthcoming. ELF in social contexts. In Jennifer Jenkins, Will Baker, Martin Dewey (eds.), *The Routledge Handbook of English as a Lingua Franca*. London: Routledge.

Ranta, Elina. 2013. *Universals in a Universal Language? Exploring Verb-Syntactic Features in English as a Lingua Franca*. Tampere: Acta Electronica Universitatis Tamperensis 1366.

Schneider, Edgar W. 2007**.** *Postcolonial English: Varieties Around the World (Cambridge Approaches to Language Contact): Varieties Around the World (Cambridge Approaches to Language Contact)*. Cambridge: Cambridge University Press.

Seidlhofer, Barbara. 2011. *Understanding English as a Lingua Franca.* Oxford: Oxford University Press.

Seidlhofer, Barbara, Ruth Osimk-Teasdale & Michael Radeka. 2014. *Parts-of-Speech Tagging and Lemmatization Manual*. 1st revised version. VOICE Project.

Szmrecsanyi, Benedikt. 2009. Typological parameters of interlingual variability: Grammatical analyticity vs. syntheticity in varieties of English. *Language Variation and Change* 21:3, 319–353.

Szmrecsanyi, Benedikt. 2012. Analyticity and syntheticity in the history of English. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Cambridge Handbook of the History of English*, 654–665. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Kortmann, Bernd. 2011. Typological profiling: learner Englishes versus indigenized L2 varieties of English. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, 167–187. Amsterdam: John Benjamins.