Using Shallow Syntactic Features to Measure Influences of L1 and Proficiency Level in EFL Writings

Andrea Horbach[⋆], **Jonathan Poitz**[⋆], **Alexis Palmer**[†]
[⋆] Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
[†]Institute for Natural Language Processing, Stuttgart University, Stuttgart, Germany
[⋆](andrea|jpoitz)@coli.uni-saarland.de, [†]alexis.palmer@ims.uni-stuttgart.de

Abstract

This paper proposes a framework for modeling and analyzing differences between texts written by different subgroups of learners of English as a Foreign Language (organized according to native language (L1) and proficiency level). Using frequency vectors of both POS-trigrams and mixed POS and function word trigrams, we compare learner language variants both to each other and to native English, German, and Chinese texts. We introduce the *trigram usage factor* metric for identifying sequences that are especially characteristic of a particular subgroup of learners. We show that distance between learner English and native English decreases with proficiency. Next we compare the distance between learner English and other native languages. Finally, we show that automatic proficiency classification benefits from using L1-specific classifiers.

1 Introduction

When learning to write in a foreign language (L2), learners tend to make some errors that arise via the transfer of properties of their native language (L1). In other words, sometimes lexical, syntactic, semantic, or pragmatic characteristics of a learner's L1 arise in L2 writing in ways that are either wrong or simply not typical for native speaker writers. We build on the notion of Selinker (Selinker, 1972), who introduced the concept of *interlanguage*, the specific language systems of individual language learners. A learner's interlanguage includes, among other influences, features of the learner's L1, and speakers of the same L1 often develop similar interlanguages.

In this paper, we propose a new way of modeling learner language that allows us to compare L2 texts produced by learners with various L1s both to each other and to texts written by native speakers of various languages. We investigate, via several different exploratory studies, the role of L1 influences on the shallow syntactic structures produced by learners of English as a Foreign Language (EFL).

Our shallow syntactic analysis consists of part-of-speech (POS) tags and certain lexical items, primarily closed-class function words. In this way we abstract away (to a large extent) from lexical biases due to topic, and instead focus on syntactic aspects of the learner language. This approach has also been used in work on Native Language Identification (Nagata and Whittaker, 2013; Wong et al., 2012). We build a vector space of trigram frequencies for different groups of learners of English, as well as for native speakers of several languages, and we use these vectors to compare language variants, using one standard similarity metric and one novel similarity metric. The models are described in more detail in Sec. 4.

The first aim of the study is to confirm the validity of this modeling approach in the language learning context (Sec. 5.1). Our model shows (not surprisingly) that native English and L2 English indeed differ in the distribution of our vector components: learners of English use structures with different frequencies than native speakers. A key finding here is that the distance between native English and L2 English, measured by distributional similarity in the trigram vector space, decreases as learners become more proficient, showing the validity of our model.

We further investigate how these deviations vary across different L1s, identifying certain patterns of deviation that can be linked to syntactic properties of the L1 (Sec. 5.2). Here we introduce our *trigram usage factor* metric, which allows us to identify particular trigrams which are either over- or underused by a particular group of

Andrea Horbach, Jonathan Poitz and Alexis Palmer 2015. Using shallow syntactic features to measure influences of L1 and proficiency level in EFL writings. *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015*. NEALT Proceedings Series 26 / Linköping Electronic Conference Proceedings 114: 21–34.

learners. Brief case studies for English written by speakers of German, Japanese, Turkish, and French show that our model picks up interesting L1-specific properties. We further find that instances of overused trigrams often represent stylistic differences rather than actual errors, and only in certain selected contexts can the usage factor help to automatically identify problematic constructions in learner text.

Next, we consider how the influence of students' L1 changes as learners become more proficient in the relevant L2, in this case English (Sec. 5.3). We investigate this by measuring the similarity between various English learner groups and texts written by native speakers of English, German, and Chinese.

This investigation requires mapping the POS tags for English, German, and Chinese into the Universal Tagset (Petrov et al., 2012), a coarse-grained tagset designed to be suitable for all languages (as the name suggests). We use existing mapping scripts to convert tagsets for the three languages into the Universal Tagset, and we build a new vector space based on the coarse-grained POS tags. In every case, even low-proficiency L2-English is closer to native English than to either native German or native Chinese. Some effects seem to be due at least in part to typological differences between L1s.

Finally, building on the observation that trigram distributions change as learner proficiency increases, we use trigram vectors as features for a simple learner-proficiency classifier (Sec. 5.4). The results of this very preliminary study are mixed: though the features are not able to beat a simple baseline, we do show that the accuracy of proficiency classification improves when we classify groups of essays written by learners with a shared L1. In other words, the changes in trigram distributions according to proficiency are at least to some extent influenced by the native language of the learner.

2 Related Work

Aspects of our approach are similar to some work in grammatical error detection that also makes use of trigrams or similar measures. For example, the ALEC system (Chodorow and Leacock, 2000) compares the local context of a specific word in an essay to the context in a native corpus to identify erroneous usages in learner texts.

Tetreault and Chodorow (2009) use region specific web counts to identify linguistic phenomena on the lexical level that are particularly problematic for a certain geographic region, i.e. speakers of a certain L1. They compare how often a certain construction that can be indicative of an error is used in comparison to its correct counterpart in that region and compare this ratio to the one in a native population. In this way, they reliably detect constructions corresponding to common errors for learners of that L1. The approach to model learner language for multiple individual L1s is not commonly integrated into Automatic Error Detection, but used also in some other works such as (Hermet and Désilets, 2009).

Sun et al. (2007) use so-called labeled sequential patterns that overcome the locality of trigrams and consist of (not necessarily consecutive) sequences of words that might be indicative of errors. They mine such patterns and use them to classify correct and erroneous sentences.

While these approaches mostly focus on lexical items and errors connected to them, we stay with our analyses on the side of POS and mixed model trigrams. In terms of error detection, we thus lack the granularity needed for this task and rather observe over- and underusages that might be indicative of errors but do not directly allow error classification. However, for the goal of comparing different language learner variants as a whole to native English, we obtain models that avoid data sparseness and filter out most of the topic bias present in lexical models.

3 Data and Preprocessing

This section describes the four corpora used in our experiments and preprocessing steps. The primary corpus is the ETS Corpus of Non-Native Written English, which contains essays in English from learners from eleven different L1s. The secondary resources used are three corpora of texts written by native speakers: LOCNESS for English, the FalkoEssayL1 corpus for German, and the Penn Chinese Treebank for Chinese.

3.1 Corpora

The ETS Corpus of Non-Native Written English. The ETS corpus (Blanchard et al., 2014) contains a total of 12,100 essays (more than 4 million tokens) from EFL learners of eleven different L1 origins, namely Arabic, Chinese, French, Ger-

	low		medium		high	
Arabic (ARA)	66146	(296)	197217	(605)	77234	(199)
German (DEU)	3711	(15)	142380	(412)	268309	(673)
French (FRA)	13839	(63)	195455	(577)	181202	(460)
Hindi (HIN)	8670	(29)	151265	(429)	263322	(642)
Italian (ITA)	37307	(164)	201745	(623)	117699	(313)
Japanese (JPN)	46451	(233)	220426	(679)	75236	(188)
Korean (KOR)	35754	(169)	228526	(678)	106199	(253)
Spanish (SPA)	19904	(79)	192858	(563)	184641	(458)
Telugu (TEL)	27968	(94)	229723	(659)	139085	(347)
Turkish (TUR)	19636	(90)	208241	(616)	158060	(394)
Chinese (ZHO)	24661	(98)	258462	(727)	114992	(275)

Table 1: Number of tokens (and essays) per language and proficiency

man, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish. The dataset is composed of responses in the TOEFL test to 8 different prompts and is mainly used for native language identification tasks. It is thus balanced over languages, i.e. 1100 essays per language. The essays also come with proficiency information on three levels (*low*, *medium* and *high*). Table 1 shows the distribution over languages and proficiency levels. We can see that the levels are not balanced and we have substantially more essays from a medium proficiency range than for low or high proficiency.

Proficiency levels are derived from 5-point essay scores assigned by human raters, who adressed various aspects of an essay in their grade, such as lexical choice, grammar, coherence and argumentative structure.

The LOCNESS corpus. The LOCNESS corpus[1] contains 410 essays from British and American high school students, amounting to 320,000 tokens of text. We use it as a comparison corpus for comparing the different variants of L2 writings to native English of the same text type, i.e. argumentative essays.

The Falko-L1 corpus. The FalkoEssayL1 corpus (Reznicek et al., 2012) is a corpus of native German argumentative essays written by students in response to four different prompts. It contains 95 texts with a total of approximately 70,000 tokens. The texts have been error-annotated and normalized. We use in our experiments the so-called target hypothesis *ZH1* that has the goal of correcting primarily orthographical and morphosyntactic errors. This version of the corpus is chosen over the raw essay texts in order to minimize POS tagging problems due to misspelled and therefore unknown words.

The Penn Chinese Treebank. The Penn Chinese Treebank (Xue et al., 2002) is a corpus of Chinese news texts that comes already with - among other annotation layers - manual annotations for word segmentation and POS tags.

3.2 Preprocessing

The ETS corpus is already tokenized, and we use this tokenization. Falko and Penn Chinese Treebank come with token and POS annotations. LOCNESS requires sentence-splitting and tokenization, for which we use the OpenNLP toolkit.[2] The final step needed to have suitable input for our models is POS tagging. We use Treetagger (Schmid, 1994), which uses a refined form of the Penn Treebank tagset (Marcus et al., 1993), to tag all English texts. For a description of these tags, refer to Table 2. The other two corpora are pretagged, and in both cases we use the existing POS tags. Falko corpus texts (as well as the normalized form we use) have been tagged with the Treetagger and the STTS tagset (Schiller et al., 1999), and the Penn Chinese Treebank comes with manual POS annotations.

4 Models

The core of our modeling approach are trigrams in learner essays. N-gram features have proven useful in many natural language processing applications, including those aiming to capture differences between non-native texts written by learners

[1] https://www.uclouvain.be/en-cecl-locness.html

[2] https://opennlp.apache.org/

POS Tag	Meaning	POS Tag	Meaning
#	”#” character	RBR	adverb, comparative
$	currency symbol	RBS	adverb, superlative
“	opening quotes	RP	particle
”	closing quotes	SENT	end punctuation
(	opening braces (”(” or ”{”)	SYM	symbol
)	closing braces (”)” or ”}”)	TO	”to”
,	”,” character	UH	interjection
:	general joiner	VB	verb be, base form
CC	coordinating conjunction	VBD	verb be, past
CD	cardinal number	VBG	verb be, gerund/participle
DT	determiner	VBN	verb be, past participle
EX	existential there	VBP	verb be, pres non-3rd p.
FW	foreign word	VBZ	verb be, pres, 3rd p. sing
IN	preposition/subord. conj.	VH	verb have, base form
IN/that	complementizer	VHD	verb have, past
JJ	adjective	VHG	verb have, gerund/participle
JJR	adjective, comparative	VHN	verb have, past participle
JJS	adjective, superlative	VHP	verb have, pres non-3rd per.
LS	list marker	VHZ	verb have, pres 3rd per.sing
MD	modal	VV	verb, base form
NN	noun, singular or mass	VVD	verb, past tense
NNS	noun plural	VVG	verb, gerund/participle
NP	proper noun, singular	VVN	verb, past participle
NPS	proper noun, plural	VVP	verb, present, non-3rd p.
NS	–	VVZ	verb, present 3d p. sing.
PDT	predeterminer	WDT	wh-determiner
POS	possessive ending	WP	wh-pronoun
PP	personal pronoun	WP$	possessive wh-pronoun
PP$	possessive pronoun	WRB	wh-abverb
RB	adverb		

Table 2: Tags used for POS-tagging English content: the Treetagger version of the Penn Treebank tagset

with different L1s. One prominent example is native language identification where many systems use some sort of n-gram features (Tetreault et al., 2013). In our case, we use trigram models to capture syntactic properties of various subgroups of language learners, grouping by both L1 and proficiency level. We concentrate on trigrams as they are long enough to capture some context of a word, but do not cause sparse data problems.

We build a model of each particular learner group – for example, medium-proficiency learners whose native language is Hindi – by collecting frequency counts for a selected set of trigrams (here, the most frequent trigrams in a native English corpus). Trigram counts are extracted from the set of English essays written by that group of learners. For most studies, we build one vector for each *sub-corpus* (in this case, HIN_medium), where the vector components are frequency counts for the given trigrams. We then can think of a vector space which contains vectors for all learner sub-corpora, which we also use for comparison in parts of study 1 (Sec. 5.1). In the final study (Sec. 5.4), and also for part of study 1 (Sec. 5.1), we build one such vector per essay.

Two different types of trigrams are used to build these models (see below). In both approaches, we count only trigrams which occur within sentences, and use <SENT> to represent the start of the sentence.[3]

POS models. In the *POS models*, vectors are constructed by extracting trigram counts from POS-tagged texts. This means that each word is tagged, and the original lexical material of the text is discarded. The aim of using POS tag sequences is to abstract away from concrete topics in the data and rely as much as possible on the grammatical structures present in the text.

Mixed models. In the *mixed models*, vectors are constructed by extracting trigram counts from texts that have been transformed into a mix of POS tags and lexical items (as done similarly by Nagata and Whittaker (2013) and Wong et al. (2012)).

The motivation for our mixed models is that many learner deviations manifest on the lexico-syntactic level rather than purely on the POS level. In other words, it often matters not just whether a preposition is used, but which one, or not only whether an article is used at all, but whether it is definite or indefinite. Those differences are captured by our mixed model, while still filtering out content-bearing material.

For open-class words, like nouns, verbs, and adjectives, words are replaced by their POS tags. Function words and closed-class words such as prepositions and articles remain unchanged.[4] Adverbs (RB) are a special case: we differentiate between those that end in *-ly*, which we treat as open-class, and all other adverbs, which we treat as closed-class. While this simple distinction is correct in most cases, there is room to further refine this heuristic. For instance, the word *only* is both an adverb and ends in *-ly*, however it is not a content word. Also the categories RBR and RBS (comparative and superlative, respectively) are not completely clear-cut. RBR can be the part of speech for function words like *more, less*, but likewise for content words like *better, faster, stronger*, and similarly for the superlative RBS tag.

5 Explorations

Having established the modeling set-up and model variants, we now describe the various studies in which we use this modeling framework to investigate L1 influences and their relation to learner proficiency level.

The first two studies (Sec. 5.1 and Sec. 5.2) examine L1-specific correlates in L2 writings, showing that essays written by EFL learners show certain properties specific to their native language. Some of the deviations from native English seen in learner essays can be attributed to specific syntactic differences between the languages, while others are characteristic of learner language in general. The third study (Sec. 5.3) compares EFL learner essays to native-speaker essays in German and Chinese, and the final study (Sec. 5.4) makes a first attempt to use our modeling framework for automatic proficiency classification.

5.1 Study 1: Measuring the distance between native and non-native English

In this study we investigate how far away from native English different learner groups are (i.e. individual combinations of L1 and proficiency level),

[3] We only allow this as the first word in a trigram.

[4] More specifically, lexical items are replaced by their POS tags when those tags are any of the following (from the Penn Treebank tagset): FW, JJ, JJR, JJS, NN, NNS, NP, NPS, RBR, RBS, UH, VB, VBD, VBG, VBN, VBP, VBZ, VH, VHD, VHG, VHN, VHP, VHZ, VV, VVD, VVG, VVN, VVP, VVZ, NS and CD.

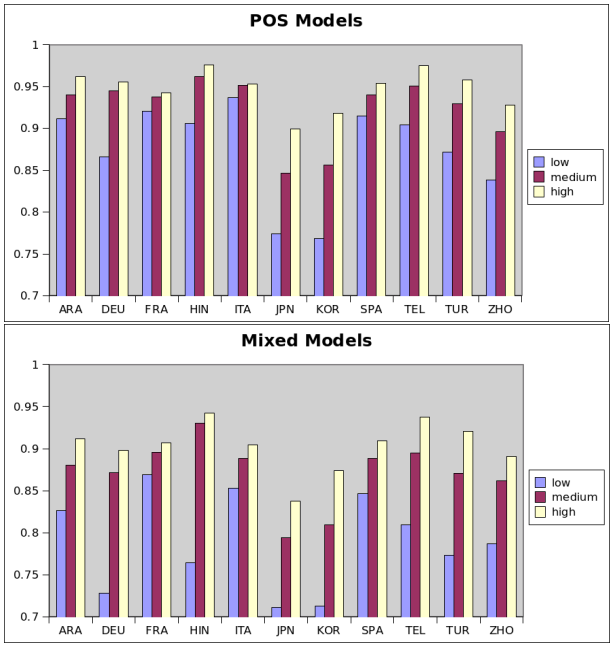

Figure 1: Cosine similarity between non-native English variants and native English, computed *per subcorpus*, for three different proficiency levels, using both POS trigrams (top) and mixed trigrams (bottom)

investigating differences both on the subcorpus level and on the essay level.

First, we model native English by building two feature vectors from the LOCNESS corpus, one with POS trigrams and one with mixed trigrams (see Sec. 4), each vector containing the 500 most frequent trigrams for that version of the corpus, with their raw frequency counts. Then, for each L1-proficiency subcorpus (i.e. for 33 subcorpora of the ETS corpus), we again build two feature vectors, each containing the absolute frequencies within the given subcorpus of the respective (POS- or mixed-trigram) top-500 native English trigrams. Finally, to measure distance between each learner language variant and native English, we compute the *cosine similarity* between each of the non-native vectors and native English. Results appear in figure 1.

We see that – as expected – for both models, and for all L1s, low-proficiency learner English variants differ the most from native English. Furthermore, the gap between low and medium proficiency is always bigger than that between medium and high. It is likely that many of the differences between medium and high proficiency are too subtle to be captured by the mixed-model trigrams and manifest rather on the side of appropriate lexical choice within the same POS category.

We see further that similarity for the POS-models is generally higher than for mixed-models, and that especially the gap between low and medium is more pronounced for mixed trigrams.

Among those languages whose low- and medium-level variants are most dissimilar to native English are mainly non-Indo-Germanic languages such as Japanese, Korean and Chinese.

One should note that the proficiency level of an essay is based on a score that also integrates aspects of an essay that cannot be grasped by a trigram model, such as discourse structure; this limits the extent to which we can capture proficiency with our model. Furthermore, while we tried to compare corpora that are as similar as possible in the sense that they both contain argumentative essays, some of the dissimilarities might stem from structural differences like e.g. the topics of the essays in the corpora.

We also compare on a per-essay level with native English, by building one feature vector per essay and comparing to the feature vector for the complete native English corpus.

The results (cf. figure 2) confirm the effects observed per subcorpus. On average, similarity per essay is lower than similarity per subcorpus, which can be explained by the high number of features; not all of the top-500 trigrams occur in every essay. In addition, when aggregating counts over a subcorpus, over- and under-usages of individual trigrams in individual essays tend to cancel each other out. The overall trend confirms that higher-proficiency individual essays are closer to native English than lower-proficiency essays.

5.2 Study 2: Identifying L1-specific deviations in trigram distributions

In Study 1, we show that low-proficiency non-native Englishes are more different from native English on the mixed-model trigram level than medium or high-proficiency variations. We next investigate how different ETS subcorpora (i.e. different combinations of L1 and proficiency level) differ from one another. More specifically, we introduce the *trigram usage factor (TUF)* metric, which computes the relative frequency for an individual trigram across two language variants. TUF allows us to identify individual trigrams which are especially characteristic of par-

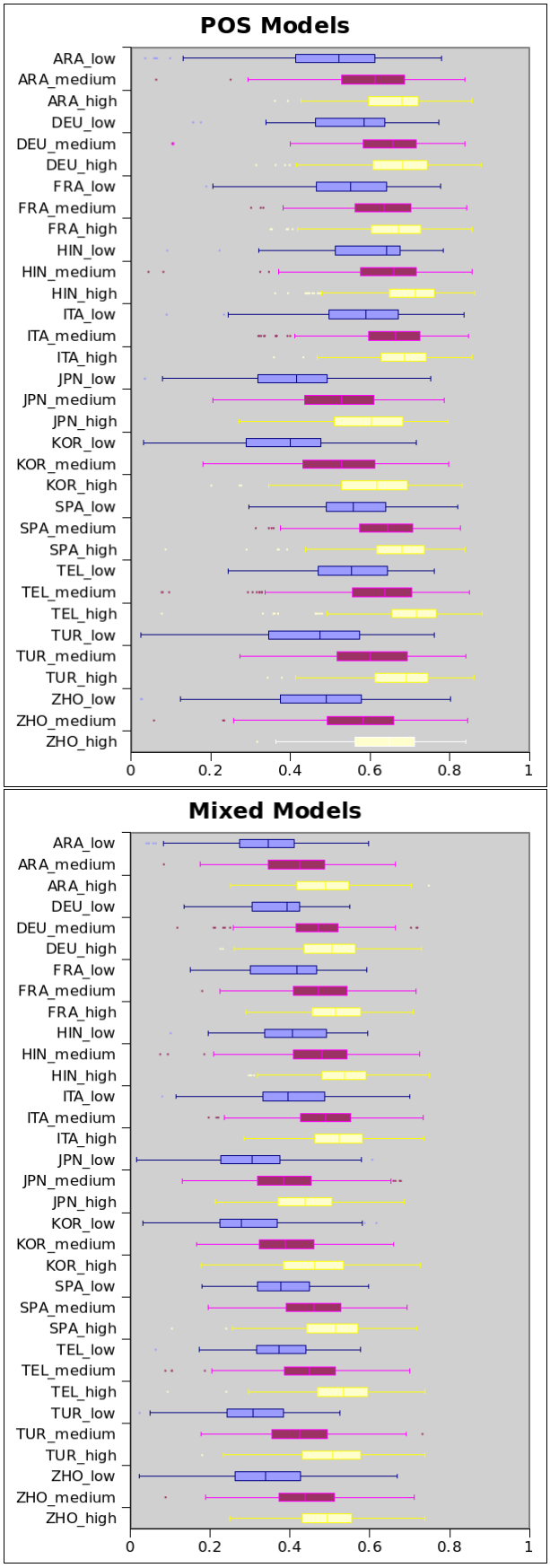

Figure 2: Cosine similarity between non-native English variants and native English, computed *per essay*, using both POS-trigrams (top) and mixed-trigrams (bottom)

ticular L1-proficiency learner groups.

Trigram Usage Factor. To measure how individual trigrams deviate from native English, we compute for each trigram the *usage factor* for a language-proficiency combination by dividing the relative frequency of the trigram t for the relevant subcorpus by the relative frequency of t in native English.

$$\text{TUF}_{Native}(t) = \frac{\text{FREQ}_{L_1\text{-}proficiency}(t)}{\text{FREQ}_{Native}(t)}$$

For example, a usage factor of 4.2 for the trigram *VVP JJ NN* for low-proficiency Japanese means that this trigram occurs 4.2 times more often in essays by low-proficiency Japanese writers than in native English essays.

In the course of this analysis, it became clear that many trigrams are generally overused by most L2 subcorpora, such as *<SENT> for NN*, where *<SENT>* stands for the start of a sentence. We interpret these trigrams as reflecting properties of learner language that are not specific to a particular L1. Table 3 shows the top 10 most overused learner language mixed model trigrams (computed by taking all ETS subcorpora together) as compared to native English. We checked small samples of 10 instances of each of the 10 trigrams for 4 languages (German, French, Japanese, Turkish) and found that they almost never indicated errors, but correspond to frequent sentence constructions such as *I think that*, *for example,* etc as well as influences from the prompt (e.g. many instances of *young people X* and *old people X* in essays asking for a statement about the sentence *Young people enjoy life more than older people do.*).

Over- and underusages for certain phenomena and learner groups are well-known from the Second Language Acquisition literature (e.g. Odlin and Jarvis (2004)). In order to see differences between individual L1s better, we perform an alternative evaluation that is not susceptible to trigrams that are generally frequent learner language. In this variant, we compute TUF relative to the average usage across all L2 essays, by dividing the relative frequency of a trigram t for a given language-proficiency subcorpus by the relative frequency of t in the complete ETS corpus.

$$\text{TUF}_{Learner}(t) = \frac{\text{FREQ}_{L_1\text{-}proficiency}(t)}{\text{FREQ}_{Learner}(t)}$$

overusage factor	trigram	example	rank in LOCNESS
6.01	<SENT> for NN	For example	446
5.15	, i VVP	, I agree	479
4.00	<SENT> i VVP	I believe	169
3.63	i VVP that	I think that	274
2.86	VVP to VV	try to explain	85
2.84	for NN ,	for instance,	179
2.75	JJ NNS VVP	young people enjoy	130
2.65	<SENT> in NN	In conclusion	206
2.62	VVP not VV	do not agree	199
2.51	<SENT> RB ,	Finally,	201

Table 3: The top-10 most overused mixed model trigrams in general learner language as compared to native English

In doing so, we are better able to pick up differences between the different L1s, by measuring whether a certain trigram is over- or underused according to the average usage by language learners.

Study 2a: Trigram Usage Factors in Comparison to Native English

Next, we check how the usage factors of trigrams compared to native English evolve over proficiency levels. We say that a usage factor for a certain trigram evolves *towards native English*, if the usage factor for that trigram moves closer to 1 (i.e. closer to the relative frequency of the trigram in native English) over the three different proficiency levels, e.g. 0.3 for low, 0.4 for medium and 0.8 for high proficiency learners, or 3.5 (low), 2.0 (medium) and 1.3 (high). To account for cases where, for a given trigram, values for the three proficiency levels are not all on the same side of 1, we map values above 1 to their inverse. (This then covers, e.g., cases where an extreme underusage for low-proficiency moves via a moderate underusage for medium, towards only a slight overusage for high proficiency (e.g. 0.3 (low), 0.8 (medium), 1.05 (high, mapped to 0.952)). We still want to consider such cases as moving towards the native distribution, in contrast to a set of usage factors like the following that does not move towards English: 0.3 (low), 0.8 (medium), 1.5 (high, mapped to 0.67)

We perform two versions of this evaluation. In the first (marked as *all* in figure 3), we consider all three proficiency levels. The second evaluation (*low/medium*) is motivated by the cosine similarity results seen in study 1, where the distances between low and medium proficiency are more pronounced than those between medium and high proficiency. In the second evaluation, we ask how often low proficiency moves via medium towards native, excluding the high-proficiency level. We call cases that evolve towards native English where the low-proficiency usage factor is below 1 an underusage, otherwise an overusage.

If we consider all three proficiency levels, we can see that for (on average) 41% of the most frequent 500 native POS trigrams and 42% of the mixed trigrams, TUFs indeed move towards native English. In the second condition, 67% of all POS-trigrams and 65% of all mixed-trigrams move towards native English. (If usage factors varied randomly, we would expect 25% for all three levels and 50% for two levels.) The improvements are similar across languages and across the two trigram models. We see more under- than overusages. We assume that this might be because language learners use a small syntactic inventory quite often, while not exploiting the complete syntactic variety of a language.

Study 2b: Trigram Usage factors Compared to General Learner English: Case studies

We next have a closer look at the most over- and underused trigrams for the medium (i.e. medium-proficiency) level for each of four languages and try to identify properties of the L1 that might account for such overusages. (For underusages, it is generally hard to find examples where a certain trigram should have been used, but wasn't.)

We select German, French, Japanese and Turkish for closer inspection, with these languages

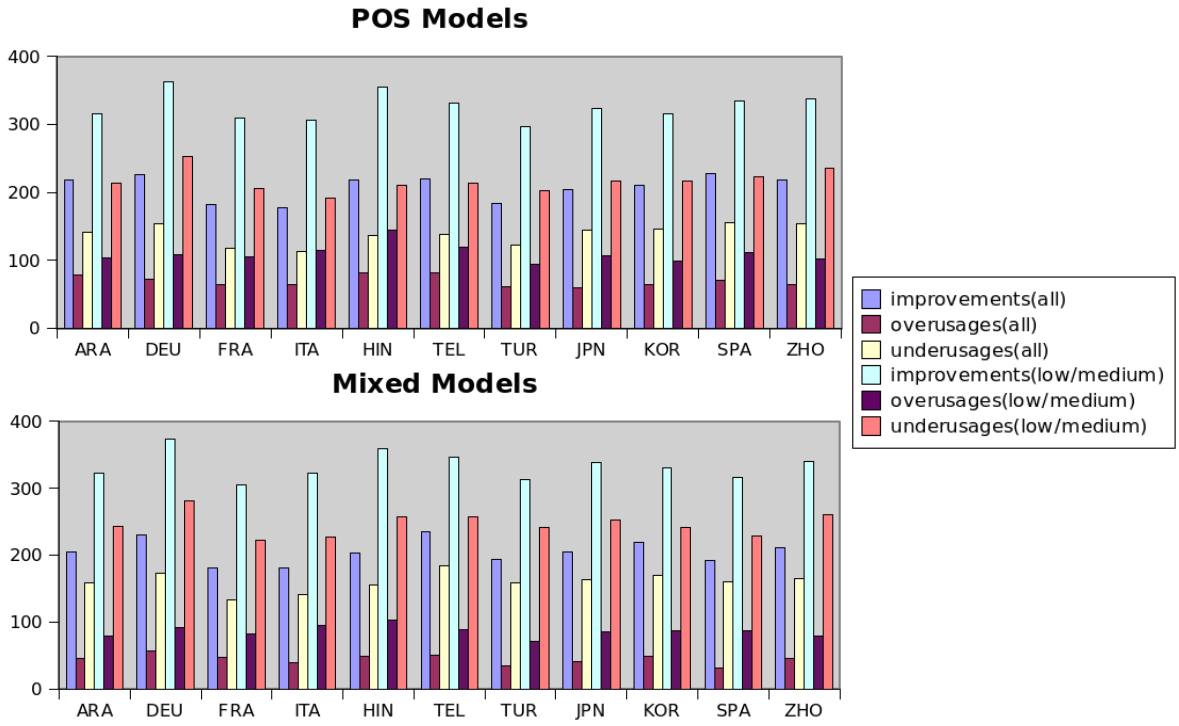

Figure 3: Number of POS and mixed model trigrams (out of the top 500) that moved to a distribution closer to native English

chosen to cover different language families and in order to pick languages of which at least one of the authors has some basic understanding. We choose the medium proficiency subcorpus for each language as it is always the largest subcorpus.

Table 4 shows the top-10 overusages per language, measured against general learner language.

German: In the case of German, the top-2 overused trigrams seem to stem from a tendency not to put a comma after an adverbial phrase at the beginning of a sentence, as in example (1) in contrast to (2).[5]

(1) At this **point it is** good ...

(2) At this point, it is good ...

Additionally, we see overusages of trigrams that correspond to certain fixed phrases such as example (3) or (4).

(3) <**SENT**> **Another (example|point|reason|...)** ...

(4) <**SENT**> **On the** other hand ...

[5]Bold print is always used for the relevant trigram in an example sentence from the ETS Corpus. Examples without bold print are constructed.

Interestingly, for low-proficiency German learners we see an underusage of some trigrams involving *will*. This could be explained by the fact that in German, the future is often expressed using a present tense verb, e.g. (5) instead of (6),[6] leading to essay sentences like (7).

(5) *Ich fahre morgen nach Frankfurt.*
∗ I go to Frankfurt tomorrow.

(6) *Ich werde morgen nach Frankfurt fahren.*
I will go to Frankfurt tomorrow.

(7) Only then **the development is** also in the future as fast as then now.

French: When looking at the top 10 overused trigrams in French, one can observe a high number of trigrams containing infinitive verb constructions like, among others, *VBZ to VV*, *to VB JJ* or *, to VV*. Such a distribution could either point to a high number of infinitive constructions in French as compared to English or to the absence of infinitive constructions and thus an exaggeration of the usage of such structures during learning. However, we could not find evidence for either of the one being the case.

[6]Examples shown with literal translations.

rank	German	French	Japanese	Turkish
1	NN it VBZ	it VBZ a	<SENT> first ,	can not VV
2	NN i VVP	<SENT> that VBZ	, there VBP	<SENT> as a
3	<SENT> another NN	, to VV	i VVP with	<SENT> JJ of
4	VH a JJ	VBZ to VV	<SENT> therefore ,	JJ of all
5	to VH a	when you VVP	i VVP to	<SENT> if you
6	not JJ to	and to VV	VVP not VH	RBS JJ NN
7	to VV this	to VV ,	can VV JJ	VVG the NNS
8	RBR JJ to	to VB JJ	NN , i	this NN ,
9	NP NP NP	NN , we	, i VVP	the NNS ,
10	<SENT> on the	NN , you	NN to VVG	as a NN

Table 4: The top-10 most overused mixed model trigrams in the medium level of four L1 variants of learner language as compared to learner language in general

One can, however, see another trend in the top-ranked trigrams. Contrary to general learner language, French speakers tend to overuse constructions with *you* and *we*. In the top 15 trigrams, two contain *you* (*when you VVP* and *NN , you*) and two *we* (*NN , we* and *we can VV*), e.g. (8) and (9). These could indicate that French speakers adopt a different perspective when writing argumentative texts. One possible reason for this could be the indefinite pronoun *on* in French that – in colloquial, spoken situations – often replaces the morphologically more complex *nous* form of the verb, e.g. (10). In written situations, its purpose is rather to refer to unknown or generalized entities or to replace the use of the passive voice, as in (11) and (12). This ambiguity could be an explanation of the learners' difficulty to choose the appropriate pronoun.

(8) When we are young it is very useful to try a lot of subject but **when you grow** up things change.

(9) In your **argumentation , we** will present some elements in order to give our own opinion.

(10) *On va / Nous allons à la plage.*
We go to the beach.

(11) *On m'a demandé de te donner cela.*
I was asked to give you this.

(12) *On ne sait jamais . . .*
One never knows . . .

Japanese: For Japanese learners, we see an overusage of trigrams involving formulaic language (*First, . . .*) and repetitions of the prompt. The third most overused trigram (*i VVP with*) arises almost exclusively from phrases like (13) and (14). The second most overused sequence covers almost exclusively existential constructions like *there is* or *there are*.

(13) **I agree with** this statement.

(14) **I disagree with** this statement.

A trigram like *can VV JJ* (together with other top 20 Japanese overused trigrams such as *not VV JJ* or *VH JJ NN*) points at problems with article usage, which can be explained by the absence of articles in Japanese. While there are of course valid instantiations of such patterns such as (15), other occurrences of these trigrams actually point at errors such as (16), (17) or (18).

(15) Young people **can do many** things.

(16) They can **get good mark**.

(17) Old people [. . .] **have long life** expectancy.

(18) Young people do **not feel strong** relationship.

Turkish: In Turkish, there are no definite articles, which results in learner texts with an interesting distribution of trigrams involving *the*. Among the top 30 overused trigrams in Turkish, 7 contain *the* (e.g. *VVG the NNS*, *the RBS JJ*, and *the NNS ,*). One can observe a steady trend for these trigrams across proficiency levels. While low proficiency learners' trigram distribution ranges between under- to slight overusage, medium and

high levels strongly overuse them. This is a possible manifestation of a learner's behavior when dealing with a grammatical feature that is nonexistent in their mother tongue – at a low level, they tend to not use it due to a lack of knowledge and confidence. At a higher level, they may overcompensate by trying to fit it in places where it is syntactically or semantically incorrect.

When looking at underused trigrams, what is striking is that half of the top 20 underused trigrams involve the use of a preposition like *for, to, in, with* or *by*. This effect is not surprising as in the Turkish language adpositional phrases are constructed differently than in English or in many Indo-European languages. First of all, Turkish is a strictly head-final language which uses postpositions instead of prepositions. Secondly, English prepositions cannot be – in many cases – directly mapped to their most obvious counterpart in Turkish. The Turkish dative and locative case, for instance, replace certain prepositional phrases in English. The dative case often conveys a sense of movement, e.g. (19), while the locative case is used to refer to a static position as in (20). These examples show how Turkish differently treats temporal and spatial relations that are conveyed by English prepositional phrases.

(19) *Ankara'ya gidiyorum.*
Ankara'[dat.] go[pres.][1.p.sg.]
I'm going to Ankara.

(20) *Ankara'da yaşıyorum.*
Ankara'[lok.] live[pres.][1.p.sg.]
I live in Ankara.

Exploring the Potential for Error Detection

The ability to identify heavily over- or underused sequences in learner language via the TUF metric suggests the potential application of automatically detecting errors in learner language. In fact though, strong over- or underusage of a particular trigram by a learner of a particular L1 might in some instances indicate an error, in other instances it is just an overusage of an otherwise correct phenomenon.

When, for example, low proficiency Japanese learners show a heavy overusage of a trigram of the form *VVP JJ NN* some of these instances might indicate a missing article, as we have seen in some of the examples above. There would be of course the alternative that Japanese low-proficiency learners overuse constructions with mass nouns such as *drink cold milk*. On the other hand, some overused constructions might be attributed to simple formulaic language, such as *i VVP* which is often used in constructions like *I think* etc.

To get a better understanding of how well the usage factor metric can be used for error detection, we perform a small, preliminary annotation study. We annotate 10 random instances each for the top-10 overused trigrams for medium German learners, and for the top 10 generally-overused trigrams (with examples taken also from the medium German learner corpus). We check the underlying learner essays for errors within the range of that phenomenon in order to determine whether they are associated with errors, or rather with non-erroneous but overused phenomena.

In this study we found almost all instances of overusages to be grammatical. Only very few pointed at actual errors, while others point at constructions where there is some preferable alternative. Despite poor results from this small pilot annotation, further investigation of this method for detecting errors may be warranted.

5.3 Study 3: Cross-checking learners against German and Chinese native language distributions using tagset mappings

We have seen that trigram deviations vary across L1s, and we have argued that these variations are due to influences from the L1. In this next study, we investigate a question that naturally arises from this claim that low-proficiency learners are indeed "closer" to their native language (even when writing in a second language) than are high-proficiency learners. The question is whether we observe the opposite trend when comparing L2 essays to texts written natively in the L1.

In order to test this hypothesis, we compute similarities between the non-native (ETS) and native (LOCNESS) English data and two additional native corpora, the German Falko corpus and the Penn Chinese Treebank (see Sec. 3), in order to compare to one language from the same language family (Germanic) and another language that is typologically (and phylogenetically) quite far from English.

The domain of the Falko essay corpus are argumentative essays written by students, making the corpus comparable to the ETS data. For Chinese, we use news texts, as we were unable to locate a

native language Asian essay corpus.

Computing similarity between L2-English essays and texts written in other languages of course requires some modifications to the model. First, mixed models are not relevant for obvious reasons; we are limited to the pure POS models. Second, because different languages generally use different POS tagsets, we need to map these tags into a common representation. For this we use the universal POS tagset proposed by Petrov et al. (2012) and existing scripts for mapping various tagsets (including Penn Treebank, STTS for German and Penn Chinese Treebank) into the following 12 coarse-grained POS tags: "NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), . (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words)".

We then evaluate by building feature vectors for native English, German and Chinese by taking all mapped POS trigrams for that corpus into consideration and computing pairwise similarity between these three corpora and the per-language subcorpora from ETS (see figure 4).

The comparison with native English via POS-mapped trigrams confirms that the increasing similarities for higher-proficiency L2 writing still show on the coarser level of mapped trigrams. We see a similar pattern to that in figure 1.

The comparison with the German data shows a slightly different picture. For the non-European languages Arabic, Hindi, Japanese and Korean, we see a similar behavior as for English: with increasing proficiency students' writing also comes closer to native German. We argue that this might be due to the close relatedness between German and English as two Germanic languages. For Telugu, Turkish and Chinese this pattern is only valid for low and medium proficiency while European languages (except for a tendency in French) do not show this behavior. Unexpectedly, high-proficiency German learners are closer to native German than low-proficiency Germans, maybe an effect of coming closer to the full expressiveness of Germanic languages.

In the comparison with Chinese, we can see that similarity is generally lower than for German or even native English, and we observe that other

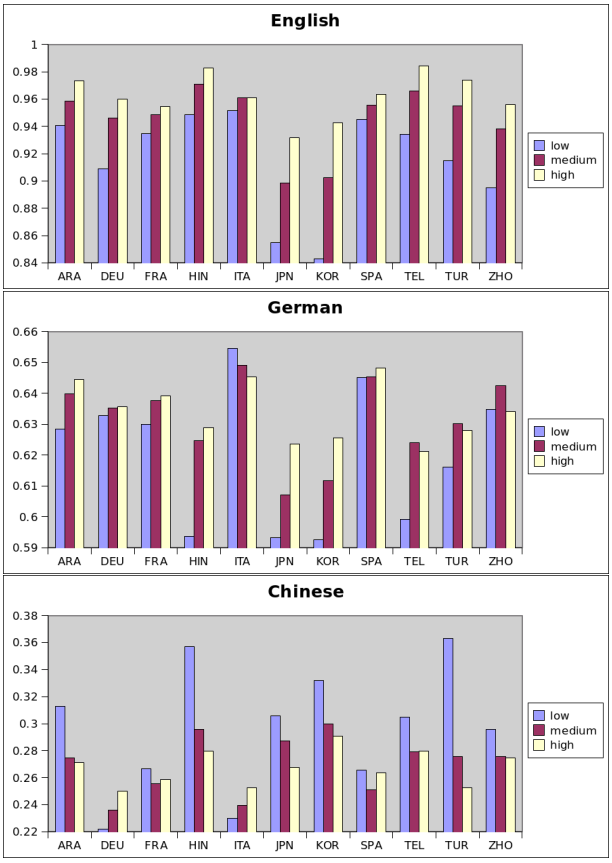

Figure 4: Cosine similarity between non-native English variants and native English (top), German (middle) and Chinese (bottom) on the level of mapped POS trigrams

Asian languages generally have a higher similarity with Chinese than do European languages. In order to exclude this lesser similarity stemming from domain effects instead of language effects, we also compared to the German TIGER corpus (Brants et al., 2004) of newspaper texts and found similarities in a range comparable to Falko (interestingly, the similarities were higher for TIGER than for Falko), with very similar tendencies for the individual L1 subcorpora.

These observations of similarities between language families echos findings by Nagata and Whittaker (2013), who reconstruct Indo-European language family relations from language models of non-native writings.

5.4 Study 4: Exploring the Use of Trigram Models for Proficiency Classification

We have shown that different L1s as well as different proficiency levels lead to different trigram frequency distributions that deviate from those for native English. As a final exploratory experi-

Features	general	L1 specific
baseline	68.8	70.5
top 500 trigrams	46.7	48.9
baseline + top 500 trigrams	46.5	49.7
selected attributes (all)	69.8	**71.5**
selected attributes (trigrams)	59.1	62.9

Table 5: Averaged classification accuracy when training on datasets for individual L1s and on mixed corpora

ment, we begin the investigation into whether vectors from our mixed models are beneficial for the task of automatic proficiency classification into the three proficiency levels low, medium and high.

While both lexical and POS trigrams have been used in related work on automatic grading of learner texts (Yannakoudakis et al., 2011), we are specifically interested in investigating the effectiveness of L1-specific classifiers.

We operationalize this question using two different feature sets. We use a baseline that consists of just 5 features: number of tokens, number of sentences, average number of tokens per sentence, number of individual types and type-token-ratio. Additionally, we use the frequencies of the most frequent 500 native English trigrams as features.

For classification, we train an out-of-the-box logistic regression model using the WEKA toolkit (Hall et al., 2009). We train and evaluate classifiers per L1, using all 1100 (per language) essays and leave-one-out cross-validation. For comparison, we additionally sample 11 disjoint "general" sets of 1100 essays from all 11 languages, with equal amounts of essays per language in each subcorpus. We use the same cross-validation procedure in order to have training corpora of compatible size. We use each of the two features sets individually and combined (cf. table 5)

This baseline is already very strong, and the new trigram features (both alone and in combination) perform far worse than the baseline. However, all feature combinations benefit from L1 specific classifiers.

A plausible reason for this degradation in performance is the excessive number of features. Thus we employ feature selection to identify the best performing features. Specifically, we use WEKA's CfsSubsetEval attribute selection method to identify the most helpful features from both the trigrams and the baseline features. If we use these features for classification (thus simulating an optimal classifier for a dataset), we get improvement from the trigram features over the baseline and again see a better performance for the L1 specific models over the general models.

We take these first results as an indicator that proficiency classification can further profit from L1 information and will investigate this classification task further in future work.

6 Conclusions and Future Work

In this paper we have shown how two important factors influencing EFL writings, L1 and proficiency level, influence the shallow syntactic structure of essays. Using frequency vectors of trigrams, we investigate various attributes of learner language, using both cosine similarity and our own trigram usage factor metric. We hope this framework will be useful for further investigations into learner language, automatic error detection, and automatic proficiency classification.

What we have not covered so far in our experiments is a third important factor: the influence of the task, in our case the essay prompt. In the course of performing the case studies and annotation pilot described here, we have seen that the prompt can be visible even on the abstraction level of POS models. For example, students that write essays in response to the prompt (21) frequently reused the prepositional phrase *In twenty years*, which resulted in higher frequency counts for the POS trigram *PP CD NNS*.

(21) Do you agree or disagree with the following statement? *In twenty years, there will be fewer cars in use than there are today.* Use reasons and examples to support your answer.

In future work we therefore plan to use clustering techniques to measure the influence that each of the three influence factors (L1, prompt and proficiency level) have on the trigram distributions of essays and to explicitly quantify the influence of the prompt.

7 Acknowledgements

We would like to thank three anonymous reviewers for their helpful comments. We also thank Annemarie Friedrich for fruitful discussions and Helmut Schmid and Richard Eckart de Castilho for their valuable comments regarding our work with Treetagger. This work was funded by the Cluster

of Excellence 'Multimodal Computing and Interaction' of the German Excellence Initiative. The third author is supported by SFB-732 'Incremental Specification in Context'.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2014. ETS corpus of non-native written English.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation, Special Issue*, 2(4):597–620.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 140–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Matthieu Hermet and Alain Désilets. 2009. Using first and second language models to correct preposition errors in second language authoring. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–72. Association for Computational Linguistics.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Ryo Nagata and Edward W. D. Whittaker. 2013. Reconstructing an Indo-European family tree from non-native English texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1137–1147. The Association for Computer Linguistics.

Terence Odlin and Scott Jarvis. 2004. Same source, different outcomes: A study of Swedish influence on the acquisition of English in Finland. *International Journal of Multilingualism*, 1(2):123–140.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proc. of LREC*.

Marc Reznicek, Anke Lüdeling, and Franziska Schwantuschke. 2012. Das Falko-Handbuch: Korpusaufbau und Annotationen: Version 2.0.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, IMS-CL, University Stuttgart.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1–4):209–232.

Guihua Sun, Xiaohua Liu, Gao Cong, Ming Zhou, Zhongyang Xiong, John Lee, and Chin-Yew Lin. 2007. Detecting erroneous sentences using automatically mined sequential patterns. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 81–88, Prague, Czech Republic, June. Association for Computational Linguistics.

Joel R Tetreault and Martin Chodorow. 2009. Examining the use of region web counts for esl error detection. In *Web as Corpus Workshop (WAC5)*, page 71.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 699–709. Association for Computational Linguistics.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated Chinese corpus. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.