Native Language Cognate Effects on Second Language Lexical Choice

Ella Rabinovich[⋆▲] **Yulia Tsvetkov**[†] **Shuly Wintner**[⋆]

[⋆]Department of Computer Science, University of Haifa
[▲]IBM Research
[†]Language Technologies Institute, Carnegie Mellon University
ellarabi@gmail.com, ytsvetko@cs.cmu.edu shuly@cs.haifa.ac.il

Abstract

We present a computational analysis of cognate effects on the spontaneous linguistic productions of advanced non-native speakers. Introducing a large corpus of highly competent non-native English speakers, and using a set of carefully selected lexical items, we show that the lexical choices of non-natives are affected by cognates in their native language. This effect is so powerful that we are able to reconstruct the phylogenetic language tree of the Indo-European language family solely from the frequencies of specific lexical items in the English of authors with various native languages. We quantitatively analyze non-native lexical choice, highlighting cognate facilitation as one of the important phenomena shaping the language of non-native speakers.

1 Introduction

Acquisition of vocabulary and semantic knowledge of a second language, including appropriate word choice and awareness of subtle word meaning contours, are recognized as a notoriously hard task, even for advanced non-native speakers. When non-native authors produce utterances in a foreign language (*L2*), these utterances are marked by traces of their native language (*L1*). Such traces are known as *transfer* effects, and they can be phonological (a foreign accent), morphological, lexical, or syntactic. Specifically, psycholinguistic research has shown that the choice of lexical items is influenced by the author's L1, and that non-native speakers tend to choose words that happen to have *cognates* in their native language.

Cognates are words in two languages that share both a similar meaning and a similar phonetic (and, sometimes, also orthographic) form, due to a common ancestor in some protolanguage. The definition is sometimes also extended to words that have similar forms and meanings due to *borrowing*. Most studies on cognate facilitation have been conducted with few human subjects, focusing on few words, and the experimental setup was such that participants were asked to produce lexical choices in an artificial setting. We demonstrate that cognates affect lexical choice in L2 spontaneous production on a much larger scale.

Using a new and unique large corpus of non-native English that we introduce as part of this work, we identify a *focus set* of over 1000 words, and show that they are distributed very differently across the "Englishes" of authors with various L1s. Importantly, we go to great lengths to guarantee that these words do not reflect specific properties of the various native languages, the cultures associated with them, or the topics that may be relevant for particular geographic regions. Rather, these are "ordinary" words, with very little culture-specific weight, that happen to have synonyms in English that may reflect cognates in some L1s, but not all of them. Consequently, they are used differently by authors with different linguistic backgrounds, to the extent that the authors' L1s can be identified through their use of the words in the focus set. The signal of L1 is so powerful, that we are able to reconstruct a linguistic typology tree from the distribution of these words in the Englishes witnessed in the corpus.

We propose a methodology for creating a focus set of highly frequent, unbiased words that we expect to be distributed differently across different Englishes simply because they happen to have synonyms with different etymologies, even though they

Transactions of the Association for Computational Linguistics, vol. 6, pp. 329–342, 2018. Action Editor: Ivan Titov.
Submission batch: 1/2018; Revision batch: 3/2018; Published 5/2018.

carry very limited cultural weight. Then, we show that simple lexical semantic features (based on the focus set of words) suffice for clustering together English texts authored by speakers of "closer" languages; we generate a phylogenetic tree of 31 languages solely by looking at lexical semantic properties of the English spoken by non-native speakers from 31 countries.

The contribution of this work is twofold. First, we introduce the *L2-Reddit corpus*: a large corpus of highly-advanced, fluent, diverse, non-native English, with sentence-level annotations of the native language of each author. Second, we lay out sound empirical foundations for the theoretical hypothesis on the cognate effect in L2 of non-native English speakers, highlighting the cognate facilitation phenomenon as one of the important factors shaping the language of non-native speakers.

After discussing related work in Section 2, we describe the L2-Reddit corpus in Section 3. Section 4 details the methodology we use and our results. We analyze these results in Section 5, and conclude with suggestions for future research.

2 Related Work

The language of bilinguals is different. The mutual presence of two linguistic systems in the mind of the bilingual speaker involves a significant cognitive load (Shlesinger, 2003; Hvelplund, 2014; Prior, 2014; Kroll et al., 2014); this burden is likely to have a bearing on the linguistic productions of the bilingual speaker. Moreover, the presence of more than one linguistic system gives rise to *transfer*: traces of one linguistic system may be observed in the other language (Jarvis and Pavlenko, 2008).

Several works addressed the translation choices of bilingual speakers, either within a rich linguistic context (e.g., given a source sentence), or decontextualized. For example, de Groot (1992) demonstrated that cognate translations are produced more rapidly and accurately than translations that do not exhibit phonetic or orthographic similarity with a source word. This observation was further articulated by Prior et al. (2007), who showed that translation choices of L2 speakers were positively correlated with cross-linguistic form overlap of a stimulus word with its target language translations. Prior et al. (2011) emphasized that "bilinguals are sensitive to the degree of form overlap between the translation equivalents in the two languages, and show a preference toward producing a cognate translation". As an example, they showed that the preferred translation of the Spanish *incidente* to English was *incident*, and not the alternative translation *event*, despite the much higher frequency of the latter.

More recent work is consistent with previous research and advances it by highlighting phonologically mediated cross-lingual influences on visual word processing of same- and different-script bilinguals (Degani and Tokowicz, 2010; Degani et al., 2017). Cognate facilitation was also studied using eye tracking (Libben and Titone, 2009; Cop et al., 2017), demonstrating that the reading of bilinguals is influenced by orthographic similarity of words with their translation equivalents in another language. Crucially, much of this research has been conducted in a laboratory experimental setup; this implies a small number of participants, a small number of target words, and focus on a very limited set of languages. While our research questions are similar, we present a computational analysis of the effects of cognates on L2 productions on a completely different scale: 31 languages, over 1000 words, and thousands of speakers whose spontaneous language production is recorded in a very large corpus.

Corpus-based investigation of non-native language has been a prolific field of recent research. Numerous studies address *syntactic* transfer effects on L2. Such influences from L1 facilitate various computational tasks, including automatic detection of highly competent non-native writers (Tomokiyo and Jones, 2001; Bergsma et al., 2012), identification of the mother tongue of English learners (Koppel et al., 2005; Tetreault et al., 2013; Tsvetkov et al., 2013; Malmasi et al., 2017) and typology-driven error prediction in learners' speech (Berzak et al., 2015). English texts produced by native speakers of a variety of languages have been used to reconstruct phylogenetic trees, with varying degrees of success (Nagata and Whittaker, 2013; Berzak et al., 2014). Syntactic preferences of professional translators were exploited to reconstruct the Indo-European language tree (Rabinovich et al., 2017). Our study is also corpus-based; but it stands out as it focuses not on the distribution of function words or (shallow) syntactic structures, but rather on the unique use of

cognates in L2.

From the *lexical* perspective, L2 writers have been shown to produce more overgeneralizations, use more frequent words and words with a lower degree of ambiguity (Hinkel, 2002; Crossley and McNamara, 2011). Several studies addressed cross-linguistic influences on semantic acquisition in L2, investigating the distribution of collocations (Siyanova-Chanturia, 2015; Kochmar and Shutova, 2017) and formulaic language (Paquot and Granger, 2012) in learner corpora. We, in contrast, address highly-fluent, advanced non-natives in this work.

Nastase and Strapparava (2017) presented the first attempt to leverage etymological information for the task of native language identification of English learners. They sowed the seeds for exploitation of etymological clues in the study of non-native language, but their results were very inconclusive.

In contrast to the learner corpora that dominate studies in this field (Granger, 2003; Geertzen et al., 2013; Blanchard et al., 2013), our corpus contains spontaneous productions of advanced, highly proficient non-native speakers, spanning over 80K topical threads, by 45K distinct users from 50 countries (with 46 native languages). To the best of our knowledge, this is the first attempt to computationally study the effect of L1 cognates on L2 lexical choice in productions of competent non-native English speakers, certainly at such a large scale.

3 The *L2-Reddit corpus*

One contribution of this work is the collection, organization and annotation of a large corpus of highly-fluent non-native English. We describe this new and unique corpus in this section.

3.1 Corpus mining

Reddit[1] is an online community-driven platform consisting of numerous forums for news aggregation, content rating, and discussions. As of 2017, it has over 200 million unique users, ranking the fourth most visited website in the US. Content entries are organized by areas of interest called *subreddits*,[2] ranging from main forums that receive much attention to smaller ones that foster discussion on niche areas. Subreddit topics include news, science, movies, books, music, fitness and many others.

Collection of author metadata We collected a large dataset of posts (both initial submissions and subsequent comments) using an API especially designed for providing search capabilities on Reddit content.[3] We focused on several subreddits (r/Europe, r/AskEurope, r/EuropeanCulture, r/EuropeanFederalists, r/Eurosceptics) whose content is generated by users who specified their country as a *flair* (metadata attribute). Although categorized as 'European', these subreddits are used by people from all over the world, expressing views on politics, legislation, economics, culture, etc.

In the absence of a restrictive policy, multiple flair alternatives often exist for the same country, e.g., 'CROA' and 'Croatia' for Croatia. Additionally, distinct flairs are sometimes used for regions, cities, or states of big European countries, e.g., 'Bavaria' for Germany. We (manually) grouped flairs representing the same country into a single cluster, reducing 489 distinct flairs into 50 countries, from Albania to Vietnam. The posts in the Europe-related subreddits constitute our *seed corpus*, comprising 9M sentences (160M tokens) by over 45K distinct users.

Dataset expansion A typical user activity in Reddit is not limited to a single thread, but rather spreads across multiple, not necessarily related, areas of interest. Once the authors' country is determined based on their European submissions, their entire Reddit footprint can be associated with their profile, and, therefore, with their country of origin. We extended our seed corpus by mining *all* submissions of users whose country flair is known, querying all Reddit data spanning years 2005-2017. The final dataset thus contains over 250M sentences (3.8B tokens) of native and non-native English speakers, where each sentence is annotated with its author's country of origin. The data covers posts by over 45K authors and spans over 80K subreddits.[4]

Focus on "large" languages For the sake of robustness, we limited the scope of this work to (coun-

[1] https://www.reddit.com/

[2] Subreddits are typically denoted with a leading r/, for example r/linguistics is the 'linguistics' subreddit.

[3] https://github.com/pushshift/api

[4] The annotated dataset will freely available at http://cl.haifa.ac.il/projects/L2. To protect the anonymity of Reddit users, the released dataset does not expose any author identifying information.

tries whose L1s are) the Indo-European (IE) languages; and only to those countries whose users had at least 500K sentences in the corpus. Additionally, we excluded multilingual countries, such as Belgium and Switzerland. Consequently, the final set of Reddit authors considered in this work originate from 31 countries, which represent the three main IE language families: *Germanic* (Austria, Denmark, Germany, Iceland, Netherlands, Norway, Sweden); *Romance* (France, Italy, Mexico, Portugal, Romania, Spain); and *Balto-Slavic* (Bosnia, Bulgaria, Croatia, Czech, Latvia, Lithuania, Poland, Russia, Serbia, Slovakia, Slovenia, Ukraine). In addition, we have data authored by native English speakers from Australia, Canada, Ireland, New Zealand, the UK and the US.

Correlation of country annotation with L1 We view the country information as an accurate, albeit not perfect, proxy for the native language of the author.[5] We acknowledge that the L1 information is noisy and may occasionally be inaccurate. We therefore evaluated the correlation of the country flair with L1 by means of supervised classification: our assumption is that if we can accurately distinguish among users from various countries using features that reflect language, rather than culture or content, then such a correlation indeed exists.

We assume that the native language of speakers "shines through" mainly in their syntactic choices. Consequently, we opted for (shallow) syntactic structures, realized by function words (FW) and n-grams of part-of-speech (POS) tags, rather than geographical and topical markers, that are reflected best by content words. Aiming to disentangle the effect of native language we randomly shuffled texts produced by all authors from each country, thereby "blurring out" any topical (i.e., subreddit-specific) or authorial trace. Consequently, we assume that the separability of texts by country can be attributed to the only distinguishing linguistic variable left: the dimension of the native language of a speaker.

We classified 200 chunks of 100 randomly sampled sentences from each country into (i) native vs. non-native English speakers, (ii) the three IE language families, and (iii) 45 individual L1s, where the six English-speaking countries are unified under the native-English umbrella. Using over 400 function words and top-300 most frequent POS-trigrams, we obtained 10-fold cross-validation accuracy of 90.8%, 82.5% and 60.8%, for the three scenarios, respectively. We conclude, therefore, that the country flair can be viewed as a plausible proxy for the native language of Reddit authors.

Initial preprocessing Several preprocessing steps were applied on the dataset. We (i) removed text by users who changed their country flair within their period of activity; (ii) excluded non-English sentences;[6] and (iii) eliminated sentences containing single non-alphabetic tokens. The final corpus comprises over 230M sentences and 3.5B tokens.

3.2 Evaluation of author proficiency

Unlike most corpora of non-native speakers, which focus on *learners* (e.g., ICLE (Granger, 2003), EF-CAMDAT (Geertzen et al., 2013), or the TOEFL dataset (Blanchard et al., 2013)), our corpus is unique in that it is composed by fluent, advanced non-native speakers of English. We verified that, on average, Reddit users possess excellent, near-native command of English by comparing three distinct populations: (i) Reddit native English authors, defined as those tagged for one of the English-speaking countries: Australia, Canada, Ireland, New Zealand, and the UK. We excluded texts produced by US authors due to the high ratio of the US immigrant population; (ii) Reddit non-native English authors; and (iii) A population of English learners, using the TOEFL dataset (Blanchard et al., 2013); here, the proficiency of authors is classified as low, intermediate, or high.

We compared these populations across various indices, assessing their proficiency with several commonly accepted lexical and syntactic complexity measures (Lu and Ai, 2015; Kyle and Crossley, 2015). Lexical richness was evaluated through type-to-token ratio (TTR), average age-of-acquisition (in years) of lexical items (Kuperman et al., 2012), and mean word rank, where the rank was retrieved from a list of the entire Reddit dataset vocabulary, sorted by word frequency in the corpus. Syntactic com-

[5] We therefore use the terms 'user country', 'native language' and 'L1' interchangeably henceforth.

[6] We used the *polyglot* language detection tool (`http://polyglot.readthedocs.io`).

plexity was assessed using mean length of T-units (TU; the minimal terminable unit of language that can be considered a grammatical sentence), and the ratio of complex T-units (those containing a dependent clause) to all T-units in a sentence.

Table 1 reports the results. Across almost all indices, the level of Reddit non-natives is much higher than even the advanced TOEFL learners, and almost on par with Reddit natives.

4 L1 cognate effects on L2 lexical choice

4.1 Hypotheses

Cognates are words in two languages that share both a similar meaning and a similar form. Our main hypothesis is that non-native speakers, when required to pick an English word that has a set of synonyms, are more likely to select a lexical item that has a cognate in their L1. We therefore expect the effect of L1 cognates to be reflected in the frequency of their English counterparts in the spontaneous productions of L2 speakers. Moreover, we expect similar effects, perhaps to a lesser extent, in the contextual usage of certain words, reflecting collocations and subtle contours of word meanings that are transferred from L1. The different contexts that certain words are embedded in (in the Englishes of speakers with different L1 backgrounds) can be captured by the means of distributional semantics.

Furthermore, we hypothesize that the effect of L1 is powerful to an extent that facilitates clustering of Englishes produced by non-natives with "similar" L1s; specifically, L1s that belong to the same language family. "Similar" L1s may reflect both typological and areal closeness: for example, we expect the English spoken by Romanians to be similar both to the English of Italians (as both are Romance languages) and to the English of Bulgarians (as both are Balkan languages). Ultimately, we aim to reconstruct the IE language phylogeny, reflecting historical and areal evolution of the subsets of Germanic, Romance and Balto-Slavic languages over thousands of years, from non-native English only.

While lexical transfer from L1 is a known phenomenon in *learner* language, we hypothesize that its signal is present also in the language of highly competent non-native speakers. Mastering the nuances of lexical choice, including subtle contours of word meaning and the correct context in which words tend to occur, are key factors in advanced language competence. The L2-Reddit corpus provides a perfect environment for testing this hypothesis.

4.2 Selection of a focus set of words

Our goal is to investigate non-native speakers' choice of lexical items in English. We address this task by defining a set of English words that have at least one synonym; ideally, we would like the various synonyms to have different etymologies, and in particular, to have different cognates in different language families. English happens to be a particularly good choice for this task, since in spite of its Germanic origins, much of its vocabulary evolved from Romance, as a great number of words were borrowed from Old French during the Norman occupation of Britain in the 11th century.

To trace the etymological history of English words we used Etymological WordNet (EW), a database that contains information about the ancestors of over 100K English words, about 25K of them in contemporary English (de Melo, 2014). For each word recorded in EW, the full path to its root can be reconstructed. Intuitively, an English word with Latin roots may exhibit higher (phonetic and orthographic) proximity to its Romance languages' counterparts. Conversely, an English word with a Proto-Germanic ancestor may better resemble its equivalents in Germanic languages.

We selected from EW all the nouns, verbs, and adjectives. For each such word w, we identified the synset of w in WordNet, choosing only the first (i.e., most prominent) sense of w (and, in particular, corresponding to the most frequent part-of-speech (POS) category of w in the L2-Reddit dataset). Then, we retained only those words that had synonyms, and only those whose synonyms had at least two different etymological paths, i.e., synonyms rooted in different ancestors. For example, we retained the synset {*heaven*, *paradise*}, since the former is derived from Proto-Germanic **himin-* , while the latter is derived from Greek παράδεισος (via Latin and Old French).

Furthermore, to capture the bias of non-native speakers toward their L1 cognate, it makes sense to focus on a set of easily interchangeable synonyms, e.g., {*divide*, *split*}. In contrast, consider an unbal-

Population	Mean TU length	Complex TU ratio	TTR	Mean word rank	AoA
Learners (low)	15.583	0.513	0.089	1172.19	5.186
Learners (medium)	16.357	0.534	0.106	1504.01	5.317
Learners (high)	17.468	0.528	0.124	1852.64	5.562
Reddit non-natives	*19.528*	*0.633*	*0.174*	*1960.62*	*5.524*
Reddit natives	20.154	0.658	0.179	2063.89	5.575

Table 1: Evaluation of the English proficiency of non-native Reddit users.

anced synset {*kiss*, *buss*, *osculation*}: presumably, the prevalent alternative *kiss* is likely to be used by all speakers, regardless of their native language. To eliminate such cases, we excluded synsets that were dominated by a single alternative (with a frequency of over 90% in our corpus), compared to other synonymous choices. Table 2 illustrates a few examples of synonym sets with their etymological origins.

Eliminating cultural bias Although our Reddit corpus spans over 80K topical threads and 45K users, posts produced by authors from neighboring countries may carry over markers with similar geographical or cultural flavor. For example, we may expect to encounter *soviet* more frequently in posts by Russians and Ukrainians, *wine* in texts of French or Italian authors, and *refugees* in posts by German users. While they may be typical to a certain population group, such terms are totally unrelated to the phenomenon we address here, and we therefore wish to eliminate them from the focus set of words.

A common way to identify elements that are statistically over-represented in a particular population, compared to another, is *log-odds ratio informative Dirichlet prior* (Monroe et al., 2008). We employed this approach to discover words that were overused by authors of a certain country, where posts from each country (a category under test) were compared to all the others (the background). We used the strict log-odds score of -5 as a threshold for filtering out terms associated with a certain country.[7] Among the terms eliminated by this procedure were *genocide* for Armenia, *hockey* for Canada and *independence* for the UK. The final focus set of words thus consists of neutral, ubiquitous sets of synonyms, varying in their etymological roots. It comprises 540 synonym sets and 1143 distinct words.

[7]The threshold was set by preliminary experiments, without any further tuning.

4.3 Model

We hypothesize (Section 4.1) that L1 effects on lexical choice are so powerful, even with advanced non-native speakers, that it is possible to reconstruct the IE language phylogeny, reflecting historical and areal evolution over thousands of years, from non-native English only. We now describe a simple yet effective framework for clustering the Englishes of authors with different L1s, integrating both word frequencies and semantic word representations of the words in our focus set (Section 4.2).

4.3.1 Data cleanup and abstraction

Aiming to learn word representations for the lexical items in our focus set, we want the contextual information to be as free as possible from strong geographical and cultural cues. We therefore process the corpus further. First, we identified named entities (NEs) and systematically replaced them by their type. We used the implementation available in the *spacy* Python package,[8] which supports a wide range of entities (e.g., names of people, nationalities, countries, products, events, book titles, etc.), at state-of-the-art accuracy. Like other web-based user generated content, the Reddit corpus does not adhere to strict casing rules, which has detrimental effects on the accuracy of NE identification. To improve the tagging accuracy, we applied a preprocessing step of 'truecasing', where each token w was assigned the case (lower, upper, or upper-initial) that maximized the likelihood of the consecutive tri-gram $\langle w_{pre},\ w,\ w_{post}\rangle$ in the Corpus of Contemporary American English (COCA).[9] For example, the tri-gram 'the us people' was converted to 'the US people', but 'let us know' remained unchanged. When a tri-gram was not found in the COCA n-gram corpus,

[8]`https://spacy.io`

[9]`https://www.ngrams.info`

Synonym set	Etymological path to root
cargo (N)	Spanish: *cargo* ← Spanish: *cargar* ← Late Latin: *carricare*
freight (N)	Mid. English: *freyght* ← Mid. Low German: *vrecht* ← Proto-Germanic **fra-* + **aihtiz*
weary (Adj)	Mid. English: *wery* ← Old English: *wērigġ* ← Proto-Germanic: **wōrīgaz*
fatigue (Adj)	French: *fatigue* ← French: *fatiguer* ← Latin: *fatigare*
exaggerate (V)	Latin: *exaggerare* ← Latin: *ex-* + Latin: *aggerare*
overdo (V)	English: *over* + *do*

Table 2: Etymological roots of example synonym sets with corresponding part-of-speech.

we employed fallback to unigram probability estimation. Additionally, we replaced all non-English words with the token 'UNK'; and all web links, subreddit (e.g., r/compling) and user (u/userid) pointers with the 'URL' token.[10]

4.3.2 Distance estimation and clustering

Bamman et al. (2014) introduced a model for incorporating contextual information (such as geography) in learning vector representations. They proposed a joint model for learning word representations in a situated language, a model that "includes information about a subject (i.e., the speaker), allowing to learn the contours of a word's meaning that are shaped by the context in which it is uttered". Using a large corpus of tweets, their joint model learned word representations that were sensitive to geographical factors, demonstrating that the usage of *wicked* in the United States (meaning *bad* or *evil*) differs from that in New England, where it is used as an adverbial intensifier (*my boy's wicked smart*).

We leveraged this model to uncover linguistic variation grounded in the different L1 backgrounds of non-native Reddit speakers. We used equal-sized random samples of 500K sentences from each country to train a model of vector representations. The model comprises representation of every vocabulary item in each of the 31 Englishes; e.g., 31 vectors are generated for the word *fatigue*, presumably reflecting the subtle divergences of word semantics, rooted in the various L1 backgrounds of the authors.

In order to cluster together Englishes of speakers with "similar" L1s, we need a measure of distance between two English texts. This measure is based on two constituents: word frequencies and word embeddings. Given two English texts originating from different countries, we computed for each word w in our focus set (i) the difference in the frequency of w in the two texts; and (ii) the distance between the vector representations of w in these texts, estimated by cosine similarity of the two corresponding word vectors. We employed the popular *weighted product model* to integrate the two arguments. The word vector component was assigned a higher weight as the frequency of w in the collection increases; this is motivated by the intuition that learning the semantic relationships of a word benefits from vast usage examples. We therefore weigh the embedding constituent proportionally to the word's frequency in the dataset, and assign the complementary weight to the difference of frequencies.

Formally, given two English texts E_{L_i} and E_{L_j}, with L_i and L_j native languages, and given a word w in the focus set, let f_i and f_j denote the frequencies of w in E_{L_i} and E_{L_j}, respectively. Let p_w be the probability of w in the entire collection. We further denote the vector space representation of w in E_{L_i} by v_i, and the representation of w in E_{L_j} by v_j. Then, the distance between E_{L_i} and E_{L_j} with respect to the word w is:

$$D_{ij}(w) = (|f_i - f_j|)^{1-p_w} \times (1 - cos(v_i, v_j))^{p_w}. \quad (1)$$

The final distance between E_{L_i} and E_{L_j} is given by averaging D_{ij} over all words in the focus set FS:

$$D_{ij} = \frac{(\sum_{w \in FS} D_{ij}(w))}{|FS|}.$$

Finally, we constructed a symmetric distance matrix (31×31) M by setting $M[i, j] = D_{ij}$. We used

[10]The cleaned, abstracted subset of the corpus is also available at `http://cl.haifa.ac.il/projects/L2`. The cleanup code is available at `https://github.com/ellarabi/reddit-l2`.

Ward's hierarchical clustering[11] with the Euclidean distance metric to derive a tree from the distance matrix M.

We considered several other weighting alternatives, including assignment of constant weights to the two factors in Equation 1; they all resulted in inferior outcomes. We also corroborated the relative contribution of the two components by using each of them alone. While considering only frequencies resulted in a slightly inferior outcome (see Section 4.5), using word representations alone produced a completely arbitrary result.

4.4 Results

The resulting tree is depicted in Figure 1. The reconstructed language typology reveals several interesting observations. First, and much expectedly, all native English speakers are grouped together into a single, distant sub-tree, implying that similarities exhibited by the lexical choices of native speakers go beyond geographical and cultural differences. The Englishes of non-native speakers are clustered into three main language families: Germanic, Romance, and Balto-Slavic. Notably, Spanish-speaking Mexico is clustered with its Romance counterparts. The firm Balto-Slavic cluster reveals historical relations between languages by generating coherent sub-branches: the Czech Republic and Slovakia, Latvia and Lithuania, as well as the relative proximity of Serbia and Croatia. In fact, former Yugoslavia is clustered together, except for Bosnia, which is somewhat detached. Similar close ties can be seen between Austria and Germany, and between Portugal and Spain.

Another interesting phenomenon is captured by English texts of authors from Romania: their language is assigned to the Balto-Slavic family, implying that the deep-rooted areal and cultural Balkan influences left their traces in the Romanian language, which in turn, is reflected in the English productions of native Romanian authors. Unfortunately, we cannot explain the location of Iceland.

A geographical view mirroring the language phylogeny is presented in Figure 3. Flat clusters were obtained from the hierarchy using the *scipy fcluster* method[12] with defaults.

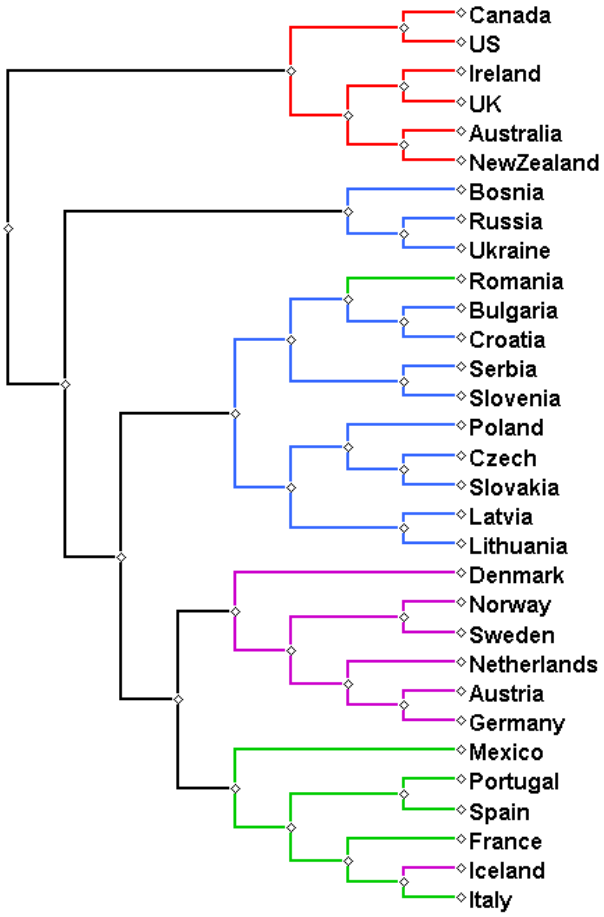

Figure 1: Language typology reconstructed from non-native Englishes using features reflecting lexical choice. Countries that belong to the same phylogenetic family (according to the gold tree) share identical color. E.g., Iceland is colored purple, like other Germanic languages, even though it is assigned to the Romance cluster.

This outcome, obtained using only lexical semantic properties (word frequencies and word embeddings) of English authored by various non-native speakers, is a strong indication of the power of L1 influence on L2 speakers, even highly fluent ones. These results are strongly dependent on the choice of focus words: we carefully selected words that on one hand lack any cultural or geographical bias toward one group of non-natives, but on the other hand have synonyms with different etymologies. As

[11] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html

an additional validation step, we generated a language tree using exactly the same methodology but a different set of focus words. We randomly sampled 1143 words from the corpus, controlling for country-specific bias but *not* for the existence of synonyms with different etymologies. Although some of the intra-family ties were captured (in particular, all native speakers were clustered together), the resulting tree (Figure 2) is far inferior.

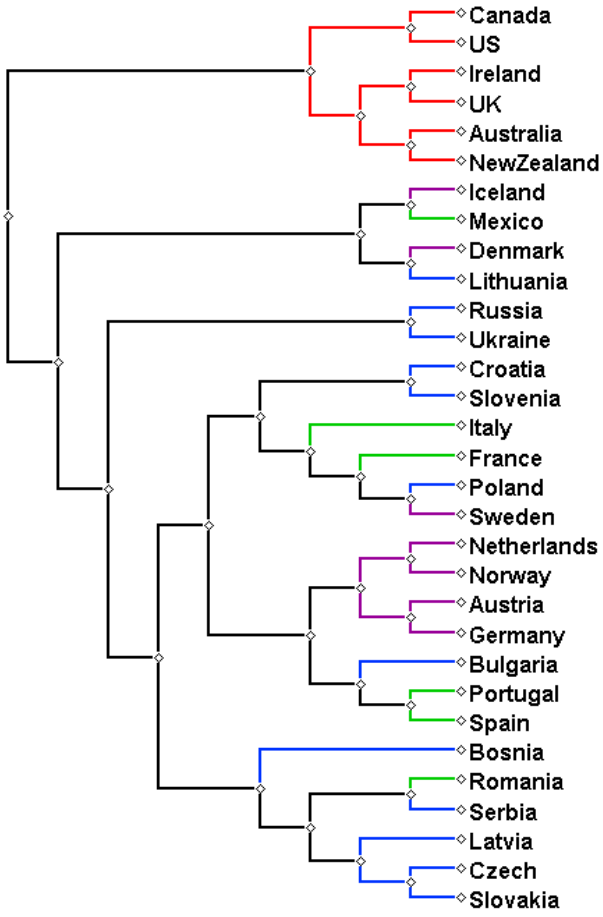

Figure 2: Language typology reconstructed from a randomly selected focus set of 1143 words.

We also conducted an additional experiment, including multilingual Belgium and Switzerland in the set of countries. While the L1 of speakers cannot be determined for these two countries, presumably Belgium is dominated by Dutch and French, and Switzerland by German and French. Indeed, both countries were assigned into the Germanic language family in our clustering experiments.

4.5 Evaluation

To better assess the quality of the reconstructed trees we now provide a quantitative evaluation of the language typologies obtained by the various experiments. We adopt the evaluation approach of Rabinovich et al. (2017), who introduced a distance metric between two trees, defined as the sum of the square differences between all leaf-pair distances in the two trees. More specifically, given a tree of N leaves, l_i, $i \in [1..N]$, the distance between two leaves l_i, l_j in a tree τ, denoted $D_\tau(l_i, l_j)$, is defined as the length of the shortest path between l_i and l_j. The distance $Dist(\tau, g)$ between a generated tree τ and the gold tree g is then calculated by summing the square differences between all leaf-pair distances in the two trees:

$$Dist(\tau, g) = \sum_{i,j \in [1..N]; i \neq j} (D_\tau(l_i, l_j) - D_g(l_i, l_j))^2.$$

We used the Indo-European tree in *Glottolog*[13] as our gold standard, pruning it to contain the set of 31 languages considered in this work. For the sake of comparison, we also present the distance obtained for a completely random tree, generated by sampling a random distance matrix from the uniform (0, 1) distribution. The reported random tree evaluation score is averaged over 100 experiments.

Table 3 presents the results. All distances are normalized to a zero-one scale, where the bounds, zero and one, represent the identical and the most distant tree with respect to the gold standard, respectively. Much expectedly, the random tree is the worst one, followed closely by the tree reconstructed from a random sample of over 1000 words sampled from the corpus (Figure 2). The best result is obtained by considering both word frequencies and representations, being only slightly superior to the tree reconstructed using word frequencies alone. The latter result corroborates the aforementioned observation (Section 4.3.2) and further posits word frequencies as the major factor affecting the shape of the obtained phylogeny.

[13] http://glottolog.org/

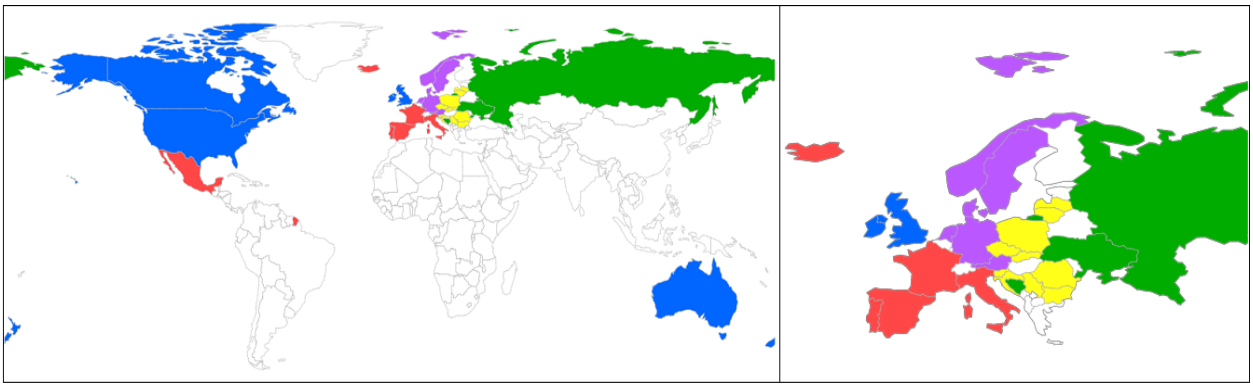

Figure 3: Countries by clusters: World (on the left) and Europe (on the right) views. Countries assigned to the same flat cluster by the *clustering procedure* (Section 4.4) share identical color, e.g., the wrongly assigned Iceland shares the red color with the Romance-language speaking countries. Countries not included in this work are uncolored.

Features used	Distance
Random tree	1.000
Randomly sampled words (Figure 2)	0.857
Focus set with frequencies only	0.497
+ embeddings (Figure 1)	**0.469**

Table 3: Normalized distance between a reconstructed and the gold tree; lower distances indicate better result.

5 Analysis

The results described in Section 4.4 empirically support the intuition that cognates are one of the factors that shape lexical choice in productions of non-native authors. In this section we perform a closer analysis of the data, aiming to capture the subtle yet systematic distortions that help distinguish between English texts of speakers with different L1s.

Quantitative analysis Given a synonym set $s \in FS$, consisting of words $\langle w_1, w_2, ..., w_n \rangle$, and two English texts with two different L1s, E_{L_i} and E_{L_j}, we computed the counts of the synset words in these texts, and further normalized the counts by the total sum, yielding probabilities. We denote the probability distribution of a synset $s = \langle w_1, w_2, ..., w_n \rangle$ in E_{L_i} by:

$$P_i^s = \langle p_i(w_1), p_i(w_2), ..., p_i(w_n) \rangle.$$

The different usage patterns of a synonym set s across two Englishes can then be estimated using the Jensen-Shannon divergence (JSD) between the two probability distributions:

$$div_{ij}(s) = JSD(P_i^s, P_j^s). \quad (2)$$

We expect that "close" L1s will have lower divergence, whereas L1s from different language families will exhibit higher divergences.

Table 4 presents the top twenty synonym sets for the arbitrarily chosen Germany–Spain country pair, ranked by divergence (Equation 2). The overuse of *hinder* by German authors may be attributed to its German *behindern* cognate, whereas Spanish users' preference of *impede* is probably attributable to its Spanish *impedir* equivalent. A Spanish cognate for *plantation*, *plantación*, possibly explains the clear preference of Spanish native speakers for this alternative, compared to the more popular choice of German authors, *grove*, which has Germanic etymological origins.

The {*weariness*, *tiredness*, *fatigue*} synset reveals the preference of Spanish native speakers for *fatigue*, whose Spanish equivalent *fatiga* resembles it to a great extent; *weariness*, however, is slightly more frequent in the texts of German speakers, potentially reflecting its Proto-Germanic **wōrīgaz* ancestor. An interesting phenomenon is revealed by the synset {*conceivable*, *imaginable*}: while both words have Latin origins, *imaginable* is more ubiquitous in the English language, rendering it more frequent in texts of German native speakers, compared to the more balanced choice of Spanish authors. Usage patterns in {*overdo*, *exaggerate*} and {*inspect*, *audit*, *scrutinize*} can be attributed to the same phe-

Synonym set s	$P^s_{Germany}$	P^s_{Spain}
⟨hinder impede⟩	(0.909, 0.091)	(0.69, 0.31)
⟨grove orchard plantation⟩	(0.643, 0.214, 0.143)	(0.227, 0.068, 0.705)
⟨weariness tiredness fatigue⟩	(0.167, 0.208, 0.625)	(0.017, 0.119, 0.864)
⟨yarn recital narration⟩	(0.55, 0.1, 0.35)	(0.22, 0.15, 0.63)
⟨bloom blossom flower⟩	(0.25, 0.143, 0.607)	(0.085, 0.098, 0.817)
⟨conceivable imaginable⟩	(0.22, 0.78)	(0.415, 0.585)
⟨overdo exaggerate⟩	(0.556, 0.444)	(0.319, 0.681)
⟨inspect audit scrutinize ⟩	(0.667, 0.25, 0.083)	(0.446, 0.429, 0.125)
⟨sharp acute⟩	(0.886, 0.114)	(0.717, 0.283)
⟨steady stiff unwavering firm⟩	(0.364, 0.172, 0.017, 0.447)	(0.278, 0.083, 0.007, 0.632)
⟨ecstasy rapture⟩	(0.593, 0.407)	(0.412, 0.588)
⟨sizeable ample⟩	(0.597, 0.403)	(0.429, 0.571)
⟨scummy abject miserable⟩	(0.167, 0.028, 0.806)	(0.067, 0.053, 0.88)
⟨drift displace⟩	(0.835, 0.165)	(0.734, 0.266)
⟨waive abandon forego⟩	(0.095, 0.845, 0.061)	(0.043, 0.899, 0.058)
⟨weigh consider count⟩	(0.028, 0.605, 0.367)	(0.024, 0.582, 0.394)
⟨quick fast rapid⟩	(0.328, 0.649, 0.024)	(0.326, 0.643, 0.031)
⟨stumble stagger lurch⟩	(0.889, 0.097, 0.014)	(0.7, 0.114, 0.186)
⟨omen presage⟩	(1.0, 0.0)	(0.9, 0.1)
⟨freight cargo⟩	(0.215, 0.785)	(0.19, 0.81)

Table 4: Top-20 examples of the most divergent usage patterns of synsets in texts of German vs. Spanish authors. Words with (recorded) Germanic origins are in blue and words with (recorded) Latin origins are in red.

nomenon, where the German equivalent for *inspect* (*inspizieren*) resembles its English counterpart despite a different etymological root.

Usage examples Table 5 presents example sentences written by Reddit authors with French and Italian L1s, further illustrating discrepancies in lexical choice (presumably) stemming from cognate facilitation effects. The French *rapide* is a translation equivalent of the English synset {*rapid, quick, fast*}, but its English *rapid* cognate is more constrained to contexts of movement or growth, rendering the collocation *rapid check* somewhat marked. The French noun *approbation* is more frequent in contemporary French than its English (practically unused) equivalent *approbation*; this makes its use in English sound unnatural. In our Reddit corpus, *approbation* appears 48 times in L1-French texts, compared to 5, 4, and 4 in equal-sized texts by authors from the UK, Ireland and Canada, respectively. One of the frequent English synonym alternatives {*approval, acceptance*} would better fit this context. Finally, while the Italian expression *sera precedente* is common, its English equivalent *precedent evening* is very infrequent, yet it is used in English productions of Italian speakers.

6 Conclusion

We presented an investigation of L1 cognate effects on the productions of advanced non-native Reddit authors. The results are accompanied by a large dataset of native and non-native English speakers, annotated for author country (and, presumably, also L1) at the sentence level.

Several open questions remain for future research. From a theoretical perspective, we would like to extend this work by studying whether the tendency to choose an English cognate is more powerful in L1s with both phonetic and orthographic similarity to English (Roman script) than in L1s with phonetic similarity only (e.g., Cyrillic script). We also plan to more carefully investigate productions of speakers from multilingual countries, like Belgium and Switzerland. Another extension of this work may broaden the analysis to include additional language families.

L1	Sentence
French	*I have to go to the Dr. to do a **rapid** check on my heart stability.*
French	*Maybe put every name through a manual **approbation** pipeline so it ensures quality.*
French	*Polls have shown public **approbation** for this law is somewhere between 58% and 65%, and it has been a strong promise during the presidential campaign.*
Italian	*The event was even more shocking because the **precedent** evening he wasn't sick at all.*

Table 5: Cognate facilitation phenomena in usage examples by Reddit authors.

There are also various potential practical applications to this work. First, we plan to exploit the potential benefits of our findings to the task of native language identification of (highly advanced) non-native authors, in various domains. Second, our results will be instrumental for personalization of language learning applications, based on the L1 background of the learner. For example, error correction systems can be enhanced with the native language of the author to offer root cause analysis of subtle discrepancies in the usage of lexical items, considering both their frequencies and context. Given the L1 of the target audience, lexical simplification systems can also benefit from cognate cues, e.g., by providing an informed choice of potentially challenging candidates for substitution with a simplified alternative. We leave such applications for future research.

Acknowledgments

This work was partially supported by the National Science Foundation through award IIS-1526745. We would like to thank Anat Prior and Steffen Eger for valuable suggestions. We are also grateful to Sivan Rabinovich for much advise and helpful comments. Finally, we are thankful to our action editor, Ivan Titov, and three anonymous reviewers for their constructive feedback.

References

David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-2134`.

Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337. Association for Computational Linguistics, 2012.

Yevgeni Berzak, Roi Reichart, and Boris Katz. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 21–29, June 2014. URL `http://aclweb.org/anthology/W/W14/W14-1603.pdf`.

Yevgeni Berzak, Roi Reichart, and Boris Katz. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 94–102, July 2015. URL `http://aclweb.org/anthology/K/K15/K15-1010.pdf`.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013 (2):i–15, 2013.

Uschi Cop, Nicolas Dirix, Eva Van Assche, Denis Drieghe, and Wouter Duyck. Reading a book in one or two languages? An eye movement study of cognate facilitation in L1 and L2 reading. *Bilingualism: Language and Cognition*, 20(4):747–769, 2017.

Scott A. Crossley and Danielle S. McNamara. Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, 20(4):271–285, 12 2011. ISSN 1060-3743. doi: 10.1016/j.jslw.2011.05.007.

Annette M. de Groot. Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5):1001, 1992.

Gerard de Melo. Etymological WordNet: Tracing the history of words. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Paris, France, 2014. ELRA.

Tamar Degani and Natasha Tokowicz. Semantic ambiguity within and across languages: An integrative review. *The Quarterly Journal of Experimental Psychology*, 63(7):1266–1303, 2010.

Tamar Degani, Anat Prior, and Walaa Hajajra. Cross-language semantic influences in different script bilinguals. *Bilingualism: Language and Cognition*, pages 1–23, 2017.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*, Somerville, MA, 2013. Cascadilla Proceedings Project.

Sylviane Granger. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *Tesol Quarterly*, pages 538–546, 2003.

Eli Hinkel. *Second language writers' text: Linguistic and rhetorical features*. Routledge, 2002.

Kristian Tangsgaard Hvelplund. Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. *MonTI. Monografías de Traducción e Interpretación*, pages 201–223, 2014.

Scott Jarvis and Aneta Pavlenko. *Crosslinguistic influence in language and cognition*. Routledge, 2008.

Ekaterina Kochmar and Ekaterina Shutova. Modelling semantic acquisition in second language learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 293–302, 2017.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628. ACM, 2005.

Judith F. Kroll, Susan C. Bobb, and Noriko Hoshino. Two languages in mind: Bilingualism as a tool to investigate language, cognition, and the brain. *Current Directions in Psychological Science*, 23(3):159–163, Jun 2014. doi: 10.1177/0963721414528511.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990, Dec 2012. ISSN 1554-3528. doi: 10.3758/s13428-012-0210-4. URL https://doi.org/10.3758/s13428-012-0210-4.

Kristopher Kyle and Scott A. Crossley. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786, 2015.

Maya R. Libben and Debra A. Titone. Bilingual lexical access in context: Evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2):381, 2009.

Xiaofei Lu and Haiyang Ai. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29, 2015. ISSN 1060-3743. doi: https://doi.org/10.1016/j.jslw.2015.06.003. URL http://www.sciencedirect.com/science/article/pii/S1060374315000405.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, 2017.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

Ryo Nagata and Edward W. D. Whittaker. Reconstructing an Indo-European family tree from non-native English texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1137–1147, August 2013. URL http://aclweb.org/anthology/P/P13/P13-1112.pdf.

Vivi Nastase and Carlo Strapparava. Word etymology as native language interference. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2702–2707. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/D17-1286.

Magali Paquot and Sylviane Granger. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32:130–149, 2012.

Anat Prior. Bilingualism: Interactions between languages. In Patricia J. Brook and Vera Kempe, editors, *Encyclopedia of Language Development*. Sage Publications, 2014. URL http://dx.doi.org/10.4135/9781483346441.

Anat Prior, Brian MacWhinney, and Judith F. Kroll. Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, 39(4):1029–1038, 2007.

Anat Prior, Shuly Wintner, Brian Macwhinney, and Alon Lavie. Translation ambiguity in and out of context. *Applied Psycholinguistics*, 32(1):93–111, 2011.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 530–540. Association for Computational Linguistics, July 2017. URL http://aclweb.org/anthology/P17-1049.

Miriam Shlesinger. Effects of presentation rate on working memory in simultaneous interpreting. *The Interpreters' Newsletter*, 12:37–49, 2003. URL http://hdl.handle.net/10077/2470.

Anna Siyanova-Chanturia. Collocation in beginner learner writing: A longitudinal study. *System*, 53:148–160, 2015.

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

Laura Mayfield Tomokiyo and Rosie Jones. You're not from 'round here, are you?: Naive Bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 2001.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. Identifying the L1 of non-native writers: the CMU-Haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 279–287. Association for Computational Linguistics, June 2013. URL http://www.aclweb.org/anthology/W13-1736.