Does Proficiency impact the use of cognates in L2?
A Computational Approach

Liat Nativ

1 Introduction

The language of non-native speakers, even highly advanced ones, is different from that of native speakers. One source of such differences is the influence of the speaker's first language (L1) on his or her second language (L2); this is referred to as *transfer effects*. Such transfer is evident, inter alia, in the lexical choices made by non-native speakers. Our hypothesis in this work is that transfer effects, as reflected by lexical choice, are correlated with the speaker's L2 proficiency. We will employ corpus-based computational methods to investigate this hypothesis.

While the main expected contribution of this work is theoretical, there is also practical motivation for such investigation. By gaining insights into the differences in the language produced by non-native speakers and their correlation with L2 fluency, we can contribute to creating better personalized, L1-based, natural language processing (NLP) solutions. A motivation for such personalization can be found, for example, in the work of Heilman et al. (2007), who suggest that since the process of language acquisition inherently differs for first and second-language learners (e.g., grammatical structures are acquired in a much slower rate by second language learners), readability assessment should be approached differently for these two populations. For similar reasons, the task of *text simplification*, both lexical and grammatical, must be solved differently for native and non-native speakers, most probably taking into account also the speaker's L1. However, to the best of our knowledge, this task is currently addressed uniformly, independently of the target audience. Summing up, understanding of non-native speakers' lexical choices, combined with the ability to assess their fluency level accordingly, is fundamental for creating "personalized" NLP applications, based on the users' L1 background.

2 Research Goal

Cognates are words in different languages that have similar forms and similar meanings due to a common ancestor in some protolanguage. Psycholinguistic research has shown that non-native speakers tend to overuse words that have cognates in their L1. This tendency has also been established recently on a large scale in a corpus-based computational work (Rabinovich et al., 2018). Our main goal in this work is to show a correlation between a non-native author's choice of cognates and his or her proficiency level in L2 (here, always English). We hypothesize that higher proficiency authors will tend to be less influenced by their L1, and therefore use cognates in a manner more similar to that of native speakers. Usually, this type of study is conducted with a small group of human participants, on which detailed information is collected. Such psycholinguistic studies are usually focused on a small set of words, and participants are requested to make lexical choices in response to a ready-made questioning (Prior et al., 2006, 2013). We aim to investigate how lexical choices reflected by the use of cognates are correlated to L2 fluency on a larger scale, focusing on dozens of cognates that occur in the spontaneous language of a large group (hundreds) of non-native authors with a variety of L1s.

3 Previous Work

Most people in the world are able to express themselves in more than one language (Grosjean and Li, 2013), thereby maintaining two or more language systems simultaneously—a task that requires considerable cognitive resources (Shlesinger, 2003; Hvelplund, 2014; Prior, 2014; Kroll et al., 2014). Traces of the mother tongue are likely to be found in the non-native speaker's second language utterances (Jarvis and Pavlenko, 2008). The differences between native and non-native speakers are so prominent that even highly advanced non-native speakers can be accurately distinguished from natives (Tomokiyo and Jones, 2001; Bergsma et al., 2012; Rabinovich et al., 2016; Goldin et al., 2018). Focusing on lexical choices, several authors have shown the preference of bilingual speakers towards cognates, for example in tasks of translation and via eye tracking while reading (de Groot, 1992; Prior et al 2011;Libben and Titone, 2009; Cop et al., 2017). *fix refs* Rabinovich et al. (2018) were able to reconstruct the phylogenetic language tree of the Indo-European language family, based solely on the tendency of non-natives to favor cognates in their native language.

The focus of this work is the correlation between L2 proficiency and cognate facilitation. Proficiency levels are estimated based on studies dealing with lexical, psycholinguistic and syntactic measures (Kuperman et al., 2012; Lu and Ai, 2015; Kyle and Crossley, 2015).

Following the psycholinguistic nature of our hypothesis, most related studies are conducted in an artificial environment, based on a limited number of participants, native languages, and target words. The setting for our work is different: we are using computational analysis to conduct a corpus-based study based on spontaneous (written) language productions by over 1600 speakers and 700 target words. Several corpus-based studies focused on investigating second language from various perspectives and with different tasks in mind (Tomokiyo and Jones, 2001; Bergsma et al., 2012; Koppel et al.,2005; Tetreault et al.,2013; Tsvetkov et al., 2013; Malmasi et al., 2017 ;Berzak et al., 2015 - Need to figure out what to mention/omit here). *fix refs Unclear how these works are related to yours: this section is intended to* situate *your work in the context of existing research.* Our work is inspired to a great extent by Rabinovich et al. (2018), and is based on the L2-Reddit corpus of highly-fluent, advanced non-natives, that has been released as part of it. *This is repeated (and belongs) in the following section*

4 Research Plan

4.1 Dataset

The dataset for this work is based on a large corpus of non-native English: the *L2-Reddit corpus*, released by Rabinovich et al. (2018). It comprises of social media posts by highly fluent authors who indicate their country as a metadata attribute. We view the country information as an accurate, albeit not perfect, proxy for the native language of the author (Goldin et al., 2018). Rabinovich et al. (2018) carefully created a culture-independent focus set of over 1000 words, forming 541 synonym sets that may reflect cognates in some L1s, but not all of them, and are hence used differently by authors with different linguistic backgrounds. We further reduced the size of the focus set to about 740 words forming 288 synonym sets, by investigating the etymology of each cognate, leaving only synonym sets that include at least one word of Germanic origin, and at least one of Romance Origin. The final dataset for this work was extracted from the L2-Reddit corpus and consists of 2168 non-native authors whose native language is either Romance or Germanic, who used cognates from the focus set at least 1000 times in their posts. The non-native group consists of two subgroups: authors whose L1 is Germanic (1627 users) and those whose L1 is Romance (541 users). A control group of approximately the same size of English native authors from 5 different countries (US, UK, New-Zealand, Australia and Ireland), with a similar distribution of cognates from the focus set as the non-native group, was extracted to complete the picture.

4.2 English proficiency measures

We use several commonly accepted measures for assessing the proficiency level of authors. These include both lexical and syntactic measures.

4.2.1 Lexical and psycholinguistic measures

Lexical measures include lexical richness, defined as type-token ratio (TTR); average age-of-acquisition (in years) of lexical items (Kuperman et al., 2012); *fix ref* 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 (Rabinovich et al., 2018). In addition, we also used the psycholinguistic measure of mean naming reaction time (in Milliseconds), based on a list generated from the English Lexicon Project (Balota et al. 2007). *fix ref* These measures were calculated on a random sample of 10000 tokens for each user.

4.2.2 Syntactic Measures

Syntactic complexity was assessed through 14 different syntactic measures, using the *L2 Syntactic Complexity Analyzer* (Lu 2010): *fix ref; leave only the ones you end up using*

MLS mean length of sentence (# of words/# of sentences),

MLT mean length of T-unit (# of words/# of T-units). T-unit is the minimal terminable unit of language that can be considered a grammatical sentence

MLC mean length of clause (# of words/# of clauses)

C/T clauses per T-unit (# of clauses/# of T-unit)

CT/T complex T-unit (containing a dependent clause) to all T-units ratio (# of complex T-units/# of T-units)

DC/C dependent clauses per clause

DC/T dependent clauses per T-unit (# of dependent clauses/# of clauses)

CP/C coordinate phrases per clause (# of coordinate phrases/# of clauses)

CP/T coordinate phrases per T-unit (# of dependent clauses/# of T-units)

T/S T-units per sentence (# of T-units/# of sentences)

CN/C complex nominals per clause (# of complex nominals/# of clauses)

CN/T complex nominals per T-unit (# of complex nominals/# of T-units)

VP/T verb phrases per T-unit (# of verb phrases/# of T-units)

C/S clauses per sentence (# of clauses/# of sentences)

These 14 measures are calculated on a random sample of 1000 sentences for each native and non-native author. [This list will be shorter based on the measures we'll decide that are meaningful]

4.3 Methodology

This is still too vague. Recall the research hypothesis and try to first explain, in simple non-technical terms, how your planned work is going to address and answer it. In particular, say something about Romance vs. Germanic, it is not at all clear from the current presentation.

To model the use of cognates and to examine the tendency towards Germanic or Romance cognates, we computed a normalized count for each user and for each synonym set. For example, if for a synonym set with one Germanic and one Romance word the count for a specific user was 15 for the Germanic word and 60 for the Romance word, than the normalized count will be 0.2 and 0.8 for Germanic and Romance cognates, respectively. For exploring the correlation between the use of cognates and the level of English proficiency, a benchmark of "good" English will be calculated based on the lexical selections made by the native authors group. The benchmark is defined as the average of the Germanic/Romance tendency—represented by the normalized count—across all the native authors. This benchmark is calculated for each synonym set independently. Once the benchmark is set for a certain synonym set, the lexical choice of each non-native author can be represented by the directed (-/+) distance from that benchmark. The direction is defined arbitrarily to be positive for Germanic and negative for Romance over-representation, respectively. The deviation from the benchmark as a function of the user proficiency level will provide an indication of the validity of the research hypothesis.

5 Preliminary Results

[Draw a few graphs for some of the synsets, based on the selected proficiency measures]

References

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.

Gili Goldin, Ella Rabinovich, and Shuly Wintner. Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/D18-1395`.

François Grosjean and Ping Li. *The Psycholinguistics of Bilingualism*. Wiley-Blackwell, 2013. ISBN 9781118349786.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, Rochester, New York, April 2007. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N07-1058`.

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.

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.

Xiaofei Lu and Haiyang Ai. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29:16–27, September 2015.

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

Anat Prior, Brian MacWhinney, and Judith F. Kroll. Does "querer" translate as "to love" or "to want"? effects of lexical properties and bilingual experience in negotiating translation ambiguity. Poster presented at the Fifth International Conference on the Mental Lexicon, October 2006.

Anat Prior, Judith F. Kroll, and Brian MacWhinney. Translation ambiguity but not word class predicts translation performance. *Bilingualism: Language and Cognition*, 16:458–474, 4 2013. ISSN 1469-1841. doi: 10.1017/S1366728912000272.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 1870–1881, August 2016. URL `http://aclweb.org/anthology/P/P16/P16-1176.pdf`.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342, 2018. ISSN 2307-387X. URL `https://transacl.org/ojs/index.php/tacl/article/view/1403`.

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`.

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.