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

Liat Nativ Advisors: Prof. Shuly Wintner and Dr. Anat Prior

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 the current 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 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 much larger scale, investigating dozens of cognates that occur in the spontaneous language of 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). 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 our work is the correlation between L2 proficiency and cognate facilitation: we expect to observe a lesser transfer effect as the level of proficiency increases. One motivation for the hypothesis can be found in the work of Prior et al. (2017), who showed reduced cross language interference with higher vocabulary. We estimate proficiency levels based on studies dealing with lexical, psycholinguistic and syntactic measures (Kuperman et al., 2012; Lu and Ai, 2015; Kyle and Crossley, 2015).

Due to 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. In contrast, we will use computational analysis to conduct a corpus-based study based on spontaneous (written) language productions by over 1600 speakers and 700 target words. Our work is inspired to a great extent by Rabinovich et al. (2018).

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 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 960 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 of equal size: authors whose L1 is Germanic and those whose L1 is Romance. 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) 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 examined the psycholinguistic measure of mean naming reaction time (in Milliseconds), based on a list generated from the English Lexicon Project (Balota et al., 2007). These measures were calculated on a random sample of 10000 tokens for each author. As L2-Reddit is based on highly fluent authors, similar scores were obtained for natives and non-natives for most of the measures. However, TTR showed statistically significant differences ($p < 0.05$) between the two populations and will be used for analyzing the results, as described in Section 4.3.

4.2.2 Syntactic Measures

Syntactic complexity was assessed through five different syntactic measures, using the *L2 Syntactic Complexity Analyzer* (Lu, 2010):

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

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

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

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

These 5 measures are calculated on a random sample of 1000 sentences for each native and non-native author, and the differences in scores between natives and non natives were found to be statistically significant ($p < 0.05$).

4.3 Methodology

As shown by Rabinovich et al. (2018), words that have cognates in the non-native speaker's L1 tend to be overrepresented in his or her L2 productions. We focus on two groups of non-native speakers whose native language belong to either the Romance or the Germanic family. Our hypothesis implies that authors whose L1 is Germanic will use more Germanic cognates, whereas authors with a Romance background will overuse Romance cognates; and that this tendency will decrease with the level of English proficiency, down to the point of frequencies similar to those of native speakers.

To model the use of cognates and to examine the tendency towards Germanic or Romance cognates, we compute, for each author i and each synonym set s, a normalized count that reflects the proportions in which

i uses the words in s. For example, for the synonym set {*thankful, grateful*} (where the former is Germanic and the latter is Romance), a specific author used *thankful* 8 times and *grateful* 3 times; the normalized count is then $8/11$ for Germanic, $3/11$ for Romance.

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. The deviation from the benchmark as a function of the user proficiency level will provide an indication of the validity of the research hypothesis.

For example, for the synonym set {*mindless,senseless*} (again, the former is Germanic and the latter Romance), native speakers average preference towards *mindless* is 0.7098; this number is the benchmark from which the non-native speaker deviation will be calculated. For a specific non-native author from Denmark (*dalsgaard*) whose Germanic normalized count for the same synonym set is 0.889, the deviation from the threshold is $0.889 - 0.7098 = 0.1792$. For a different author from France (*CH4F*) who used the Germanic and Romance word equally (i.e., the normalized count is 0.5), the diversion from the threshold will be $0.5 - 0.7098 = -0.2098$.

If the hypothesis holds, we expect to see positive distances from the calculated threshold for users whose L1 is of Germanic origin (indicating overuse of the Germanic word(s) in comparison to native speakers), and negative distances for users whose L1 is of Romance origin. For both non-native groups, we expect the absolute distance values to decrease as the level of English proficiency rises.

This analysis is strongly affected by the specific synonym set and proficiency measure and should be interpreted considering these choices. Different frequency levels might affect the analysis: for example, the synonym set {*affluence, richness*} occurs only 500 times in non-native speakers posts, while {*attend, hang*} has over $30,000$ occurrences, with very different frequency level for each member of the synonym set (8026 occurrences of *attend* vs. $22,917$ occurrences of *hang*); this is another factor that might influence the analysis.

5 Preliminary Results

Figure 1 presents preliminary results of the analysis described in Section 4.3. The blue dots represent the distance of "Germanic" users from the native speakers threshold (the tendency towards words of Germanic origin, as defined in the previous section), and the red dots represent "Romance" user distances. The x-axis is the proficiency level of authors according to one measure, and the y-axis stands for the distance, where the native speaker threshold is of course 0.

Since each group has close to 500 data points (one per author), it is hard to visualize the analysis when each point is represented individually. Therefore, each of the two populations was aggregated into 25 groups, bases on their English proficiency scores in the relevant measure. Figure 1-left shows directed distances from native threshold for the synonym set {*sink, pass, lapse*} as a function of the syntactic measure DC/C (dependent clauses per clause). Figure 1-right shows directed distances from native threshold for the synonym set {*belittle, denigrate*} as a function of the lexical measure TTR (type token ratio).

The figures mathches, to some extent, our expectaion of seeing a greater prefrence towards Germanic words among "Germanic" users, shown by the fact that blue dots ("Germanic" users) in general are placed higher along the y-axis compared to the red dots ("Romance" users). In addition, although including quite a

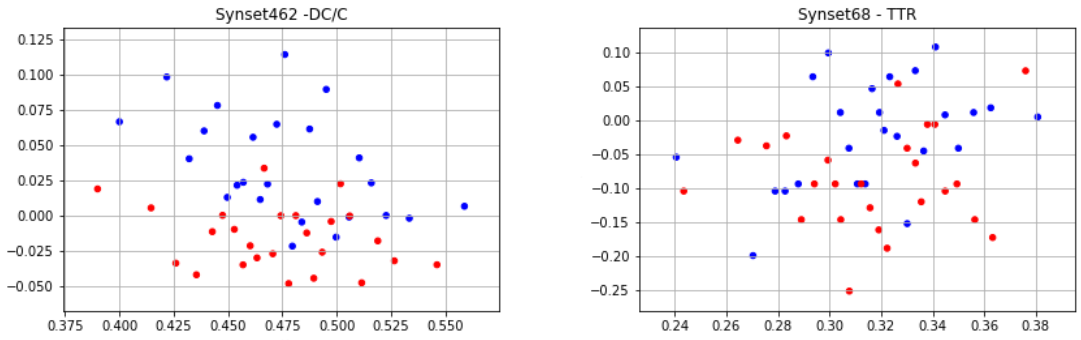

Figure 1: Over- and under-use of Germanic words as a function of proficiency. Left: synset 462 vs. English syntactic proficiency measure DC/C; right: synset 68 vs. English Lexical proficiency measure TTR.

few outliers, bilinguals with higher proficiency scores (along the x-axis) seem to diverge less from the native speaker pattern (manifest by a greater for these individuals around 0 values on the y-axis). This pattern lends preliminary support to the hypothesis underlying this research proposal.

References

David A. Balota, Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. The english lexicon project. *Behavior Research Methods*, 39(3):445–459, Aug 2007. ISSN 1554-3528. doi: 10.3758/BF03193014.

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.

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.

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

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.

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. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496, 2010.

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

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.

Anat Prior, Tamar Degani, Sehrab Awawdy, Rana Yassin, and Nachshon Korem. Is susceptibility to cross-language interference domain specific? *Cognition*, 165:10 – 25, 2017. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2017.04.006. URL http://www.sciencedirect.com/science/article/pii/S0010027717301002.

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.