DO NON-NATIVE SPEAKERS CREATE A PRESSURE TOWARDS SIMPLIFICATION? CORPUS EVIDENCE

ALEKSANDRS BERDICEVSKIS[*1]

[*]Corresponding Author: aleksandrs.berdicevskis@lingfil.uu.se
[1]Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden

An influential paradigm within evolutionary linguistics is that languages change in response to socioecological pressures. Language complexity is a common parameter to test for such adaptation (Beckner et al., 2009: 12; Lupyan & Dale, 2010). Strong claims have been made about the evolution of complexity. e.g. that large proportion of non-native speakers in a population facilitates morphological simplification (Trudgill, 2011). While there exists evidence in favour of this claim (Bentz & Winter, 2013; Szmrecsanyi & Kortmann, 2009; Bentz & Berdicevskis, 2016), it still rests on several assumptions that have not been rigorously tested.

One such assumption is that the linguistic production of non-native speakers tends to be simpler than that of native speakers, thus creating a pressure towards simplification. Some quantitative comparisons of the complexity of L1 and L2 production have been made (Brezina & Pallotti, 2016 and references therein), but no studies involved large corpora of natural written production.

To address this issue, I create large corpora of native and non-native English, French, Italian and Spanish by using data from WordReference forums. At these forums, users have to indicate their native language in their profiles. For each of the four languages, a forum exists where the rules permit discussions solely in this language. I download the content of these forums, noting for every post the nickname of its author and whether the author is a native speaker of the language the post is written in. The resulting corpus sizes are reported in Table 1.

As a proxy of complexity I use lexical diversity (LD), operationalized in three different ways: type-token ratio (TTR) and related, but more sophisticated measures called HD-D and MTLD. TTR has been criticized for a number of shortcomings (Jarvis, 2002), but has an advantage of being easily interpretable. HD-D and especially MTLD are claimed to be more robust and less sensitive to

DOI:10.12775/3991-1.007

text size measures (McCarthy & Jarvis 2010). A comparison of some measures of morphological complexity (Bentz et al. 2016) suggests that TTR is doing reasonably well. Bentz et al. (2016) worked with parallel texts, which cannot be done in my situation, but I perform all comparisons on texts of equal lengths.

Table 1. Number of tokens in every corpus (in millions).

	Italian	*French*	*Spanish*	*English*
L1	3.5	6.6	22.4	49.8
L2	1	3.7	5.5	38.0

I define three thresholds: 200, 300 and 400 tokens. LD measures may be unreliable at lower thresholds (Koizumi & In'nami 2012), while higher thresholds yield too few datapoints. For every threshold *n*, the following procedure is repeated: posts shorter than *n* are discarded, for every other post, all three measures are calculated using the first *n* tokens. I fit then a mixed-effect regression model with an LD measure as the response variable, Speaker type (L1 vs. L2) as a main effect and Author as a random effect and perform a likelihood-ratio test against a null model without the Speaker type predictor.

Speaker type is a significant predictor for English (TTR and HD-D; all thresholds) and French (MTLD; threshold 200). For English, both TTR and HD-D show that L2 production is less complex, for French, MTLD shows that it is more complex. In all cases, the slopes are small, but not negligible. All other combinations of language, measure and threshold do not give significant results, but the observed differences suggest that Italian behaves like English, while Spanish behaves like French.

In the talk I review these results and potential reasons for differences between English and French. I also discuss implications for typological theories outlined in the first paragraph (can it be that L2 speakers create a pressure towards simplification in some cases, but not others?). Finally, I turn to methodological issues of measuring complexity of natural written production using unannotated corpora.

Speaking of methodological issues, it should be noted that there are several potential confounds. First, the results can be affected by orthographical variation. Second, different L1 backgrounds of L2 speakers may play a role. Factoring these parameters into analysis and including other types of measures than LD-based ones are natural further steps.

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M., Croft, B., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman. D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language learning 59(s1)*, 1–26.

Bentz, C.; Ruzsics, T.; Koplenig, A & Samaržić, T. (2016). A comparison between morphological complexity measures: typological data vs. language corpora. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics, Osaka, Japan, 11 December 2016.

Bentz, C., & Berdicevskis, A. (2016). Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC) , 26th International Conference on Computational Linguistics, Osaka, Japan, 11 December 2016.

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change 3(1)*, 1–27.

Brezina, V., & Pallotti, G. (2016). Morphological complexity in written L2 texts. *Second Language Research*, 0267658316643125.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19(1)*, 57-84.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40(4), 554-564.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, 5 (1), e8559.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2), 381-392.

Szmrecsanyi, B., & Kortmann, B. (2009). The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua 119(11)*, 1643–1663.

Trudgill, P. (2011). Sociolinguistic typology: social determinants of linguistic complexity. Oxford: Oxford University Press.