Далеко пойти на малом материале: стратегическое усвоение иностранных языков

Getting more out of limited input: strategic mastery of foreign languages

Laura A. Janda, UiT The Arctic University of Norway

laura.janda@uit.no

In the learning of foreign languages (L2), a major challenge is presented by languages that have rich inflectional morphology. In this article I describe a resource that exploits corpus data to provide maximally strategic input for L2 learners of Russian, and a project to extend this resource to a myriad of other languages and uses.

By “rich inflectional morphology”, I mean large nominal and verbal paradigms that use morphophonemic alternations and affixes to mark grammatical categories such as case, number, definiteness, gender, person, tense, aspect. By “large”, I mean more than five word forms in a paradigm, since five is approximately the number that can be easily subitized by human beings (Kaufman et al. 1949). In Eurasia, for example, the following groups of languages have a strong tendency toward rich inflectional morphology: languages from the Baltic, Celtic, and Slavic subfamilies of Indo-European, from the Uralic, Turkic, and Semitic families, as well as isolates such as Basque.

Because Zipf’s law (1949) applies also to the distribution of word forms, most of the word forms that constitute the large paradigms of languages with rich inflectional morphology are infrequent, and many are vanishingly rare. The facility with which native (L1) speakers of such languages navigate their large paradigms is a bit of a mystery, termed the “Paradigm Cell Filling Problem” (Ackerman et al. 2009), since native speakers routinely both comprehend and produce forms that are rare or even unattested. For example, a native speaker of Russian will have no difficulty understanding or, given the appropriate context, producing the gerund *недокармливая* ‘while underfeeding’ and present passive participle *недокармливаемый* ‘being underfed’ forms of the verb *недокармливать* ‘underfeed’. However, neither of these forms appear in the 360-million-word Russian National Corpus (RNC, <http://ruscorpora.ru/>), a language sample roughly equivalent in quantity to the lifetime exposure of a person between 40 and 70 years of age to their native language. Given this fact, it is rather unlikely that most Russians have ever encountered these forms, but lack of input does not hinder their proficiency.

Acquisition of rich inflectional morphology is essential to achieve L2 linguistic proficiency, but also a “bottleneck”, since morphology is acknowledged to be more difficult to master than both syntax and semantics (Slabakova 2009 & 2014, Jensen et al. 2019). There is reason to believe that strategically focusing L2 input on just the most frequent forms is actually more effective in promoting native-like proficiency than expecting L2 learners to master entire paradigms. A computational learning experiment (Janda & Tyers 2018) pitted two models against each other in a word form production task in Russian. In the first model, the computer learned the entire paradigm of each word (noun, adjective, or verb) it encountered. The second model, called the “single form model”, was provided only with the single most frequent form for each word. Both models were given the same production task: to predict one word form, specified by a combination of values of grammatical categories, like for example “Genitive Plural” or “Third Person Singular Present Tense”, for each of 100 previously unseen (new) words. After learning 1100 words, the single form model consistently outperformed the model that learned entire paradigms, both in terms of the percentage of correct forms (only the single form model achieved better than 85% correct), and in terms of the egregiousness of errors (the errors that the single form model made required less repair to arrive at the correct answer when measured in Levenshtein distance). It seems that it is more strategic to triangulate to a new word form from a number of overlapping partially filled paradigms than from an equally large number of complete paradigms.

Thanks to the existence of language corpora such as the RNC, we now have an empirical basis to determine which word forms are most common in language. Until now, L2 learners have, with notable exceptions (Sinclair 2004), received little benefit from corpus linguistics. The task of extracting and presenting the most strategic word forms to L2 learners is more complicated than simply harvesting high-frequency word forms from a corpus. The distribution of highest-frequency word forms is different for every lexeme, and the high frequency of every single form is uniquely motivated by the collocations and grammatical constructions it is associated with. For example, the highest frequency form of *балерина* ‘ballerina’ is the Instrumental Singular that appears in the construction *стать балериной* ‘become a ballerina’. By contrast, the highest frequency form of *бизнесмен* ‘businessman’ is the Genitive Plural *бизнесменов* because we often speak of the interests of businessmen or count them in large numbers. The highest frequency form of *опаздывать* ‘be late’ is the Gerund used in the common phrase *опаздывая* *на* ‘when running late for’. And by contrast, the highest frequency form of *ждать* ‘wait’ is the Third Person Singular Present Tense *ждёт* ‘s/he waits’.

A proof-of-concept resource already exists for Russian, the SMARTool (Strategic Mastery of Russian Tool), freely available at <https://smartool.github.io/smartool-rus-eng/>. This resource is housed in GitHub, which means that the source files are likewise freely available. The SMARTool presents the three most frequent word forms (or only two or one form if those constitute over 90% of the frequency distribution for a word) for over three thousand lexemes stratified according to CEFR (Common European Framework of Reference for Languages) Levels A1, A2, B1, and B2. Users can filter words for Level and Topics (such as “food” or “transport”), and can also access words from a Dictionary. In addition, users can filter words according to the “Analysis” (grammatical parse) of their most frequent forms, searching for example for words that appear often in Dative Plural or Imperative or any other combination of grammatical category values. The SMARTool displays all words in the context of a corpus-inspired sentence that illustrates a typical collocation and/or grammatical construction for that word, and users can click to listen to audio of the Russian pronunciation of the sentence and to get an English translation.

The programming of the SMARTool has been specifically designed to make it maximally generic so that it will be possible to reuse the architecture to create more SMARTools, both for other L2s (Croatian, Czech, Finnish, North Saami, Polish etc.) and for user languages other than English. There is already a consortium of over sixty university researchers who are planning the development of SMARTools and the group is open to additional members, especially if we receive funding for the project.

The project will both create a pipeline for SMARTool development and address a wide variety of issues pertaining to Users, Linguistic facts, Technology, and Theory. I will briefly describe each set of issues.

*Users*. From the perspective of Europe, we plan to develop an optimal set of combinations of L2 and user languages for the development of SMARTools. We also want to find ways to adapt the SMARTool for both traditional (classroom) and non-traditional learning environments, as well as for heritage, distant/digital, and special-needs learners. Collection of backend clickstream data will make it possible to empirically evaluate the efficacy of SMARTools.

*Linguistic facts*. We will need to design adaptations of the SMARTool in order to accommodate various typological differences across languages, for example, in order to account for differences between agglutinative vs. fusional morphology. Selection of the lexicon is also an issue since a corpus does not necessarily reflect the language that an L2 learner needs most. Most corpora are deficient in their representation of spoken language, and there are a host of other challenges that present themselves such as how to deal with homonymy, polysemy, taboo language, diglossia, and dialects. In languages that mark gender on nouns, directly harvesting the highest frequency lexemes and word forms from a corpus will result in a gender bias (Kuznetsova 2015, Elmerot 2017), with many of the words for females under- or unrepresented. For example, in Russian, *студент*, the word for a male (or generic) student is over fourteen times more frequent in the RNC than the female equivalent *студентка*. Once the lexemes and words forms have been selected based on corpus and other considerations, it is necessary to determine what grammatical constructions and collocations are associated with each word form. This requires further corpus analysis and can also be aided by resources such as the Collocations Colligations Corpora for Russian (<http://cosyco.ru/cococo/>).

*Technology*. We will need to devise workarounds for languages that do not have large corpora, as well as adjustments for the overrepresentation of written genres in existing corpora. We hope to implement both crowdsourcing and gamification in the development of SMARTools.

*Theory*. We hope to learn enough to suggest a realistic cognitive model of inflectional morphology, as well as of how morphology is acquired in both L1 and L2. There are scholars (e.g., Albright 2002) who claim that there is one word form that is the most informative for predicting other forms in a paradigm, and some have attempted to measure informativity (a.k.a. “conditional entropy” Ackerman & Malouf 2016). Other scholars (e.g., Bochner 1994) disagree, suggesting that no word form has privileged status. We hope to test these hypotheses.

If funded, our consortium will tackle all of these issues, present our findings on a website, and in webinars, podcasts, and workshops, and we will produce a series of SMARTools that will enable L2 learners to reap the full benefits of scientific insights from corpora and e-learning.

Acknowledgement

The SMARTool has been supported by the author’s employer, UiT The Arctic University of Norway, and by grant number CPRU-2017/10027 from DIKU, the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (<https://diku.no/en/about-diku>).

Works Cited

Ackerman, Farrell, James P Blevins & Robert Malouf. 2009. “Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter.” In James P. Blevins and Juliette Blevins (eds.), *Analogy in Grammar: Form and Acquisition*, 54–82. Oxford: Oxford University Press.

Dissertation, University of California, Los Angeles.

Ackerman, F. and R. Malouf. 2016. Implicative relations in word-based morphological systems. In Hippisley A, Stump G (eds.), *Cambridge Handbook of Morphology*, Cambridge: Cambridge University Press, 272– 296.

Albright, Adam. 2002. *The Identification of Bases in Morphological Paradigms*. PhD

Bochner, Harry. 1993. *Simplicity in Generative Morphology*. Mouton de Gruyter.

Elmerot, Irene. 2017. *These women’s verbs. A combined corpus and discourse analysis on reporting verbs about women and men in Czech media 1989–2015.* MA thesis, Stockholm University. http://www.diva- portal.org/smash/get/diva2:1161582/FULLTEXT01.pdf.

Janda, Laura A. & Francis M. Tyers. 2018. Less is More: Why All Paradigms are Defective, and Why that is a Good Thing. *Corpus Linguistics and Linguistic Theory* 14(2), 33pp. doi.org/10.1515/cllt-2018-0031.

Jensen, I. N., Slabakova, R., Westergaard, M., & Lundquist, B. 2019. The Bottleneck Hypothesis in L2 acquisition: L1 Norwegian learners’ knowledge of syntax and morphology in L2 English. *Second Language Research*. DOI: 10.1177/0267658318825067

Kaufman, E.L.; Lord, M.W.; Reese, T.W. & Volkmann, J. 1949. The discrimination of visual number. *The American Journal of Psychology*. 62 (4): 498–525. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.2307/1418556](https://doi.org/10.2307%2F1418556)

Kuznetsova, Julia. 2015. *Linguistic Profiles: Going from Form to Meaning via Statistics.* Berlin: De Gruyter.

Sinclair, John McH. (ed.) 2004. How to use corpora in language teaching (= *Studies in Corpus Linguistics*, 12).

Slabakova, Roumyana. 2009. How is inflectional morphology learned? *Eurosla Yearbook*, *9*, 56–75. DOI: 10.1075/eurosla.9.05sla

Slabakova, Roumyana. 2014. The bottleneck of second language acquisition. *Foreign Language Teaching and Research*, 46 (4), 543–559.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Reading, MA: Addison-Wesley.