Title:
Do "Particles" Deserve to be Classified as a Part of Speech? A View from Russian

Abstract:
Our usage-based analysis of corpus data on Russian particles addresses both theoretical questions about the cognitive status of parts of speech and practical concerns about how particles should be represented in models of grammar used for natural language processing, intelligent computer assisted language learning, and rule-based machine translation.

Langacker (2013: 96, 115-117, 122) accepts that the use of part-of-speech terms such as *noun*, *verb*, *adjective*, etc. is unavoidable, but cautions that, like other linguistic cognitive categories, parts of speech have fuzzy overlapping boundaries and are necessarily motivated by a conceptual basis (i.e., they fulfill the "content requirement"). In Langacker's Cognitive Grammar, for example, a noun profiles a *thing* (prototypically a conceptually autonomous stable material object in space), whereas a verb profiles a *relationship* (prototypically an immaterial event in time conceptually dependent on its participants). Langacker's analysis can be extended to categories such as adjectives, adverbs, prepositions, etc. and most of these categories are also descriptively valuable for computational modeling of grammar, although significant challenges in part-of-speech tagging remain (Manning 2011).

However, *particle* lacks a coherent conceptual basis as a part-of-speech category. The Academy Grammar (Švedova et al. 1980: 723-731) defines *particle* for Russian as the class of "non-lexical" words that perform a broad spectrum of functions, including expressing "the widest possible range" of characteristics. The authoritative Grammatical Dictionary of Russian (Zaliznjak 1980) lists over 100 "particles" in Russian, many of which have multiple part-of-speech designations. Given the high frequency of some "particles", the resulting ambiguities pose a major challenge to computational modelling of Russian.

We focus on nine high-frequency particles that are ambiguous across various parts of speech (*ved'*, *da*, *daže*, *ešče*, *že*, *li*, *net*, *slovno*, *tak*). A ten-fold cross-validation of hand-tagged corpus data using a Hidden Markov Model trigram tagger achieves very poor accuracy, suggesting that the tagging of "particles" is inconsistent. This is probably not surprising given that the guidelines for tagging of "particles" in the Russian National Corpus (Sičinava 2005) are inadequate.

We offer a new set of annotation guidelines for Russian particles that is based upon corpus data and eliminates many of the problems posed by this group of words, largely by reassigning them to other parts of speech. We then show how this can improve tagging so that an automatic tagger can achieve better accuracy while also better reflecting our theoretical understanding of parts of speech.

References:
Langacker, Ronald W. 2013. Essentials of Cognitive Grammar. Oxford: Oxford U Press.

Manning, C. D. (2011) "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In Alexander Gelbukh (ed.), Computational Linguistics and

Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608, pp. 171--189.

Sičinava, D. V. 2005. Obrabotka tekstov s grammatičeskoj razmetkoj: instrukcija razmetčika. http://ruscorpora.ru/sbornik2005/09sitch.pdf

Švedova, N. Ju. 1980. Russkaja grammatika, vol. I. Moscow: Nauka.

Zaliznjak, A. A. 1980. Grammatičeskij slovar' russkogo jazyka. Moscow: Russkij jazyk.

Authors (affiliation for all is UiT, The Arctic University of Norway): Francis M. Tyers (francis.tyers@uit.no), Anna Endresen (anna.endresen@uit.no), Robert Reynolds (robert.reynolds@uit.no), Laura A. Janda (laura.janda@uit.no)

Title:
Do "Particles" Deserve to be Classified as a Part of Speech? A View from Russian

Presentation preference: Oral presentation

Session: Main session