Ambiguity in Russian Morphology

Carole Tiberius, Dunstan Brown, Greville Corbett
Surrey Morphology Group, School of Arts
University of Surrey
Guildford GU2 7XH
{c.tiberius,d.brown,g.corbett}@surrey.ac.uk

The nature of the relationship between frequency of use and grammar in natural language is poorly understood. In order to understand this relationship better, we will look at textual frequency distributions in a language which encodes a reasonable number of grammatical distinctions in its word forms, namely Russian. In this paper, we will focus specifically on the relationship between ambiguity and frequency.

To do this we shall undertake a corpus analysis of Russian texts. We will use two corpora, the Uppsala corpus (Lönngren 1993, Maier 1994) and a corpus of Russian newspapers from the late 1990's, to check whether we find similar distributions across both corpora.

We will convert the corpora to unicode encoding to increase reusability and sharability. The corpora will then be segmented and lemmatised. The lemmatisation will be based on a predictive, formal model of Russian morphology, namely Network Morphology (Corbett and Fraser 1993, Brown 1998). This model has been implemented in the inheritance-based formalism DATR (Evans and Gazdar 1996). The lemmatiser will be based on theorem dumps derived from the DATR theory, augmented with a larger set of fully inflected forms derived from the electronic version of Zaliznjak's (1977) dictionary (Ilola and Mustajoki 1989).

Rather than going for a fully tagged corpus, we will take the output of the lemmatiser and use this labelled unresolved data to determine the amount of ambiguity in Russian text. In the first place, we will focus on the analysis of nouns and adjectives, since these are the categories where we expect to find most ambiguity in Russian. We will be able to establish how many words get more than one morphological analysis, how many of those are two-way ambiguous, how many are three-way or more ambiguous. We can then determine whether there is a relationship between ambiguity and the frequency distributions of a given feature value (e.g. dative case) within a subset of nouns from the inheritance hierarchy of the Network Morphology model.

Insight into these ambiguity distributions in the two Russian corpora will be of value for practical purposes and will also contribute to our understanding of the effects that frequency has on grammar.

References

Brown, Dunstan 1998. From the General to the Exceptional: A Network Morphology Account of Russian Nominal Inflection, PhD, University of Surrey.

Corbett, Greville G. and Norman M. Fraser (1993) Network morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* **29**. 113-42.

Evans, Roger and Gerald Gazdar. 1996. DATR: A Language For Lexical Knowledge Representation. *Computational Linguistics* **22**. 167-216.

Lönngren, Lennart (ed.) 1993. *Častotnyj slovar´ sovremennogo russkogo jazyka*. Uppsala: Uppsala University. (=Studia Slavica Upsaliensia 32)

Ilola, Eeva & Mustajoki, Arto. 1989. *Report on Russian Morphology as it appears in Zaliznyak's Grammatical Dictionary*. Helsinki: Helsinki University Press.

Maier, I. 1994. Review of Lönngren (ed.) Častotnyj slovar´ sovremennogo russkogo jazyka. *Rusistika Segodnja* **1**. 130-136

Zaliznjak, A. A. 1977. *Grammatičeskij slovar' russkogo jazyka*. Moscow: Russkij jazyk.