Building a Tagged Corpus of Russian: A Bazaar Approach

A Proposal for Honors in Russian

Christopher Tessone
tessone@knox.edu

September 2003

Department of Modern Languages
Knox College
Galesburg, Illinois
Abstract
The natural language processing (NLP) community has been able to improve many language processing systems by applying statistical methods to large electronic corpora. Corpus-based linguists use such texts to study aspects of languages that until now have been explained based on the personal experiences of the linguist. However, few annotated corpora of Russian are currently available to researchers; none are based on unedited Internet texts. To develop such a corpus would normally require paid annotators and a significant amount of time. However, using methods already applied to the open development of large-scale software projects, frequently referred to as the “bazaar” model, I will develop an annotated corpus of Russian texts to be used in NLP and linguistics research at virtually no cost.

1 Introduction
Recent years have seen a tremendous revival in the field of corpus linguistics, largely due to the advent of increasingly fast computers with greater storage capacity. With the appearance of large, computer-searchable corpora, the natural language processing (NLP) community has turned to statistical methods for exploiting annotated corpora to improve machine translation systems, question answerers, speech generation and recognition systems, and so on. Linguists have used them for demonstrating the relative rarity of the subject–intransitive verb sentence type in English (Sampson 2003), teaching
modal particles in German (Möllering 2001), and studying the genitive in Middle English (Allen 2002).

However, very little work has been done in applying statistical methods to Russian or improving Russian language processing systems using results from NLP. This is largely due to a shortage of annotated corpora of Russian. TeCoRus, a corpus of recorded speech from telephone conversations, has limited use for general NLP research, and the remaining corpora (including the Uppsala Corpus of Russian Texts from Uppsala University, Sweden, and the Corpus of Interviews from Tübingen University, Germany) are little better than the mass of text already available on the websites of dozens of Russian newspapers. Available annotated corpora are still in early stages of development, and do not address some very large spheres of Russian usage. In particular, they do not address a growing variety of written Russian—that of the Russian-language Internet.

2 Developing a New Corpus of Russian

Once we have realized the need for a new corpus, the question arises: how should it be built? In the first incarnation of the Penn Treebank, a corpus of 4.5 million English words, part of speech tagging of the full corpus and skeletal syntactic tagging of half of the corpus took nearly three years (Marcus 1994). A recent Japanese project, which lasted three and a half years, produced a corpus of just under one million tokens (Kurohashi 2003).\textsuperscript{1} Clearly corpus development is labor-intensive and potentially a very long-term project. The annotation task is often reduced to correcting computer-generated POS and syntactic tags, but the sheer number of words to be tagged, combined with the technical nature of the task, still makes annotating a large corpus very time-consuming.

However, a number of strategies may be used to reduce the time and money required to develop a large corpus of Russian. First, if a highly accurate POS and morphological tagger for Russian can be developed similar to the Japanese tagger, new text can be annotated quickly, with minimal proof-reading by a human being. The tagging of the NEGRA treebank of German suggests POS tagging can get off the ground with only a few thousand hand-tagged tokens (Brants 2003). Secondly, reported annotation methods and associated speeds from various other corpus development projects can be

\footnote{A token is the same as a word in many cases. However, in some cases where a word is composed of two clearly distinct lexical elements (for instance, \textit{it} and \textquotesingle\textquotesingle{}s in the word “it’s”), they are distinct concepts.}
compared to find the optimal way of automatically assigning tags to speed up the human annotation process as much as possible, such as using the tagger’s internal probabilities for highlighting questionable assignments (Brants 2003).

Finally, and perhaps most importantly, using an open-source model of development will reduce the cost, and hopefully the time, required to annotate the corpus, as well as improving the overall quality of the resulting data. This “bazaar” methodology (as opposed to the bottom-up, centralized “cathedral” approach used in traditional software development—see (Raymond 2001)) may help to transform NLP, bringing together researchers and linguists with the people who apply results in the field to everyday life. The key is that while the actual make-up of a given open-source development team may be quite volatile, if the project provides something worthwhile to the community, the number of capable developers will always be sufficient to advance the project.

The advantages of using text from the Internet are legion. Weblogs are updated frequently and are immediately available in electronic form. Because the particular community from which the text will be taken is international, usage patterns among the various diaspora communities in North America, Israel, and Europe can be studied in detail. Internet texts are also interesting because they are unedited and give a better picture of how people actually write.

### 3 The Plan

The target size for the annotated corpus is 500,000 tokens. This is smaller than many corpora (though not exceptionally so in comparison to other non-English parsed corpora), but it will ensure that I can obtain a useful product in one academic year, even if a community of volunteer annotators cannot be developed in that time. Additional texts can be easily added at a later date, and the work of writing tools for online annotation will in any case be completed by then. At the same time, I plan to develop a much larger untagged corpus, perhaps a hundred million tokens.\(^3\)

During the fall term of 2003, I will acquire electronic texts to be annotated, primarily from the Russian community in LiveJournal, a popular weblog site.\(^2\) Texts will be donated by users at LiveJournal, a popular weblog site. This text will later be annotated and joined with the rest of the corpus, but more raw text is better than a small amount of parsed text for the purposes of creating the frequency dictionary.

\(^{2}\)Texts will be donated by users at LiveJournal, a popular weblog site.

\(^{3}\)This text will later be annotated and joined with the rest of the corpus, but more raw text is better than a small amount of parsed text for the purposes of creating the frequency dictionary.
weblog community. The next step is to develop a set of tags to be applied by
the POS and morphological tagger. Since this will have an effect on the way
the data are presented and the type of data the corpus will hold, this is an
important step and will require coordination with Russian Academy of Sci-
ences standards and some study of the tagsets used by other Slavic-language
corpora.

During the winter and spring terms of 2004, I will hand-tag some 50,000
tokens of the corpus with POS and morphological information and develop
or adapt for Russian a tagger to automatically annotate the corpus with
this information. Also, an annotation interface for the volunteer annotators
will be developed, and I will begin to recruit a team of annotators. Once
all this is completed, some preliminary work can be done on the syntactic
annotation of the corpus—a tagset must be chosen, an online tree editor
must be developed or adapted to the Unicode character set, and preliminary
studies must be done for writing a syntactic parser. However, the actual
annotation of the corpus with syntactic information is outside the scope of
this project.

4 Conclusion

The intended result of this Honors project is a corpus of Russian text con-
sisting of 500,000 tokens annotated with part of speech and morphological
information, and a suite of open-source tools for annotating Russian text.
If all of these goals are achieved, derivative corpora annotated with other
information can easily be developed, new text can be annotated, corrected,
and added to the corpus for future releases and eventual use by developers
and NLP researchers, and existing text can be corrected in an open envi-
ronment. With luck, this model will encourage others to develop corpora
in new languages and will enrich the study of natural languages both inside
and outside academia.

References


Allen, Cynthia L. “The Early English ‘his Genitives’ from a Germanic
Perspective.” Proceedings of the 2002 Conference of the Australian


