11-712: NLP Lab Report

Maya Tydykov

April 25, 2014

1 Basic Information about Russian

The Russian language is an Indo-European language spoken primarily in Russia and in other parts of the world by approximately 162 million people. It belongs to the Eastern branch of the Slavic language family (Lewis et al., 2013). Russian is a free word order language, although according to (Dryer and Haspelmath, 2013), it is primarily Subject-Verb-Object (SVO). Russian has three genders and six cases, which are marked via suffixes on words. It is written using the Cyrillic script, which was originally created for 9th-10th century Slavic language speakers in order to translate the Bible along with other church texts ("Cyrillic alphabet").

2 Past Work on the Syntax of Russian

A wide range of phenomena concerning the syntax of Russian has been studied in recent years. (Franks, 2005) presents an overview of several issues which have recently been a focused on in the course of these studies. These issues include that of the second dative and nearest conjunct agreement phenomena in Russian. Another subject which has recieved a great deal of attention is that of numerals in Slavic languages. One issue in the domain of Russian syntax that has been studied extensively and that is particularly relevant to the problem of parsing is word order. There is a large body of work concerning Russian's free word order, which is referred to as scrambling in literature on syntax. An overview of the most influential works on the subject of free word order in Russian (as well as other Slavic languages) over the last several decades can be found in (Franks, 2005), and some recent work on this topic includes (Bailyn, 2008).

Russian word order has had a significant impact on the prefered methods of parsing. The fact that Russian has free word order, as well as Russian's rich morphology, make it difficult to parse using a constituency framework because of the increase in the number of rules that would result from such an attempt. Thus, instead of constituency parsing, dependency parsing has been the standard method used for parsing Russian (Skatov et al., 2013). In 2012, the NLP Evaluation forum RU-EVAL held a Russian syntactic parsing evaluation task in which seven dependency parsing systems were evaluated. The purpose of the conference was to get an overview of the current state of the art in parsing for Russian. The top two systems at the conference (ranked by F1-score) were Compreno and ETAP-3, both of which use primarily rule-based approaches (Gareyshina et al., 2012). The third-place parser in the competition, SyntAutom, also a rule-based parser, is an automata-based system (Antonova and Misyurev, 2012; Gareyshina et al., 2012). According to (Skatov et al., 2013), none of the rule-based parsers evaluated as part of the task are openly available to the public. Another parsing method, implemented in the DictaScope Syntax system, which is itself incoporated into a commerical product, was recently described in (Skatov et al., 2013). This method combines constituency and dependency parsing, attempting to eliminate disadvantages of each.

Few dependency parsers for Russian have been made openly available to the public. One such system is the Russian Link Grammar parser, based on the Link Grammar formalism introduced in (Sleator and Temperley, 1993). This formalism is similar to the dependency structure formalism in that it focuses on creating links between words rather than on grouping words into constituencies. However, the Link Grammar formalism differs in that the links are undirected (i.e., there is no head or child word), links can form cycles, and there is no root word. Another system is Russian Malt, a machine learning system that does not incorporate any rules. This system achieved a score in the RU-EVAL task which would have put it into third place, but did not formally participate in the competition (Sharoff and Nivre, 2011; Gareyshina et al., 2012).

3 Available Resources

While some well-developed resources such as annotated corpora exist for Russian, to the best of the author's knowledge, few of them are openly available. One unannotated corpus that is freely available to use is the MultiUN corpus, consisting of cleaned data in XML format, extracted from the United Nations Website (Eisele and Chen, 2010). This corpus could be handy because it has already been preprocessed. One problem with using this corpus, however, is that it is limited to a specific domain. In lieu of using a corpus that has already been prepared for use in NLP tasks, one can use Wikipedia to develop an annotated corpus, since many articles are available in Russian. Specifically, there are currently 1,085,000 articles in Russian on Wikipedia (Wik, 2014). Using Wikipedia would solve the aforementioned problem of having a limited domain.

A lexicon is another important resource in building a parser. The Russian version of WordNet is a one such lexicon that may be useful(Balkova et al.), although it seems that project development has not progressed over the last several years and that it may not have been completed, as no recent information seems to be available about its current status. Another potential resource that can be used as a kind of lexicon is Wiktionary. Wiktionary is a free online, collaborative dictionary available in multiple languages, including Russian (Wik, 2014). The Java-based Wiktionary Library (JWKTL), described in (Zesch et al., 2008), is a freely available Java API that provides access to Wiktionary in multiple languages and will help in processing the large amount of information available on Wiktionary.

4 Survey of Phenomena in Russian

The "second dative" phenomenon in Russian concerns constructions with two special semipredicatives , where a semipredicative is "an adjective that makes an adjunct predication of some item in the sentence, auxiliary to the main subject-predicate relation(80 Greenberg and Franks, 1991). These two special semipredicatives are "odin", which means "alone", and "sam", which means "oneself". They differ from other semipredicatives in their declension and case marking and in that their case always agrees with an antecedent in the same clause that they appear in if the clause is simple. The antecedent with which they agree can be a subject or an object, whereas with normal semipredicatives, the antecedent they agree with must be a subject. Furthermore, when the semipredicatives do not agree with their antecedent, their case in these situations is dative, rather than instrumental, which is the non-agreeing case for other semipredicatives. The second dative appears in infinitive phrases where the subject of the infinitive is not in nominative case or it is in nominative case, but there is an overt complementizer between the infinitive phrase and the matrix clause (Greenberg and Franks, 1991). Several differing explanations have been proposed for the phenomena of the second dative; according to (Franks, 2005), one adequate explanation is provided by checking theory.

Another phenomenon is that of nearest conjunct agreement, in which the verb can agree with the conjunct nearest to it. This phenomenon is usually seen in sentences that start with a prepositional phrase and where the verb is not in accusative case and comes before the subject, which is conjoined. (Franks, 2005) mentions one possible explanation for this phenomenon could be "LF feature lowering". The phenomenon of numerals in Russian involves the fact that the nominals quantified by numerals greater than "five" are in genitive case. However, when in an oblique phrase, the numeral and nominals are in oblique case.

A phenomenon which is particularly relevant to the choice of using a depedency framework to parse Russian is Russian's free word order. Although Russian is free word order, as mentioned above, some orders are prefered over others in neutral situations. Specifically, the neutral ordering of a sentence is generally SVO. Adjectives and demonstratives usually come before the noun they modify, though the order can be changed for various reasons (Bivon, 1971). Adpositions in Russian come before the noun phrase they modify (prepositions)(Dryer and Haspelmath, 2013). In generative literature, the free word order phenomenon is frequently referred to as "scrambling"(Franks, 2005). (Bailyn, 2008) questions the existence of scrambling as a way of accounting for free word order and instead proposes that a certain syntactic processes can be used to explain it. He argues that Russian is "underlyingly SVO", and that most alternative orders come about from a syntactic "Generalized Inversion" process and from Dislocation, rather than from a generalized scrambling process.

The phenomena in Russian of the second dative, nearest conjunct agreement, numerals and case, and free word order are only a handful of a many more various phenomena which characterize the Russian language. Several more of these are discussed in (Franks, 2005).

5 Initial Design

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

References

Cyrillic alphabet. Encyclopaedia Britannica Online Academic Edition. URL `http://www.britannica.com/EBchecked/topic/148713/Cyrillic-alphabet`.

Russian Wikipedia. Web, January 2014. URL `http://en.wikipedia.org/wiki/Russian_Wikipedia`.

Wiktionary. Web, January 2014. URL `http://en.wikipedia.org/wiki/Wiktionary`.

A. A. Antonova and A. V. Misyurev. Russian dependency parser syntautom at the dialogue-2012 parser evaluation task. In *Computational Linguistics and Intellectual Technologies*, 2012.

John Frederick Bailyn. *Word Order and Scrambling*. Blackwell Publishing Ltd, 2008.

Valentina Balkova, Andrey Sukhonogov, and Sergey Yablonsky. Russian WordNet. pages 31–38. URL `http://www.fi.muni.cz/gwc2004/proc/127.pdf`.

R. Bivon. *Element Order*, volume 7 of *Studies in the Modern Russian Language*. Cambridge University Press, Cambridge, 1971.

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL http://wals.info/languoid/lect/wals_code_rus.

Andreas Eisele and Yu Chen. Multiun: A multilingual corpus from united nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5 2010.

Steven Franks. Slavic languages. In *Handbook of Comparative Syntax*, 2005.

Anastasia Gareyshina, Maxim Ionov, Olga Lyashevskaya, Dmitry Privoznov, Elena Sokolova, and Svetlana Toldova. RU-EVAL-2012: Evaluating dependency parsers for Russian. In *Proceedings of COLING 2012: Posters*, pages 349–360, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL http://www.aclweb.org/anthology/C12-2035.

Gerald R. Greenberg and Steven Franks. A parametric approach to dative subjects and the second dative in slavic. *The Slavic and East European Journal*, 35(1):pp. 71–97, 1991. ISSN 00376752. URL http://www.jstor.org/stable/309034.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world, seventeenth edition. Web, 2013. URL http://www.ethnologue.com/statistics/size.

Serge Sharoff and Joakim Nivre. The proper place of men and machines in language technology processing russian without any linguistic knowledge. In *Computational Linguistics and Intellectual Technologies*, 2011.

Dan Skatov, Sergey Liverko, Vladimir Okatiev, and Dmitry Strebkov. Parsing russian: a hybrid approach. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 34–42, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2406.

Daniel D. Sleator and Davy Temperley. Parsing English with a link grammar. In *Proc. Third International Workshop on Parsing Technologies*, pages 277–292, 1993. URL http://www.cs.cmu.edu/afs/cs.cmu.edu/project/link/pub/www/papers/ps/LG-IWPT93.ps.

S. Ju. Toldova, E. G. Sokolova, I. Astaf'eva, A. Gareyshina, A. Koroleva, D. Privoznov, E. Sidorova, L. Tupikina, and O. N. Lyashevskaya. Nlp evaluation 2011-2012: Russian syntactic parsers. In *Computational Linguistics and Intellectual Technologies*, volume 2, 2012.

Torsten Zesch, Christof MÃŒller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC), electronic proceedings*. Ubiquitious Knowledge Processing, UniversitÃ€t Darmstadt, Mai 2008. URL http://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2008/lrec08_camera_ready.pdf.