O.V. Mudraya, B.V. Babych, S.S. Piao, P. Rayson, A. Wilson

DEVELOPING A RUSSIAN SEMANTIC TAGGER FOR AUTOMATIC SEMANTIC ANNOTATION [1]

1. Introduction

Semantic lexical resources play an important part in both corpus linguistics and natural language processing. Semantic annotation – semantic field analysis, in particular – is increasingly being used, with promising results, in computer text analysis, as a complementary disambiguation procedure to distinguish between different senses of the same word. As Jackson and Zé Amvela[2] highlight, a «semantic field arrangement brings together words that share the same semantic space», and thus provides «a record of the vocabulary resources available for an area of meaning».

Over the past years, large semantic lexicons such as WordNet[3], EuroWordNet[4], HowNet[5], have been built and applied to various tasks. During the same period of time, another large semantic lexical resource has been in construction at Lancaster University, UK, as a knowledge base of USAS multilingual semantic tagging system[6].

[1] This work is supported by the UK-EPSRC funded ASSIST project grant (EP/C004574 for Lancaster, EP/C005902 for the University of Leeds).

[2] *Jackson H. and Zé Amvela E.,* Words, meaning and vocabulary: an introduction to modern English lexicology. London / New York, 2000. P.112.

[3] *Fellbaum C. (ed),* WordNet: an electronic lexical database. Cambridge, Mass., 1998.

[4] *Vossen P.,* Introduction to EuroWordNet // N. Ide, D. Greenstein, P. Vossen (eds.) Special issue on EuroWordNet. Computers and the humanities. 1998. № 32(2-3). P. 73-89.

[5] http://www.keenage.com

[6] The semantic lexicon and the USAS tagger are accessible for academic research as part of the Wmatrix tool, for more details see **http://www.comp.lancs.ac.uk/ucrel/wmatrix/.**

Different from WordNet, EuroWordNet and HowNet, in which lexemes are clustered and linked via the relationship between word senses or definitions of meaning, the Lancaster semantic lexicon employs a semantic field taxonomy and maps words and multiword expression (MWE) templates to their potential semantic categories, which are disambiguated according to their context in use by a semantic tagger called USAS (UCREL semantic analysis system). Its lexicon is classified with a set of broadly defined semantic field categories, which are organised in a thesaurus-like structure.

First developed in projects for analysing interview transcripts[1] in English, the USAS semantic tagger has undergone development and improvement over more than a decade. In particular, during the Benedict[2] and ASSIST[3] projects, it has been improved along two dimensions: lexical resource expansion and multilinguality – the English Semantic Tagger (EST) has been ported to Finnish and Russian. In the following sections, we describe the semantic tagger, focusing on the functions relating to English and Russian, such as the semantic tagset, the lexical resources, the lexical coverage and the applications of the tool. In particular, we elaborate on the Russian Semantic Tagger (RST), part of USAS multilingual semantic tagging system, which is a software tool for undertaking the automatic semantic analysis of Russian texts with an online user-friendly interface.

[1] *Wilson A. and Rayson P.,* Automatic content analysis of spoken discourse // C. Souter and E. Atwell (eds.) Corpus based computational linguistics. Amsterdam, 1993. P. 215-226.

[2] *Löfberg L., Piao S., Rayson P., Juntunen J.-P., Nykänen A., and Varantola K.,* A semantic tagger for the Finnish language // Proceedings of the Corpus Linguistics 2005 conference. Birmingham, 2005.
http://www.corpus.bham.ac.uk/PCLC/cl2005_fst_fullpaper_final.doc

[3] *Sharoff S., Babych B., Rayson P., Mudraya P. and Piao S.*, ASSIST: Automated Semantic Assistance for Translators // Proceedings of the EACL 2006. Posters & Demonstrations. Trento, Italy. P. 139-142.

2. USAS semantic tagger

2.1. The semantic tagset

The USAS semantic annotation scheme was initially derived from McArthur's Longman Lexicon of Contemporary English[1] of approximately 15 000 words, relating to “the central vocabulary of the English language” and arranged into 14 semantic fields (or major codes), then further divided into a total of 127 group codes and 2441 set codes. The Lancaster semantic field taxonomy initially utilised the same basic format, but it has since been significantly modified in the light of practical tagging problems met in the course of ongoing research[2]. The current semantic tagset[3] reflects 21 major semantic categories, denoted by capital Latin letters; these top level domains are further sub-divided into 232 semantic sub-categories indexed by digit numbers and points.

For comparison, it is worth noting here a Russian work by Shatalova[4], who adopts a similar (although on a much smaller scale) approach to classifying English vocabulary, with Russian translations, in her English-Russian Thesaurus that contains about 3500 words classed by nine themes, each of which is further sub-divided into up to 30 sub-themes. Interestingly, she cites McArthur's Longman Lexicon of Contemporary English as one of her sources.

The core of the USAS semantic tagger is the semantic lexicon knowledge base, in which single words and MWEs are mapped to

[1] *McArthur T.,* Longman Lexicon of Contemporary English. London, 1981.

[2] *Archer D., Rayson P., Piao S., and McEnery T.,* Comparing the UCREL semantic annotation scheme with lexicographical taxonomies // G. Williams and S. Vessier (eds.) Proceedings of the EURALEX 2004. Lorient, France. P. 817–827.

[3] For full tagset, see **http://www.comp.lancs.ac.uk/ucrel/usas/**.

[4] *Шаталова Т. И.,* Англо-русский идеографический словарь. Москва. 1993.

their potential semantic categories. In addition to the basic tagset, some extra codes are used for denoting minor attributes. For example, +/- sign is used to denote positive and negative aspects of meanings. Another set of similar codes are *m*, *f* and *n* for male, female and neutral genders respectively. Often a lexical item is mapped to multiple semantic categories, reflecting its potential multiple senses. In such cases, the tags are arranged by the order of likelihood of meanings, with the most prominent one at the head of the list[1]. In each entry, the word is also mapped to its part-of-speech (POS) category for the purpose of reducing ambiguity. Certain lexemes show a clear double (or even triple) membership of categories. A slash is used to combine the double/triple membership categories into a so-called portmanteau category[2] :

rebel	*VV0*	*G1.2/A6.1- S8- A6.1-*
waiter	*NN1*	*I3.1/F1/S2.2m*
адмирал	*S*	*G3/S7.1+/S2mf L2mf*
больничный	*A*	*B3/H1 Q1.2/B2-*

2.2. Semantic lexicon resource

As was indicated, the USAS semantic categories and tags were originally employed in the EST, and later successfully ported to Finnish and Russian by largely reusing the framework of the English tool with necessary adjustments. In the ASSIST project, we have been developing a parallel tool for Russian – RST. Conveniently, the USAS semantic categories are compatible with the semantic categorizations

[1] For English, Collins COBUILD on CD-ROM 2001 Lingea Lexicon, ver. 3.1, and occasionally Encarta World English Dictionary 1999 Microsoft Corporation are consulted. For Russian, ABBYY Lingvo 10 English-Russian Electronic Dictionary 2004 and ГРАМОТА.РУ **http://www.gramota.ru/** are used.

[2] *Leech G., Garside R., and Bryant M.,* CLAWS4: The tagging of the British National Corpus // Proceedings of the COLING 1994. Kyoto, Japan. P. 622-628.

of objects and phenomena in Russian, as in the following example[1]:

poor	*JJ*	*I1.1- A5.1- N5- E4.1- X9.1-*
бедный	*A*	*I1.1- A6.3- N5- O4.2- E4.1-*

However, unlike English, Russian is a highly inflected language: generally, what is expressed in English through syntactic structures, is expressed in Russian via morphological inflections, such as case endings and affixation. To analyse the complex morpho-syntactic structure of Russian words, we adopted a Russian morpho-syntactic analyser *mystem*[2] that is used as the equivalent of the CLAWS[3] POS tagger in the USAS framework. Furthermore, as its output is encoded in Cp1251, which is commonly used for Cyrillics, a Cp1251-to-UTF8 encoding converter was employed to make *mystem* output compatible with the existing USAS components. Despite these modifications, the architecture of the RST software mirrors that of the EST components.

Similar to the EST, the main lexical resources of the RST include a single-word lexicon and an MWE lexicon. However, due to the highly inflectional nature of Russian words, only lemmas of the words are included in the single-word lexicon, as opposed to word forms in the English semantic lexicon. This occasionally presents a problem, as more than one word can share the same lemma, in which case the lemma-based entry may match wrong words. Word disambiguation techniques will be needed to deal with this problem[4].

[1] I1.1- = Money: lack; A5.1- = Evaluation: bad; N5- = Quantities: little; E4.1- = Unhappy; X9.1- = Ability, intelligence: poor; A6.3- = Comparing: little variety; O4.2- = Judgement of appearance: bad

[2] *Segalovich I.,* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the MLMTA 2003. USA. P. 273-280.

[3] *Garside R. and Smith N.,* A hybrid grammatical tagger: CLAWS4 // R. Garside, G. Leech, and A. McEnery, (eds.) Corpus annotation: linguistic information from computer text corpora. London, 1997. P. 102-121.

[4] Such disambiguation has not yet been implemented in the RST. Semantic disambiguation methods used in the EST are described in *Rayson*

Another major modification in the Russian single-word lexicon was incorporating a separate sub-lexicon of proper names, such as personal and geographical names, that had to be separated from the main single-word lexicon because *mystem* does not differentiate between proper and common names. As a result, a corresponding proper noun component is added to the RST. When a proper noun and common noun have the same form, the proper noun is given priority. The overall work flow can be described as follows: raw Russian text → *mystem* morpho-syntactic analyser → Russian semantic component (single words/proper nouns + MWEs) → semantic annotation.

We are building the Russian lexical resources by exploiting both dictionaries and corpora. We use readily available resources, e.g. lists of proper names, which are then semantically classified. To bootstrap the system, we have hand-tagged the 3000 most frequent words from the Russian National Corpus[1], and are now expanding our coverage within specific semantic fields, using online resources[2]. Subsequently, the lexicons will be further expanded by feeding texts from various sources into the RST and classifying words that remain unmatched.

Currently, the Russian lexicon contains 16 103 lemmas, of which 11 671 are common names and 4432 are proper names, and 713 MWEs. We aim at coverage of around 30 000 single lexical items and up to 9000 MWEs by the end of the on-going ASSIST project in March 2007 (for comparison, the EST currently contains 54 953 single word forms and 18 921 MWEs). Many of the MWE entries are

P., Archer D., Piao S. L., McEnery T., The UCREL semantic analysis system // Proceedings of the workshop on Beyond named entity recognition semantic labelling for NLP tasks in association with the LREC 2004. Lisbon, Portugal. P. 7-12.

[1] **http://ruscorpora.ru/** and **http://corpus.leeds.ac.uk/list.html**. Also, see *Sharoff S.* Methods and tools for development of the Russian Reference Corpus // D. Archer, A. Wilson, P. Rayson (eds.) Corpus Linguistics Around the World. Amsterdam, 2005. P. 167-180.

[2] For example, **http://www.terms.ru/**.

templates, capable of matching variations of MWE lexemes:

*follow*_* {Np/P*/R*} through_RP*	*A1.1.1 M1/K5 X2.4*
*без_* видим*_* {на/то} причин*_**	*X2.5- A2.2-*

2.3. RST evaluation

In the ASSIST project, we have evaluated the lexical coverage of the RST on a specially collected for the project 70-million-word Russian News Corpus which includes three major Russian newspapers, i.e., *Trud, Izvestiya* and *Strana.Ru*, published in 2002-2004. We achieved coverage of 79% (compared with 96% for the EST). For the RST lexical coverage evaluation, the Russian News Corpus was lemmatised and tagged with the help of *mystem*. Then rough disambiguation between different lexemes was performed by selecting the most frequent variant of the given word form found in the 1.6-million-word manually tagged part of the Russian National Corpus. Coverage of the RST was evaluated on the lemmatised corpus with punctuation. Unknown to the RST high-frequency words in the Russian News Corpus appear to be largely related to current political and social affairs; therefore, the Russian semantic lexicon is going to be enhanced with this vocabulary to reach the target lexical coverage of 90%.

3. RST user interface

A web-based user interface[1] has been designed for the RST. The web interface incorporates three web pages. The first page is a log-on page that requires the input of the user name and password. The main page allows a user to type or copy and paste Russian text into a text area for its subsequent semantic tagging. The output is displayed in a table, listing POS and semantic tag(s) for each word in the original text. There is also a special column for marking members of MWEs. The third web page is for retrieving lexicon entries for a given semantic tag, when a user wants to examine the composition of the lexicon.

[1] **http://148.88.224.86:8080/nlp_tools/rus_sem_tagger**

If the user types in a semantic tag, the interface lists all the entries containing this tag. It also provides an option for the user to choose between single-word and MWE sub-lexicons.

4. RST applications

The most obvious application of the RST is in performing computer-aided semantic analysis of Russian texts. Another related application is computer content analysis which is concerned with the statistical analysis of the semantic features of texts by grouping words and phrases into semantic field categories and counting word frequencies and semantic frequencies in the texts. The RST is also used in the development of tools for practising translators. In ASSIST, the RST is employed in the semantic annotation of Russian corpora in order to find non-literal solutions to difficult translation problems in comparable corpora[1] – collections of texts in different languages in the same genre, written approximately at the same time, which are not translations of each other, such as for example English News Corpus and Russian News Corpus or British National Corpus (BNC) and Russian National Corpus. Translation equivalents are found by matching similar situations described in terms of semantic tags. The ASSIST set of tools is still under development, which involves the extension of the EST, further development of the RST in order to reach the target lexical coverage of 90% of source texts, the improvement of the procedure for retrieval of semantically similar situations and the completion of the ASSIST user interface[2].

[1] *Sharoff S., Babych B. and Hartley A.,* Using comparable corpora to solve problems difficult for human translators // Proceedings of the COLING/ ACL 2006 Main Conference. Poster Sessions. Sydney. P. 739-746. **http://www.aclweb.org/anthology/P/P06/P06-2095**

[2] **http://corpus1.leeds.ac.uk/assist/v05/**