On automated semantic and syntactic annotation of texts for lexicographic purposes

Vladimir Selegey

Keywords: *automated linguistic annotation, syntax and semantic analysis, corpus-based lexicography, Internet as a corpus.*

Abstract

The main idea of this paper is that automatic annotation is the only means to secure an efficient access to the whole set of linguistic productions rather than merely a small subset of such productions annotated manually.
Why is it necessary for a lexicographer to turn to open unannotated corpora? There are two valid concurrent reasons for that: the ever-growing rate of linguistic changes, on the one hand, and, on the other, the regional, social and professional 'segmentation' of the language, requiring a differential approach to the language phenomena under analysis.

For the past 10 years or so, the line of research based on the 'Internet as a Corpus' approach has seen booming growth. As far as technologies are concerned, the means of access available to the researcher are much more modest in this case. The methods currently used for indexing the World Wide Web by search engines are based on principles that are far from being linguistic. In spite of the fact that there are projects like Semantic Web, the Internet remains so far a raw text corpus with rather unreliable data about the frequency of occurrence.

We are presenting ongoing project ABBYY Syntactic and Semantic Parser that offers technologies for the automated linguistic annotation of text corpora. These technologies make a seamless addition to the technologies for the production of representative sub-corpora relating to the major Internet segments. ABBYY Syntactic and Semantic Parser (SSP) is built on linguistic technologies developed within the scope of the ABBYY Compreno project. It is planned to be part of LingvoPro portal (http://lingvopro.abbyyonline.com/en). Compreno is a multi-language (at the moment English, Russian, German, Spain, French, Chinese) ongoing NLP project based on the combination of sophisticated linguistic modeling and modern methods of language structure analysis (recognition). It is a scalable linguistic technology to use at a basic level for a range of NLP applications. As far as lexicography is concerned, the most important feature of this system is that automatic linguistic annotation is derived from a thorough syntactic and semantic analysis of a sentence.

1. Two Approaches in Corpus-Based Lexicography

In modern corpus-based lexicography a corpus of texts is at the same time an object of investigation and an effective tool for extracting the data about syntactic and semantic properties of words.

Text corpora become an object of investigation due to the growing understanding that objective lexicography cannot rely on linguistic intuitions of an individual researcher or even a group of researchers.

Text corpora become a tool of investigation once a researcher is equipped with effective means of access to the data contained in the corpora with the help of linguistic technologies. It is a full-text search that was historically the first available technology of this type, and, as regards the languages with a well-developed system of inflexion, e.g. Russian, it was particularly important to take account of morphology (of all the inflected forms of a particular verb, noun, or adjective).

Roughly speaking, one could single out two areas in the corpus-based lexicography, dealing with the following:

1. Using specially prepared 'exemplary' text corpora that provide a researcher with a certain guarantee as far as the fullness of representation of the language phenomena under study is concerned. In fact, 'corpus' becomes equivalent to the notion of 'language'. There are many lexicographic projects that are based on such corpora as

BNC or the Russian National Corpus. Such corpora usually make it possible for a researcher to work with sub-corpora wherein homonymy has been eliminated manually or linguistic markup has been added, including syntactic and semantic one ((Hovy, Marcus and Palmer 2007), (Nivre 2010), (Atkins 2010)). Text indexing as well as the possibility to perform searches for various tags of the markup language have become a relatively simple issue in technical terms, due to which the lexicographer is given the opportunity to search for and calculate the frequency of occurrence of not only of individual words but also of word meanings and even semantic fields.

2. Turning to 'open' lexicographic resources, particularly the Internet. For the past 10 years or so, the line of research based on the 'Internet as a Corpus' approach has seen booming growth. As far as technologies are concerned, the means of access available to the researcher are much more modest in this case. The methods currently used for indexing the World Wide Web by search engines are based on principles that are far from being linguistic. In spite of the fact that there are projects like Semantic Web, the Internet remains so far a raw text corpus with rather unreliable data about the frequency of occurrence.

Why is it necessary for a lexicographer to turn to open unannotated corpora? There are two valid concurrent reasons for that: the ever-growing rate of linguistic changes, on the one hand, and, on the other, the regional, social and professional 'segmentation' of the language, requiring a differential approach to the language phenomena under analysis.

Thus, some research of the Russian language (Belikov 2010) has shown that where the periphery of the language is concerned, it is rather a linguist (grammar specialist) than a lexicographer whom the national corpus provides relevant data. It is particularly necessary to turn to open resources such as, for example, internet-blogs, where one is involved in research of the current idiom of those strata of the population that are most active socially, and where one is taking geographical factors into account.

In the following paper, we shall not dwell on the problems pertaining to the use of the internet as a corpus. We shall rather focus on the problems of the linguistic annotation of open corpora.

2. Linguistic annotation of open corpora

As authors in (Belikov 2010) conclude: *The main problem in applying the method of segmental statistics is the lack of a suitable instrument for automatic data processing.* A complete manual annotation of such corpora is impossible a priori, and a partial one makes no sense, unless some narrow individual segments of the Internet are meant that make it possible to produce a compact but relatively representative sub-corpus.
Available today there are two types of technologies for work with open corpora:

1. Means of access provided by search engines. Extended query languages of search engines are insufficient from a linguistic point of view and the statistical data are unreliable particularly in case of units with high frequency of occurrence.
2. Technologies for the dynamic production of sub-corpora and the collocational analysis of such sub-corpora similar to those that are used in the Sketch Engine (Kilgarriff 2004). The results of statistical analysis and clusterization on the basis of the collocation patterns provide a lexicographer with a rather valuable material, however these results are not equivalent to the linguistic annotation enabling, for example, to

solve tasks of machine learning or to research certain semantic and syntactic properties of individual word senses by means of syntactic or semantic queries (filters) as it is possible in the corpora with manual linguistic annotation.

What we offer are technologies for the automated linguistic annotation of text corpora. These technologies make a seamless addition to the technologies for the production of representative sub-corpora relating to the major Internet segments which we shall not touch upon herein (e.g. Sharoff (2005)).

3. ABBYY Syntactic and Semantic Parser

ABBYY Syntactic and Semantic Parser (SSP) is built on linguistic technologies developed within the scope of the ABBYY Compreno project. It is planned to be part of LingvoPro portal (http://lingvopro.abbyyonline.com/en). Compreno is a multi-language (at the moment English, Russian, German, Spain, French, Chinese) ongoing NLP project based on the combination of sophisticated linguistic modeling and modern methods of language structure analysis (recognition). It is a scalable linguistic technology to use at a basic level for a range of NLP applications.

One of such applications is ABBYY SSP, the system for automatic syntactic and semantic parsing of texts. As far as lexicography is concerned, the most important feature of this system is that automatic linguistic annotation is derived from a thorough syntactic and semantic analysis of a sentence.

Traditionally, in the analysis of sentences of natural language there have been singled out, corresponding to the levels of the language, the following stages:

- Lexical analysis: identifying lexemes, punctuation marks, other objects
- Morphological analysis: lemmatization, identifying the part of speech as well as other components of grammatical meaning (with disambiguation)
- Syntactic analysis: building the system of relations between words as well as between higher-order constituents (building syntactic trees)
- Semantic analysis (at the level of lexis and grammar) including semantic interpretation of syntactic relations and substituting universal semantic representation entities viz. semantic classes for word meanings of the given natural language.

ABBYY SSP can perform the parsing of natural language input at different levels: from morphological through semantic. However, it is well-known that disambiguation at any level of the language system, as a rule, requires recourse to a deeper level of analysis. That is why elements of syntactic and semantic analyses are used even for annotation at morphological level (disambiguation of grammatical meanings of the words in a sentence).

Such an approach gives rather good results. While the project is still under way at present, we have already started intensive testing of ABBYY Compreno technologies, mainly on texts in Russian and in English.

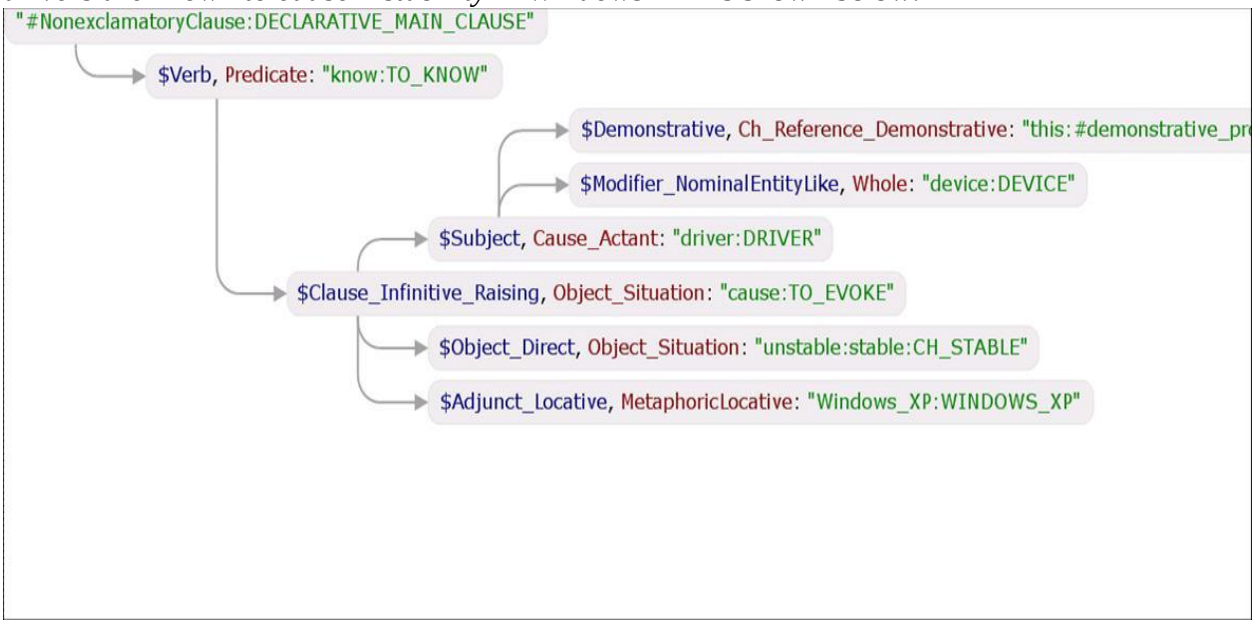
In the Parsing Cup at ‘Dialogue 2010’, which is a major computational linguistics conference organized annually in Russia, ABBYY SSP proved the best Russian-language parser, taking the lead in all events it participated, viz. ABBYY SSP scored 97,3% in grammatical meaning disambiguation and 98,1% in lemmatization (Astafjeva et al. 2010).

It is particularly important that the percentage of the words correctly annotated by the program exceeded the results shown by a human annotator on a text corpus (more than 3 mistakes per 100 words, mainly through fatigue and lack of attention). Therefore, ABBYY SSP can surely be used for morphological parsing of open corpora.

As far as syntactic and semantic parsing is concerned, we are well aware that its results depend to a certain extent on ABBYY Compreno's linguistic model. Nevertheless, the analysis undertaken shows that it is possible to map the structures that we obtain into representations that are popular with lexicographers, viz. dependency grammars in syntax, frame-based predicate models, WordNet synsets.

3.1. *Examples of Automatic Annotation and of Queries on the Basis of Such Annotation*

A sample deep structure produced by ABBYY SSP for the sentence "*The following device drivers are known to cause instability in Windows XP*" is shown below.



Same but written as text (with tags) looks as follows:

```
"#NonexclamatoryClause:DECLARATIVE_MAIN_CLAUSE"
   $Verb, Predicate: "know:TO_KNOW"
      $Clause_Infinitive_Raising, Object_Situation: "cause:TO_EVOKE"
         $Subject, Cause_Actant: "driver:DRIVER_AS_DEVICE"
            $Modifier_NominalEntityLike, Whole: "device:DEVICE"
               $Modifier_Attributive, OrderInTimeAndSpace:
"following:RELATIVE_ORDER"
         $Object_Direct, Object_Situation: "unstable:stable:CH_STABLE"
         $Adjunct_Locative, MetaphoricLocative: "Windows_XP:WINDOWS_XP"
```

What is important here is that knowing what the semantic structure of the sentence is like allows obtaining results on queries that are impossible to obtain on initial sentence. For example, taking into account the fact that 'drivers' is the result of raising and, as far as studying lexicographic co-occurrence is concerned, is related to the predicate 'cause', but not to the verb 'know'.

In case it is the research of lexicalized syntax phenomena that is necessary, an intermediate variant of syntactic annotation can be used wherein the movements had not been restored yet.

Various levels of annotation (XML encoded) can be subjected to indexing followed by a search with queries based on any element of linguistic structure:

- grammemes of grammatical categories

- names of syntactic and semantic relations
- lexemes, either specific lexical meanings or generalized semantic classes (names of lexical-semantic fields).

A study of syntactic and semantic models of specific lexical meanings based on the corpora of varying size is one important application of such annotation.
For example, it is possible to make queries of the following kind:

[$Adjunct_Locative:"COMPUTER"] "найти:TO_SEEK_FIND" (Russian)

or its English counterpart

[$Adjunct_Locative:"COMPUTER"] "find:TO_SEEK_FIND"

Such query will result in the sentences wherein the verb *find/искать* is accompanied by a locative adjunct, whatever the surface representations may be.
Thus, in the Russian output there could be found the following example:

> *Разбираясь с причинами неисправности* ***компьютера****, программисты* ***нашли в нем*** *крупного мотылька, замкнувшего какую-то цепь, и с тех пор вину за все компьютерные проблемы стали сваливать на насекомых.*

in which the query is realized by means of anaphora resolution, and it is irrelevant how far from each other the grammatical substitute (pronoun) and its antecedent are in the sentence. This peculiar feature has been preserved in its English counterpart (the example has been translated into English with the help of ABBYY Compreno):

> *Sorting out causes of malfunction of a* ***computer****, the programmers* ***found in it*** *a large moth who looped some chain and since those times they began to shift blame to insects for all the computer problems.*

Working with corpora of varying size, it is particularly important that one can significantly reduce the amount of information retrieved during searching by means of syntactic and semantic query-filters, thus limiting the retrieved information to relevant data alone.

Anyone using the internet as a lexicographic resource understands how difficult it is to solve the problem of 'parasitic information' retrieved by search engines.

Also possible are negative queries in which it is the regular patterns and meanings that are filtered, so that only 'deviations' are left, which are often very important to lexicographer. However, a detailed examination of how such annotation could be used in lexicography is not among the objectives of this paper.

4. Conclusion

The main idea of this paper is that automatic annotation is the only means to secure an efficient access to the whole set of linguistic productions rather than merely a small subset of such productions annotated manually. It is particularly important in case of new language phenomena which without fail must be represented in any up-to-date (electronic) lexicographic product.

Though ABBYY Compreno is a profit-driven project, it offers wide opportunities for scientific cooperation in the field of creation Internet-based annotated corpora.

The following issues remain open to discussion:

- To what extent are current state-of-the-art systems for automatic annotation adequate for lexicographic purposes?
- What should the evaluation of such systems be like? There can be no doubt that quality evaluation is on average relatively adequate for computational linguistics applications, however, in case of tasks requiring research the average rates of performance are unreliable and differential approaches need to be applied to the evaluation of the quality of annotation, taking into account various linguistic phenomena.
- The further we are from morphology and word sense disambiguation (based on certain canonical systems of word senses like WordNet), the greater is the number of problems having to do with the difference in approaches to the description of syntax and semantics.

References

Astafjeva I., A. Bonch-Osmolovskaya, A. Garejshina, et al. 2010. 'NLP Evaluation: Russian Morphology Parsers.' In *Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"*.

Atkins, B. T. S. 2010. 'The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data.' In G.-M. de Schryver (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Kampala: Menha Publishers

Belikov V. 2011. 'What are sociolinguists and lexicographers lacking in a digitized world?' In *Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011"*.

Hovy E., M. Marcus, M. Palmer, S. Pradhan, et al. 2007. 'OntoNotes: A Unified Relational Semantic Representation.' *International Journal of Semantic Computing* 1.4: 405–419.

Nivre J. 2010. 'Harvest Time. Explorations of the Swedish Treebank.' In M. Dickinson, K. Müürisep and M. Passarotti (eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*. Tartu: NEALT.

Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell 2004. 'The Sketch Engine.' In G. Williams and S.Vessier (eds.), *Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6-10, 2004*. Lorient: Université de Bretagne-Sud, 105–116.

Sharoff, S. 2005. 'Creating General-Purpose Corpora Using Automated Search Engine Queries.' In M. Baroni and S. Bernardini (eds.), *WACKY Papers*. Bologna: GEDIT.