# BONCH-OSMOLOVSKAYA / Skorinkin — Automatic

# <3 figures, 1 table>

# Automatic Semantic Tagging of Leo Tolstoy’s Works

Bonch-Osmolovskaya, A. and Skorinkin, D.

This paper presents an attempt to apply state-of-the-art text mining techniques currently employed by business to the task of literary research. The research in question is the preparatory part of a project called Tolstoy Digital. The main goal of this project is to convert the recently digitized 90-volume collected works of Leo Tolstoy into a novel digital humanities resource. We intend to create the so-called semantic edition of Tolstoy’s works by providing it with a complete semantic markup consistent with TEI schema1—a common standard of digital text encoding. The markup is supposed to include a wide spectrum of tags, from exact named entities and events to things like ‘editor’s note’, ‘author’s correction’, ‘draft version of the same text’, and so on. Named entities are also expected to be interlinked (co-reference resolution) and have external links to standard knowledge bases like DBpedia or Freebase.

With more than 46,000 pages of text that collectively contain 14.5 million words, Tolstoy is famed as one of the most productive writers ever. The sheer size of the material suggests that some automation of the markup is not only desirable but inevitable. In this paper we demonstrate how the use of an advanced language analyzer helps us extract named entities, various relations between them, and certain facts/events mentioned in Tolstoy’s texts.

The technology we apply to this task is called Compreno. It is a powerful text analysis platform being developed by ABBYY—a software company specializing in OCR, text analytics, and computational lexicography. Compreno converts text into a forest of syntactic-semantic trees that comprise dependency links and constituency structure. The analysis is based on the universal semantic hierarchy—a complex WordNet-like2 ontological structure that stores meanings rather than words.

The resulting trees contain nodes with all sorts of linguistic information attached to them: semantic classes from the said hierarchy, surface syntax slots, deep syntax slots, as well as sets of grammemes. Here is an example of a tree that the Compreno parser yields for the phrase ‘*28-го октября Кутузов с армией перешел на левый берег Дуная и в первый раз остановился, положив Дунай между собой и главными силами французов’* (*On the twenty-eighth of October Kutuzov with his army crossed to the left bank of the Danube and took up a position for the first time with the river between himself and the main body of the French*) from Tolstoy’s *War and Peace*:



Figure 1. Sample Compreno tree (automatically generated, no manual corrections).

Semantic classes from the hierarchy are green and capitalized, surface syntax slots are blue, deep semantic slots (they can be compared to Charles Fillmore’s [1968] deep cases) are dark red. Note that just like in English, in Russian there are several meanings for the word ‘силы’ (forces), but the parser performed disambiguation correctly, choosing the ‘FORCES\_AS\_PEOPLE’ semantic class. The parser is also capable of anaphora resolution (for more information on that, see Bogdanov et al. [2014]).

We realize that the example above is but a glimpse of the employed technology. Unfortunately, a thorough description of Compreno is well beyond the scope of this proposal. More details on that technology can be found in Anisimovich et al. (2012). So for now we will just note that for our research we used the information extraction system built upon and powered by Compreno.

The system in question allows writing sets of production rules to extract facts and entities from unstructured texts. The main advantage is that deep semantic representation of text provided by Compreno enables us to describe a whole range of different variants of a phrase in a very concise manner. For instance, we do not need to care about the word order (which is flexible in Russian), since the syntactic roles of different words remain the same. And even in case of voice transformation (*He loved her* → *she was loved by him*), only surface syntax slots change, while deep slots remain unchanged. Here is an example of a simple production:

"VERBS\_OF\_COMMUNICATION"

[Agent: active\_side "HUMAN"]

[Addressee: passive\_side "HUMAN"]

=>;

In this case we demand the system to find any Compreno subtree that has a node with a semantic class ‘VERBS\_OF\_COMMUNICATION’ or any of its descendant classes (since ‘VERBS\_OF\_COMMUNICATION’ is a very high-level class within our hierarchy and there are many lower classes that inherit from it) and at least two children nodes—one (or more) with ‘Agent’ deep syntax slot and another with ‘Addressee’ slot. Both children must also belong to / be inherited from a semantic class ‘HUMAN’ (which contains all sorts of subclasses that define people—names of occupations, social roles, relation terms, known proper names, etc.). Despite its obvious simplicity, this rule will extract many examples of communication between people (or, in our case, characters) like the ones below:

*Дмитрий, — обратился Ростов к лакею на облучке. — Ведь это у нас огонь?* (*‘Dmitri’,* ***addressed Rostov*** *to his* ***valet*** *on the box, ‘those lights are in our house, aren’t they?’*)

*Ну же пошел, — кричал он ямщику.* (*‘Now then, get on’,* ***he shouted*** *to the* ***driver***).

*Никаких извинений, ничего решительно, — говорил Долохов Денисову* (*‘No apologies, none whatever’,* ***said Dolokhov*** *to* ***Denisov***)

Formal representation of entities and facts being extracted is performed by means of ontology engineering. We develop ontologies using OWL language3 developed by the W3C. In the executable right-hand side of a production we can either create a new information object of a certain class of an ontology or modify the existing ones (add a surname attribute to a Person-class object, for instance). After the implementation of the rules, we receive the result of the information extraction process in the form of an XML document consistent with the RDF schema.4 Facts and entities also have links to their annotation—i.e., exact fragments within text. Here is a description of a ‘SpeechActivity’ fact, one of the many that were extracted from the second volume of *War and Peace*:

<BasicFact:SpeechActivity rdf:nodeID="bnode53D90FDA-F8F1-4DCB-8E8B-353456BEA164">

<BasicFact:theme rdf:nodeID="bnode261C3F83-0E71-4D4A-B60B-9268129D59F6"/>

<BasicFact:text rdf:datatype="http://www.w3.org/2001/XMLSchema#string" xml:lang="ru">Однако я тебя стесняю,</BasicFact:text>

<BasicFact:listener rdf:nodeID="bnode99CCE0DA-BDF2-4ECB-8CA1-8222D19F5641"/>

<BasicFact:type rdf:resource="http://www.abbyy.com/ns/BasicFact#TOS\_Quotation"/>

<BasicFact:speaker rdf:nodeID="bnodeF9F72F9B-9C30-43DD-B4F2-9E2EE3BE10DD"/>

</BasicFact:SpeechActivity>

‘Listener’ and ‘speaker’ attributes contain links to information objects of the Person class. In this case they point to Boris Drubetskoy (Person with surname ‘Друбецкой’ and name ‘Борис’) and Nicholas Rostov (Person with surname ‘Ростов’ and name ‘Николай’), respectively. Text attribute contains a string with the exact text of the speech. Below is the piece of text upon which the fact was extracted:

*Ростов сделался не в духе <. . .> Он встал и подошел к Борису.*

*— Однако я тебя стесняю, — сказал он ему тихо, — пойдем, поговорим о деле, и я уйду.*

(*Rostov became sullen <. . .> He got up and approached Boris.*

*‘I’ve come at a bad time I think,’ he said to him in a low voice. ‘Let us talk business, and then I’ll leave’.*)

So with the help of this system we can easily research the speech characteristics of each hero. Note also that this particular example clearly demonstrates the importance of correct anaphora resolution for the tasks of our in-depth text research.

The information extraction system used for this research has been in development for several years. It contains dozens of rule libraries with hundreds of extraction and identification (i.e., object merging) rules. Therefore, the first stage of our research consisted of applying these already existing rules to Tolstoy’s works (we limited ourselves to *War and Peace* at this point) with further analysis of the results.

Even the initial results seem very promising. For instance, the system correctly extracted the age of certain heroes, the parent-child relations between prince Vasili and Helene and many other family relations, the ‘illness’ of Dolokhov when he was wounded in a duel and the ‘Termination’ of the said illness when he was finally cured, the ‘Friendship’ between Dolokhov and Helene, the making of bast shoes by a servant of the Rostov family named Prokofy, and many more facts that to some extent represent the plot of the book. The co-reference relations between Person-class objects were also established with great accuracy: ‘Vasili Kuragin’, ‘prince’, and ‘Vasili’ are recognized as one individual; the same is true for ‘Andrew Bolkonski’, ‘prince Andrew’, ‘Andrew Nikolaevich’, and the diminutive ‘Andrysha’.

We were also able to find out quite a lot about the characters by exploring the positions they occupy in the predicate-argument structure. For instance, our data shows that prince Vasili Kuragin occurs more often in Agent and Possessor positions, while Bolkonskaya demonstrates inclination towards the roles of Experiencer, Object, and Addressee. This might be a reflection of character traits—the cunning and intrigue of profit-seeking prince Vasili versus the sensitivity and timidity of the shy and awkward princess Mariya.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Character** | **Agent** | **Object** | **Experiencer** | **Addressee** | **Possessor** |
| Andrey Bolkonsky | 705 | 183 | 225 | 59 | 83 |
| Pierre Bezukhov | 387 | 101 | 103 | 28 | 50 |
| Nikolai Rostov | 330 | 102 | 128 | 19 | 56 |
| Vasili Kuragin | 284 | 72 | 68 | 16 | 46 |
| Mariya Bolkonskaya | 225 | 92 | 132 | 24 | 39 |
| Anna Drubetskaya | 224 | 39 | 29 | 3 | 15 |
| Mikhail Kutuzov | 217 | 49 | 49 | 15 | 42 |
| Nikolay Bolkonsky (the old count) | 164 | 34 | 36 | 8 | 18 |
| Boris Drubetskoy | 157 | 77 | 55 | 15 | 22 |
| Elisabeth Bolkonskaya (the little princess) | 147 | 55 | 48 | 15 | 27 |
| Anna Scherer | 142 | 24 | 38 | 9 | 26 |
| Natasha Rostova | 142 | 33 | 25 | 10 | 22 |
| Pyotr Bagration | 136 | 27 | 33 | 13 | 16 |
| Anatole Kuragin | 113 | 32 | 26 | 8 | 14 |

Table 1.Most frequent occupants of different syntax deep slots (first volume only).

The resulting table also shows that in the first volume of the epic, princess Anna Mikhailovna Drubetskaya is a much more active character than her son Boris, although he is mentioned no less often than her. This is also clearly the reflection of the plot, where a determined, businesslike mother takes care of the career of her still-shy son (who’d become just as pragmatic later on).

Speech activity statistics also promises to be quite informative, though we have not analyzed it enough to come to conclusions yet. Below is some data we have obtained so far:

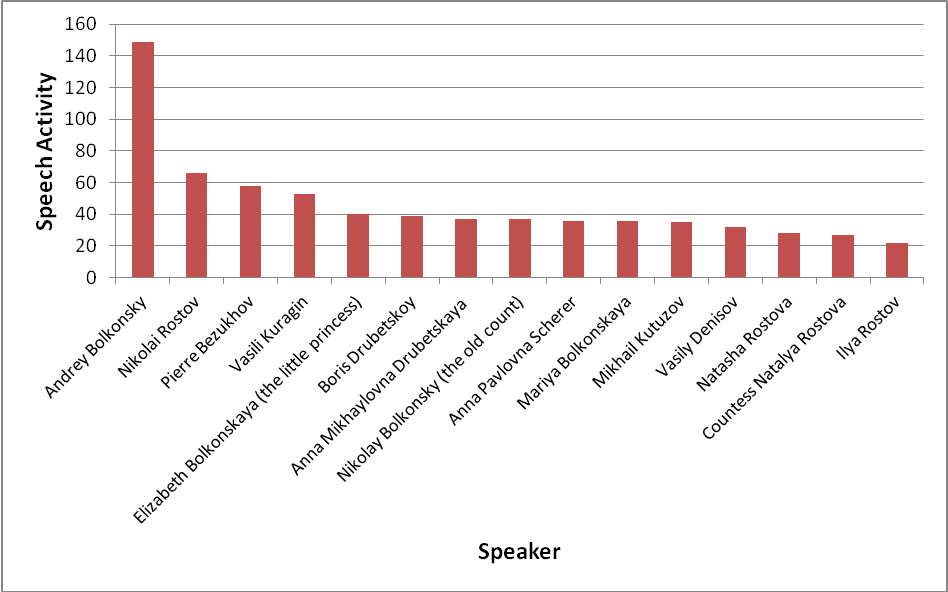


Figure 2. Speech activity frequency of different characters.

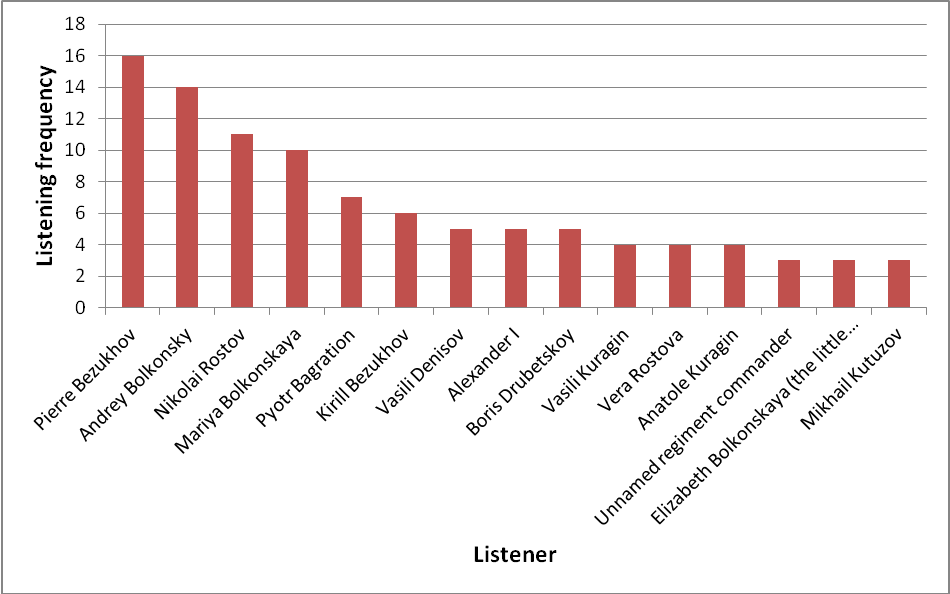


Figure 3. Listening frequency of different characters.

The next step of the research is the creation of our own extraction model within the existing system. This model, already in the making, is going to be designed and adjusted specifically to meet the needs of our research and is expected to help us extract much more information about Tolstoy’s characters, their description by the author, and their relations between each other.

## Notes

1. TEI: Text Encoding Initiative, http://www.tei-c.org.

2. WordNet, http://wordnet.princeton.edu/.

3. OWL Web Ontology Language Overview, http://www.w3.org/TR/2004/REC-owl-features-20040210.

4. Resource Description Framework, http://www.w3.org/RDF.

## References

**Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova, M. A., Selegey, V. P. and Zuev, K. A.** (2012). Syntactic and Semantic Parser Based on ABBYY Compreno Linguistic Technologies, Computational Linguistics and Intellectual Technologies. *Proceedings of the International Conference ‘Dialog’,* Bekasovo, Russia, http://www.dialog-21.ru/digests/dialog2012/materials/pdf/anisimovich.pdf, pp. 90–103.

**Bogdanov A. V., Dzhumaev, S. S., Skorinkin, D. A. and Starostin, A. S.** (2014). Anaphora Analysis Based on ABBYY Compreno Linguistic Technologies, Computational Linguistics and Intellectual Technologies. *Proceedings of the International Conference ‘Dialog’,* Bekasovo, Russia, http://www.dialog-21.ru/digests/dialog2014/materials/pdf/BogdanovAV.pdf, pp. 89–102.

**Fillmore, C. J.** (1968). The Case for Case. In Bach, E. and Harms, R. T. (eds), *Universals in Linguistic Theory*. New York: Holt, Rinehart and Winston, pp. 1–88.