Digital Scholarship in the Humanities Advance Access published December 31, 2016

Text mining *War and Peace*: Automatic extraction of character traits from literary pieces

Anastasia Bonch-Osmolovskaya and Daniil Skorinkin
National Research University 'Higher School of Economics', Russia

Abstract

This article presents a study of Leo Tolstoy's *War and Peace* by means of automatic syntactic and semantic analysis. Using a parser that extracts syntactic dependencies and semantic roles we were able to compare different characters of the novel in terms of the semantic roles they tend to occupy. Our data show that there are certain dependencies between the apparent personal traits of a character and his or her positions within the predicate structures. We hope that further research will help us gain more insights into the 'literary technique' of Tolstoy and enable us to create a semantic mark-up of his works.

Correspondence:
Skorinkin Daniil, National Research University 'Higher School of Economics', Myasnitskaya, 20, Moscow, 101000 Russia.
E-mail: skorinkin.danil@gmail.com

1 Introduction

The idea that natural language processing tools and techniques might be used for literary research can hardly be called novel. There have been a number of studies dedicated to applying such tools to works of fiction of different genres and languages (Elson *et al.*, 2010, Kokkinakis and Malm *et al.*, 2011). Most of such works are focused on relation discovery between characters and use either simple co-occurrence metrics or slightly more sophisticated lexico-syntactic patterns.

This article describes an attempt to apply state-of-the-art text mining technologies to the works of Leo Tolstoy. This study is the preparatory part of a project called 'Tolstoy Digital'. The ultimate goal of this project is to convert the ninety-volume collected works of Leo Tolstoy into a digital humanities resource (Bonch-Osmolovskaya, 2016). We intend to create a kind of a 'semantic edition' of Tolstoy's works by providing it with a markup consistent with Text Encoding Initiative (TEI) schema. The markup is expected to include a wide spectrum of tags, from persons, relations, and events to editorial notes and critical apparatus entries.

With more than 46,000 pages of text that contain about 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 desirable. In this article we demonstrate how the use of an advanced language analyser might help us extract information objects (entities and facts) which can be used for semantic mark-up of the text later on. We also show that automatic extraction of lexical patterns associated with different characters of the novel may improve our understanding of the 'literary technique' (see Shklovsky, 1925) used by the author for verbal distinction of personal properties of the main characters

2 Tools and Method

The technology we apply to this task is called COMPRENO (see Anisimovich *et al.*, 2012; Selegey, 2012). COMPRENO parser automatically converts text into a forest of syntactic-semantic trees which comprise dependency links and constituency structure. The analysis is based on the universal semantic hierarchy—a complex WordNet-like ontological

doi:10.1093/llc/fqw052

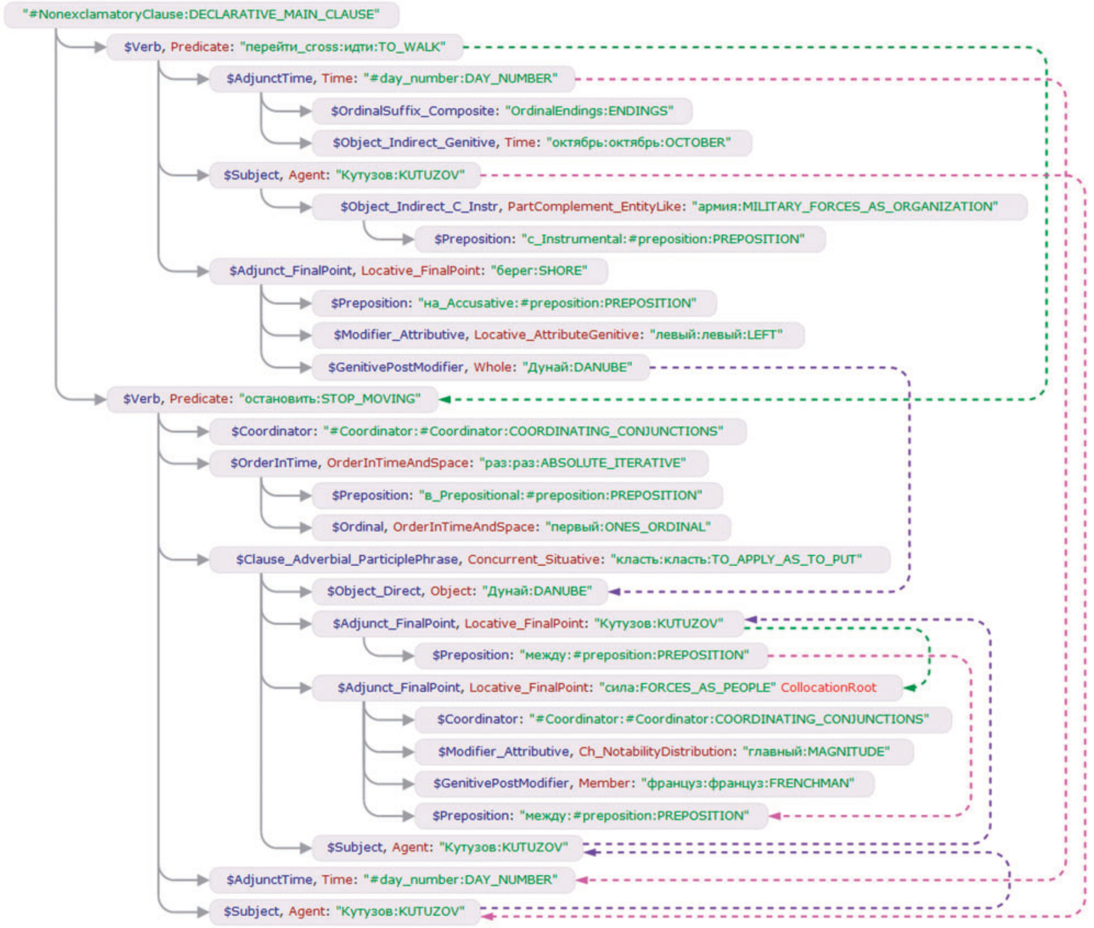
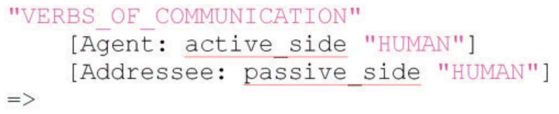

"#NonexclamatoryClause:DECLARATIVE_MAIN_CLAUSE"
$Verb, Predicate: "перейти_cross:идти:TO_WALK"
$AdjunctTime, Time: "#day_number:DAY_NUMBER"
$OrdinalSuffix_Composite: "OrdinalEndings:ENDINGS"
$Object_Indirect_Genitive, Time: "октябрь:октябрь:OCTOBER"
$Subject, Agent: "Кутузов:KUTUZOV"
$Object_Indirect_C_Instr, PartComplement_EntityLike: "армия:MILITARY_FORCES_AS_ORGANIZATION"
$Preposition: "c_Instrumental:#preposition:PREPOSITION"
$Adjunct_FinalPoint, Locative_FinalPoint: "берег:SHORE"
$Preposition: "на_Accusative:#preposition:PREPOSITION"
$Modifier_Attributive, Locative_AttributeGenitive: "левый:левый:LEFT"
$GenitivePostModifier, Whole: "Дунай:DANUBE"
$Verb, Predicate: "остановить:STOP_MOVING"
$Coordinator: "#Coordinator:#Coordinator:COORDINATING_CONJUNCTIONS"
$OrderInTime, OrderInTimeAndSpace: "раз:раз:ABSOLUTE_ITERATIVE"
$Preposition: "в_Prepositional:#preposition:PREPOSITION"
$Ordinal, OrderInTimeAndSpace: "первый:ONES_ORDINAL"
$Clause_Adverbial_ParticiplePhrase, Concurrent_Situative: "класть:класть:TO_APPLY_AS_TO_PUT"
$Object_Direct, Object: "Дунай:DANUBE"
$Adjunct_FinalPoint, Locative_FinalPoint: "Кутузов:KUTUZOV"
$Preposition: "между:#preposition:PREPOSITION"
$Adjunct_FinalPoint, Locative_FinalPoint: "сила:FORCES_AS_PEOPLE" CollocationRoot
$Coordinator: "#Coordinator:#Coordinator:COORDINATING_CONJUNCTIONS"
$Modifier_Attributive, Ch_NotabilityDistribution: "главный:MAGNITUDE"
$GenitivePostModifier, Member: "француз:француз:FRENCHMAN"
$Preposition: "между:#preposition:PREPOSITION"
$Subject, Agent: "Кутузов:KUTUZOV"
$AdjunctTime, Time: "#day_number:DAY_NUMBER"
$Subject, Agent: "Кутузов:KUTUZOV"

Fig. 1 Sample COMPRENO tree (automatically generated, no manual corrections)

structure that stores meanings rather than words (Manicheva *et al.*, 2012; Petrova, 2013). The resulting trees contain nodes with all sorts of linguistic information attached to them: semantic classes from the said hierarchy (e.g. 'PERSON_BY_FIRSTNAME' or 'VERBS_OF_ ADDRESSING'), purely syntactic 'surface slots' (e.g. $Subject or $Modifier_Adverbial), syntactic-semantic 'deep slots' (e.g. Agent or Experiencer; the 'deep slots' in the COMPRENO model are quite similar to Fillmore's 'deep cases'; Fillmore, 1968) and sets of grammemes. Figure 1 presents a sample tree that COMPRENO parser yields for a phrase in example (1) from Tolstoy's *War and Peace*:

> 1) 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 28th 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)

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). The information extraction system built upon COMPRENO 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. Figure 2 shows an example of a simple production:



```
"VERBS_OF_COMMUNICATION"
     [Agent: active_side "HUMAN"]
     [Addressee: passive_side "HUMAN"]
=>
```

Fig. 2 A rule for the extraction of Speech activity instances

In this case we demand the system to find any subtree which 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, and so on). Despite its simplicity, this rule will extract many examples of communication between people (or, in our case, characters) like the ones below in examples (2–4):

2) Дмитрий, — обратился Ростов к лакею на облучке. — Ведь это у нас огонь?

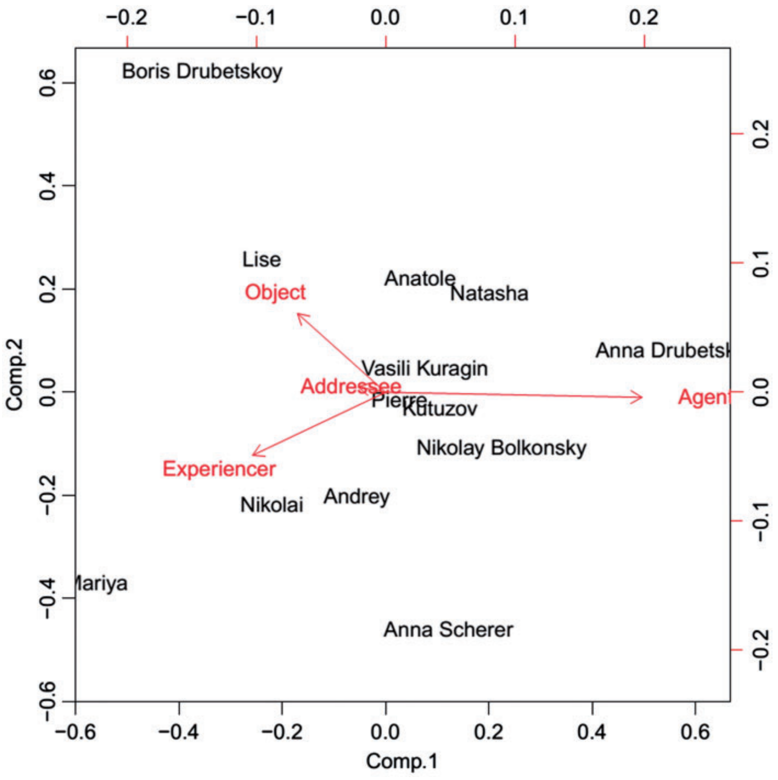


Fig. 3 Visualization of PCA of the semantic role distribution in the first volume of *War and Peace* (books 1–3 of the English translation)

‘Dmitri’, **addressed Rostov** to his **valet** on the box, ‘those lights are in our house, aren’t they?’

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

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

5) 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’)[1]

Entities and facts can be represented formally as parts of an ontological model. We develop ontologies using OWL language[2] 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 Resource Description Framework (RDF) schema.[3]

3 Experiment

In our experiment we intended to disclose the mechanisms of verbal contrast used by Tolstoy to distinguish the characters of his novel and to follow the evolution of their behaviour through the novel. The semantic role analysis presupposes a very high level of abstraction of both verbal classes and syntactic position of the argument. The verbal classes include experiential verbs of mental, perceptional, and emotional sphere (think, feel, fear, see, etc.), agentive transitive verbs (kill, break, create, etc.), agentive (unergative) intransitives (walk, laugh), patientive (unaccusative) intransitive (die, fall, sleep), and ditransitive verbs (give, address, show). The verbal arguments differentiate on the basis of their syntactic position within the verbal semantic frame—Experiencer, Agent, Patient, Addressee, and Possessor. Therefore the idea of the experiment was to capture the relations between characters within each volume[4] of the War and Peace with the help of the very abstract model which measures aptness of the characters to occupy specific syntactic positions in the context of the verbs of different semantics. We used COMPRENO parser to extract semantic roles of the predicate structure associated with the most prominent characters. The final list of roles included Agent, Object (equivalent to Patient), Experiencer, Addressee, and Possessor. Table 1 demonstrates the standardized results of the semantic role distribution for fourteen characters of the first volume of the novel.

For instance, Anna Drubetskaya is the most agentive character contrasted to her son Boris being the most object-like character. This is clearly a reflection of the plot, where a determined business-like mother takes care of the career of her yet shy son (who would become just as pragmatic later on). Sensitive Mariya (Marya) Bolkonskaya is allocated as a prototypical experiencer but one can also notice some inclination to experiential vector of her brother Andrey (prince Andrew) and Nikolay Rostov due to the emotional stress of their first battle experiences. The most striking property of the visualization is the similarity of the semantic roles between Anatole and Natasha. One may speculate that thus the reader gets a vague expectation of the major love intrigue of the novel which has been set long before it actually happens in the plot.

Principal Component Analysis (PCA) diagrams for other volumes (excluding epilogue, which is much less statistically significant) also seem to provide some meaningful insights into Tolstoy’s text. In the second volume (Fig. 4) Anatole is detached from Natasha (following their unsuccessful elopement and Anatole’s subsequent flight from Russia), and all the male characters are generally much more ‘agentive’ than the females, possibly due to the duels, gambling, and other supposedly ‘manly’ affairs. Natasha, who is undergoing her first serious moral crisis and is no longer the lively and joyful kid she used to be (see

Table 1 The distribution of semantic roles in predicate argument structure associated with the fourteen prominent characters of the first volume of *War and Peace* (books 1–3 of the English translation)

Character	Agent	Object	Experiencer	Addressee	Possessor
Natasha	0.61	0.14	0.11	0.04	0.09
Anatole	0.59	0.17	0.13	0.04	0.07
Anna Scherer	0.59	0.10	0.16	0.04	0.11
Nikolai Bolkonsky	0.63	0.13	0.14	0.03	0.07
Lise	0.50	0.19	0.16	0.05	0.09
Mariya (Marya)	0.44	0.18	0.26	0.05	0.08
Anna Drubetskaya	0.72	0.13	0.09	0.01	0.05
Boris Drubetskoy	0.48	0.24	0.17	0.05	0.07
Kutuzov	0.58	0.13	0.13	0.04	0.11
Vasili Kuragin	0.58	0.15	0.14	0.03	0.09
Nikolai	0.52	0.16	0.20	0.03	0.09
Pierre	0.58	0.15	0.15	0.04	0.07
Andrey	0.56	0.15	0.18	0.05	0.07

We then used PCA to visualize and analyse the data. As may be seen from Fig. 3, the PCA diagram of the very first volume already can be very well interpreted in terms of the main actors and 'undergoers'.

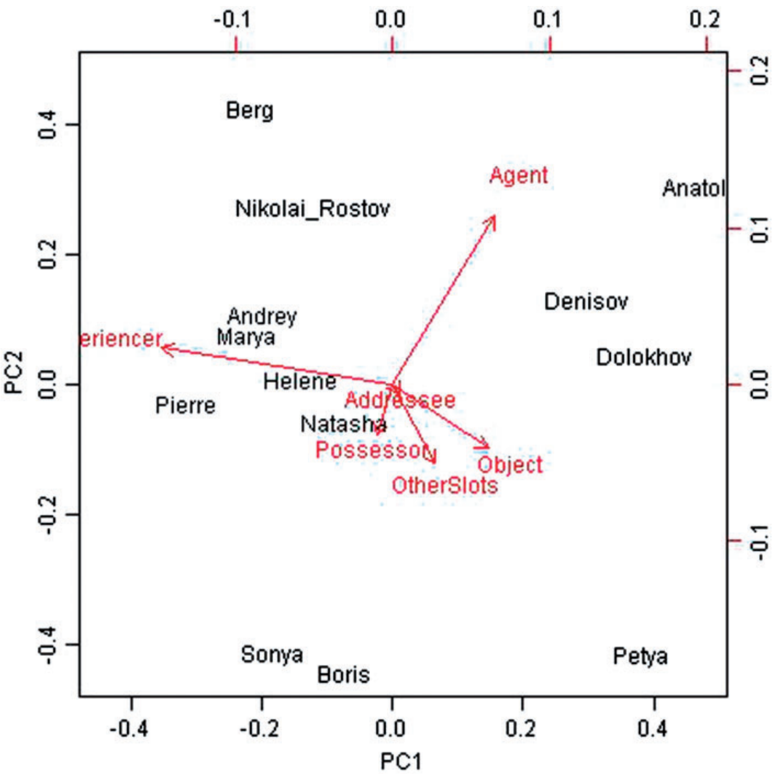


Fig. 4 Visualization of PCA of the semantic role distribution in the second volume of *War and Peace* (books 4-8 of the English translation)

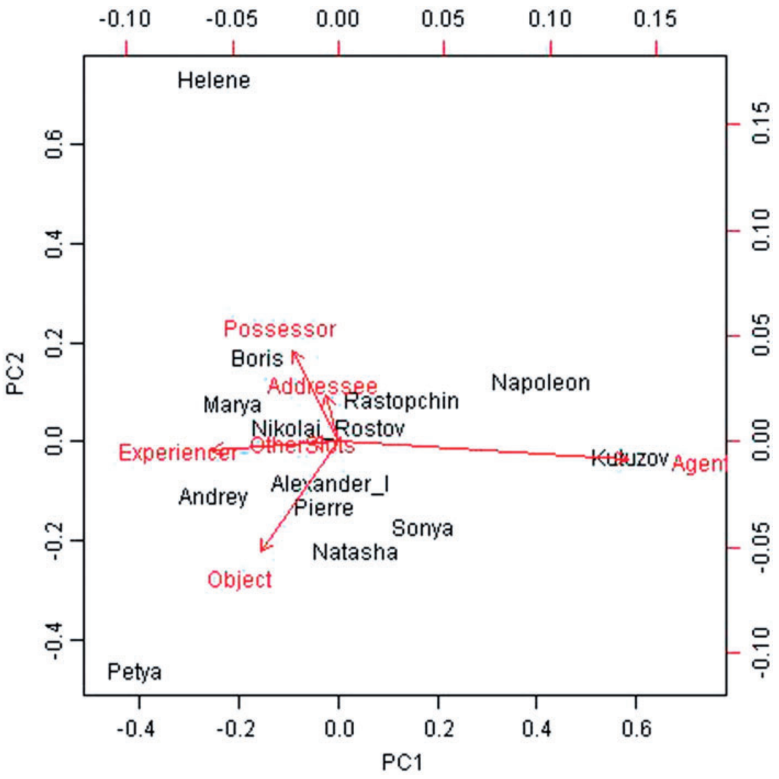


Fig. 5 Visualization of PCA of the semantic role distribution in the third volume of *War and Peace* (books 9–11 of the English translation)

Clay, 1998), becomes much closer to Pierre now, which could also be viewed as a reflection of the plot. But so does Helene, his newlywed wife, a rather cynical woman of high society and (supposedly) low moral standards, whose existence at this point forbids any possibility of romance between Pierre and Natasha. Note also that both women, Natasha and Helene, are located near the Addressee zone as far as they are the goals of the agentive actions of the three male characters. Here we might point that this whole section of the novel is to a large extent about the two women, Natasha and Helen, and the men they attract (Anatole, Dolokhov, Boris, Denisov). All these complex interactions can be formalized with the semantic role analysis in a new and potentially insightful way.

In contrast, Natasha's fiance?e Andrey Bolkonsky in this period is rethinking his life after his wife's death and his wounding in Austerlitz. Andrey is not agentive but is located now very close to his sister Marya in the Experiential zone (Fig. 4).

The third volume is rather distorted by the affairs of war, and the two opposing military commanders—Napoleon and Kutuzov—suddenly come to the forefront and become two most agentive characters (Fig. 5).

Helene leaves her husband to follow the imperial court to Vilna, where she apparently has yet another affair, now with Boris, who is also the closest to her on the diagram (Fig. 5). Pierre and Natasha remain close to each other, and Andrey keeps leaning towards Experiential positions.

The data for the last books of the novel (volume 4 in the Russian canonical edition) is interesting mainly due to the apparent culmination of prince Andrey's story. Severely wounded in the battle of Borodino, he ends up with Rostov family and eventually dies in Natasha's care. As he awaits his death

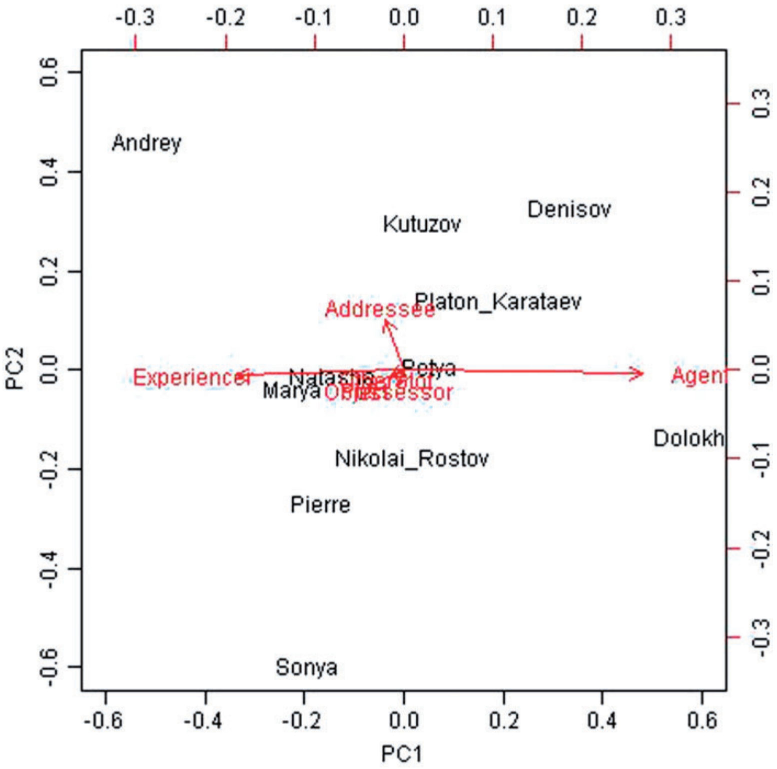


Fig. 6 Visualization of PCA of the semantic role distribution in the fourth volume of *War and Peace* (books 12–15 of the English translation)

and rethinks his life, he is unable to act anymore, but experiences a lot of feelings and is being nursed by others, which is reflected in his Objective/ Experiential position on the diagram (Fig. 6).

Andrey's sister Marya makes it in time to say her goodbyes, and both women witness Andrey's last minutes together. This common sorrow changes their formerly hostile relationship, and the two women become quite close to each other, which, apparently, is also reflected through semantic roles.

4 Conclusion

Our study shows that automatic semantic roles labelling could be applied to literary research. This technique appears to have some potential to reveal the techniques authors use to construct the complexity of relations along a linear narrative. Semantic roles proved to be quite informative of the characters' personal traits, providing us with 'objective' quantitative confirmation of something that was obvious to a reader or a critic but was not expressed explicitly in the text.

Some of our findings have direct relation to certain published critical interpretations of the novel. For instance, princess Marya's strong disposition towards the Experiencer role obviously correlates to the claims by some prominent Tolstoy scholars (Eichenbaum, 2009) that this shy and sensitive character was borrowed by Tolstoy directly from the XVIII century sentimentalists.

At the next stage of our research we intend to develop a dedicated information extraction model for literary research based on the system we are working with. This model, already in the making, is to be designed and adjusted specifically to meet the needs of such research and is expected to help us extract much more information about characters, their description by the author and their relations between each other. The first obvious improvement that could be made is to split the existing semantic roles into a bit more fine-grained and less abstract set that could, for example, distinguish between different types of ‘agentivity’ or ‘experientiality’ (sensitivity) of a character.

We also plan to pay special attention to the instances of direct speech within the novel since Tolstoy himself placed high emphasis on developing ‘character-specific language’. Our goal is to find out whether each character actually possesses his personal style of speech. Then we hope to broaden the scope of the study and include other authors to allow comparison between their writing styles and techniques.

Acknowledgement

This work was supported by grant 15-06-99523 from the Russian Foundation for Basic Research.

References

Anisimovich K., Druzhkin K., Minlos F., Petrova M., Selegey V., and Zuev K. (2012). Syntactic and Semantic Parser Based on ABBYY Compreno Linguistic Technologies, Computational Linguistics and Intellectual Technologies. In *Proceedings of the International Conference Dialogue*, Bekasovo: Russian State University for the Humanities Publishing House, pp. 90–103.

Bogdanov A., Dzhumaev S., Skorinkin D., and Starostin A. (2014). Anaphora Analysis Based on ABBYY Compreno Linguistic Technologies. Computational Linguistics and Intellectual Technologies. In *Proceedings of the International Conference Dialogue*, Bekasovo: Russian State University for the Humanities Publishing House, pp. 89–102.

Bonch-Osmolovskaya A. (2016) Digital edition of Leo Tolstoy works: contributing to advances in Russian literary scholarship. *Journal of Siberian Federal University, Humanities and Social Sciences*, 9(7): 1605–14.

Clay G. R. (1998) *Tolstoy’s Phoenix. From Method to Meaning in War and Peace*. Evanston, Illinois: Northwestern University Press.

Eichenbaum B (2009) *Works on Leo Tolstoy*. Saint-Petersburg: SPBSU Faculty of Philology and Arts.

Elson, D., Dames, N., and McKeown, K. (2010). Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics – Uppsala*: Uppsala University, pp. 138–47.

Fillmore C. J. (1968), *The Case for Case. Universals in Linguistic Theory edited by Emmon Bach and Robert T. Harms, Holt, Rinehart and Winston*, New York: Academic Press, pp. 1–88.

Kokkinakis D. and Malm, M. (2011). Character Profiling in 19th Century Fiction. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage – Hissar*, Bulgaria, pp. 70–77.

Manicheva, E., Petrova M., Kozlova E., and Popova, T. (2012) The Compreno Semantic Model as Integral Framework for Multilingual Lexical Database. In Zock, M. and Rapp, R. (eds), *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012*, Mumbai: Curran Associates, Inc., pp. 215–29

Petrova, M. (2013). The compreno semantic model: the universality problem. *International Journal of Lexicography*, 27: 105–29. 10.1093/ijl/ect038

Selegey V. (2012). On Automated Semantic and Syntactic Annotation of Texts for Lexicographic Purposes In: Ruth Vatvedt Fjeld and Julie Matilde Torjusen. *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012*, Oslo, Department of Linguistics and Scandinavian Studies, University of Oslo.

Shklovsky, V. (1925). Art as a technique. In Shklovsky, V. (ed.), *Theory of Prose*. Moscow: Krug, pp. 7–20.

Notes

[1] Note also that example (5) clearly demonstrates the importance of correct anaphora resolution for the tasks of in-depth text research.

[2] OWL Web Ontology Language Overview (http:// www. w3. org/ TR/ 2004/ REC- owl- features- 20040210), accessed 30 October 2015.

[3] Resource Description Framework http:// www. w3. org/ RDF, accessed 30 October 2015.

[4] Russian canonical edition of the WP and its English translation differ in their structure. While the English translation follows the first edition and consists of fifteen books, the Russian canonical text is divided into four volumes, cf. the Wikipedia reference (https:// en. wikipedia. org/ wiki/ War_ and_ Peace).