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

(Anastasia Bonch-Osmolovskaya and Daniil Skorinkin)

Цель

- Данная статья – подготовительный этап проекта “Digital Tolstoy” (см.:http://tolstoy.ru/projects/tolstoy-digital/).
- **Цель проекта – создать «семантическое издание» произведений Л. Н. Толстого**, в котором будут распознаны и помечены не только слова, но и значения, факты, даты, цитаты, связи, контексты. Таким образом **у текста появится дополнительный понятный компьютеру слой — состоящая из определенных тегов разметка**, которая сможет облегчить работу над текстом (например, поиск связей между произведениями разных лет можно предоставить компьютеру вместо того, чтобы делать это вручную).

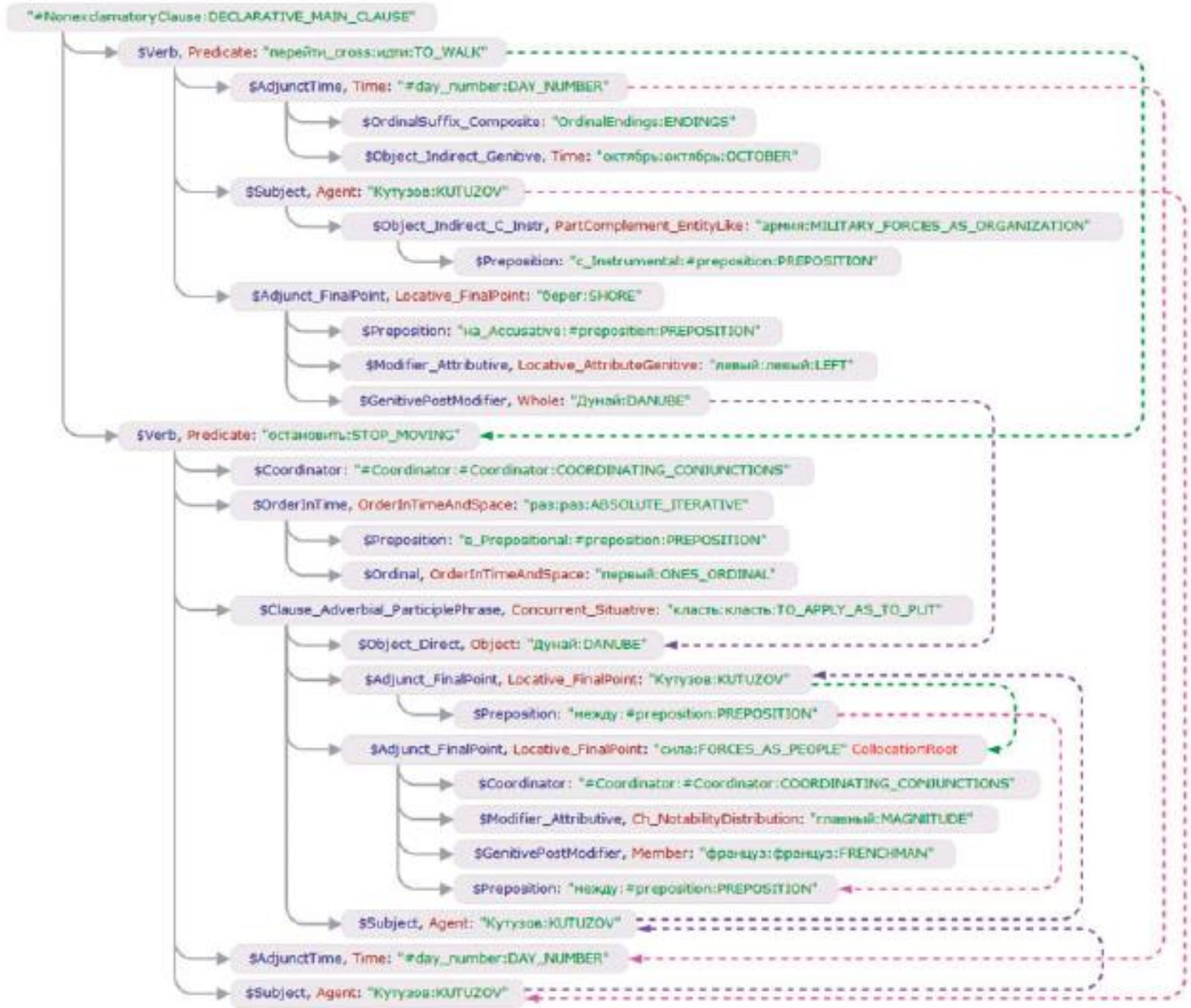
Цель

- Показать, как использование “advanced language analyser” может быть полезно для выделения из текстов объектов, которые позднее могут быть использованы для семантической разметки текста.
- Показать, как автоматическое извлечение «лексических схем», связанных с различными персонажами романа, может улучшить понимание «литературной техники» Толстого.

Средства и методы

- В работе используется технология ABBYY Compreno, которая трансформирует текст в совокупность синтаксическо-семантических деревьев, в которых отображена вся необходимая лингвистическая информация о тексте.

На следующем слайде представлено дерево, построенное на фразе: «28-го октября Кутузов с армией перешёл 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).


"#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: "с_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"

Принцип работы

Когда мы просим систему найти поддерево, которое включает в себя узел с семантическим классом «глаголы общения» и с как минимум двумя дочерними узлами «агент» и «адресат» (**то есть фрагменты, когда герои коммуницируют друг с другом**), он найдёт нам различные варианты. **Не только те, в которых «агент» и «адресат» будет выражен именами или фамилиями, но также местоимениями, социальным классом, профессией и т.д.**, так как это прописано в алгоритме.

- 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.
- 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’).

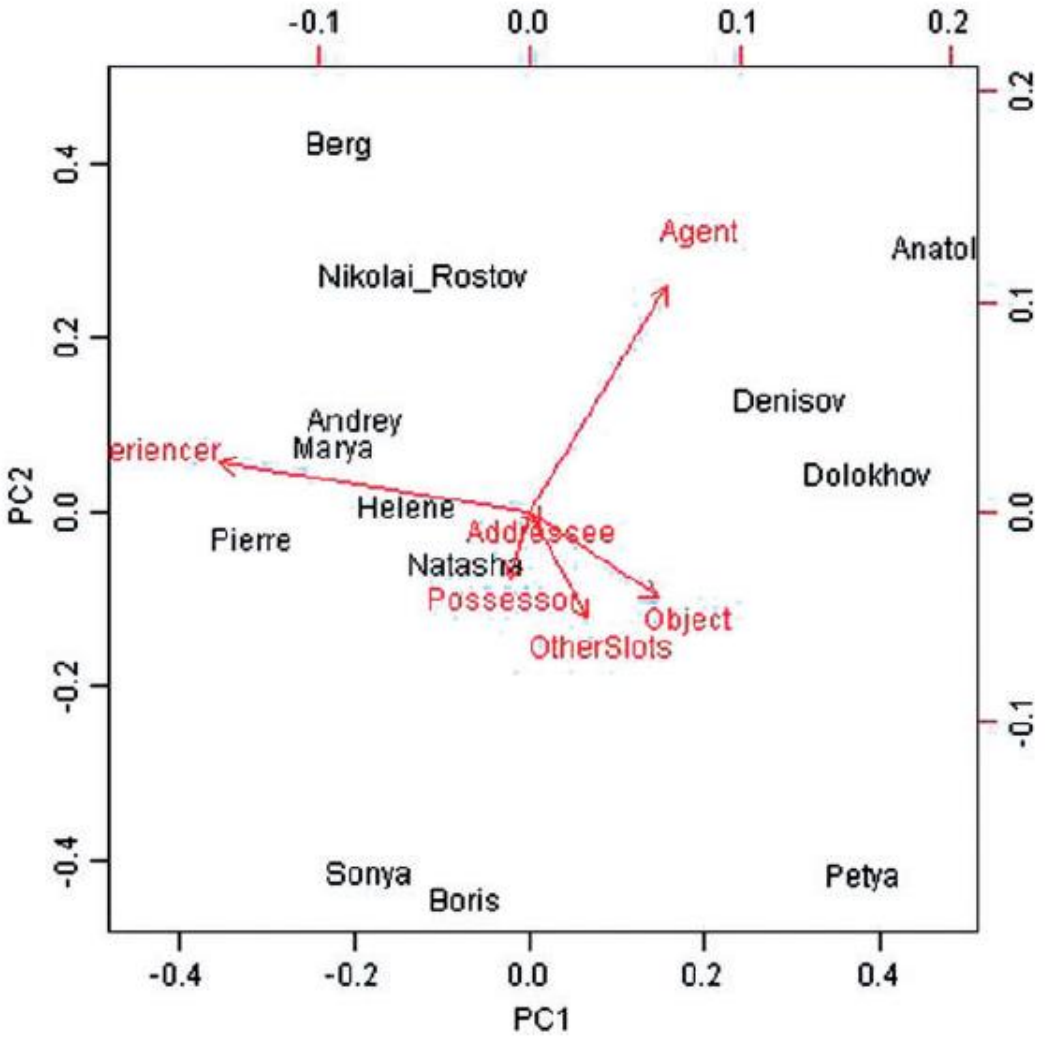
Эксперимент

Авторы попытались определить отношения между самыми выдающимися персонажами романа «Войны и мир» (14 героев), исходя из занимаемой ими синтаксической позиции в контексте глаголов с различной семантикой.

Используя упомянутый алгоритм COMPRENO они смогли создать окончательный список ролей – agent, object, experiencer, addressee, possessor – и проанализировать, как часто тот или иной персонаж выступает в той или иной роли.

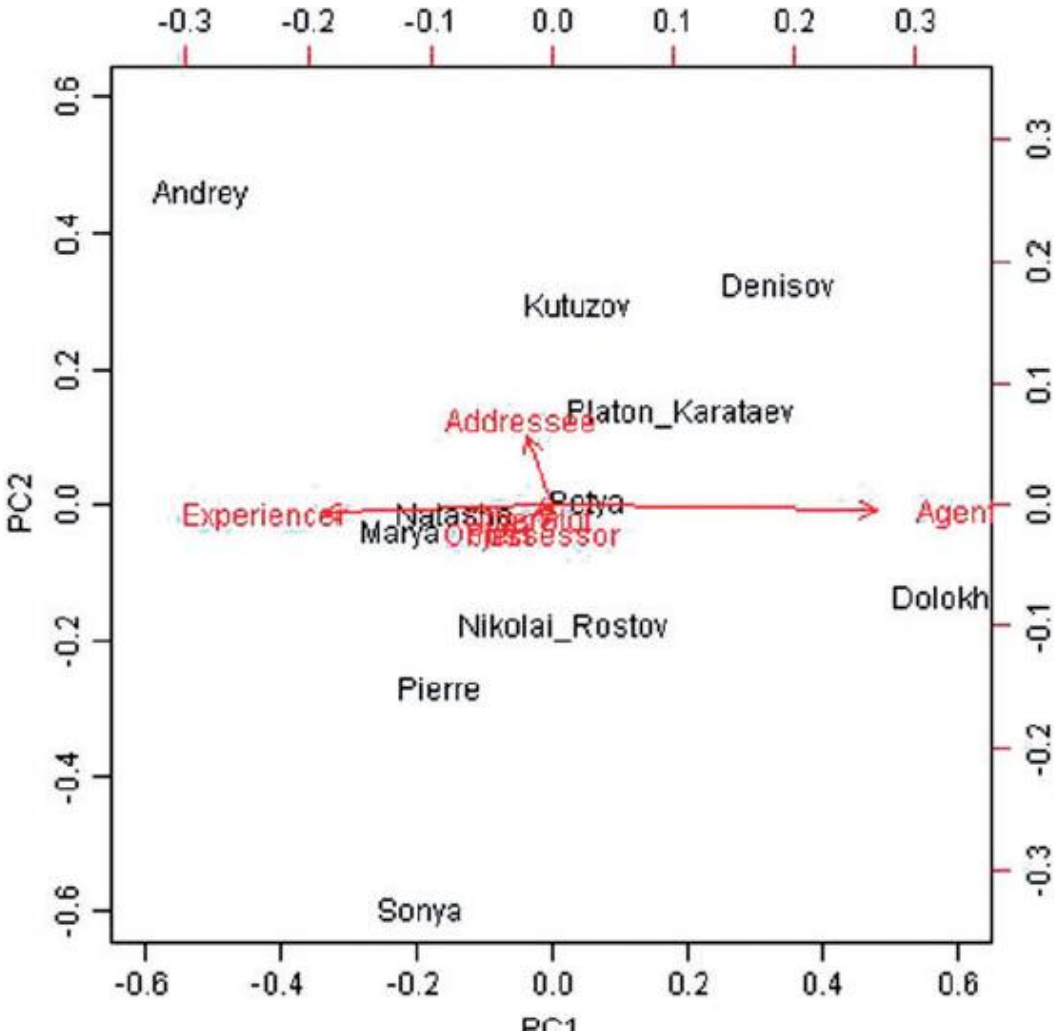
Визуализация распределения семантических ролей во втором томе романа «Война и мир»

- Мужские персонажи (Анатоль, Денисов, Долохов) на схеме представлены как более деятельные, что согласуется с сюжетом: многочисленные «мужские дела» и борьба за внимание женских персонажей.
- Наташа и Элен расположены рядом с зоной «адресата»: в романе на них направлены активные действия мужских персонажей.
- Андрей Болконский расположен рядом со своей сестрой на линии “experiencer”: переживающий смерть жены и раздумывающий над собственной жизнью, князь Андрей становится менее деятельным, ближе к размышляющей и переживающей Марье.



Визуализация распределения семантических ролей в четвёртом томе романа «Война и мир»

- Смертельно раненный князь Андрей, ждущий смерти и размышляющий о своей жизни, оказывается расположен в зоне “experiencer”.
- Марья и Наташа, бывшие рядом с князем Андреем в последние минуты его жизни, не случайно расположены рядом.



Итоги и пути развития

- На возможную критику авторы статьи отвечают в заключении. Они признают, что методика, на данном этапе разработки, смогла предоставить нам лишь количественные данные, подтверждающие факты, которые были очевидными для читателей и критиков, но вместе с этим не были представлены в тексте эксплицитно. Однако они убеждены, что будущие исследования такого типа могут многое нам сказать о стиле автора (не говоря о преимуществах, которые может дать семантическая разметка текстов в принципе).
- Исследователи планируют разбить роли на подтипы, чтобы выделить разные виды «активного» и «переживающего» поведения, и обратить особое внимание на прямую речь, попытавшись понять, действительно ли каждый персонаж обладает уникальным стилем речи.