# **Онтология маркеров академического стиля**

Набор маркеров представляет собой неструктурированную информацию, а создание онтологии при обработке неструктурированных данных имеет ряд преимуществ. Онтология позволяет наиболее целостно представить сведения об изучаемой предметной области. Данные, описанные в онтологии, представляют собой материал, представленный в единой форме, с описанием одних терминов с помощью других, отражая различные типы взаимосвязей между ними, что делает процесс восприятия данной предметной области легче.

Было решено, что основная онтология маркеров стиля не может существовать автономно. Поэтому онтологию маркеров стиля стоит рассматривать в терминах лингвистики, отражающих лингвистические понятия, в рамках которых и существуют маркеры стиля, помогая обеспечивать понятийную базу. На данный момент готовых онтологий лингвистики не было найдено, поэтому создавать ее придется самостоятельно. Так, онтология маркеров стиля будет связана с понятиями лингвистики, что позволит ей существовать в определенном контексте.

Разработка онтологии производилась в версии Protégé 4.3.0.

Проектирование онтологии включает в себя ряд шагов:

Шаг 1. Определение области и масштаба онтологии.

Для этого нужно ответить на ряд основных вопросов:

1. Какую область будет охватывать онтология?

2. Для чего она будет использовать использоваться?

3. На какие типы вопросов должна давать ответы спроектированная онтология?

4. Кто будет использовать и поддерживать онтологию?

Эти вопросы необходимы для того чтобы ограничить масштаб модели.

Итак, охватываемая область – письменный академический английский язык. Главная цель проектируемой онтологии – представить онтологию маркеров стиля в виде классификации со всеми вытекающими связями и отношениями. Онтология должна давать определение принадлежности атрибутов онтологии маркеров стиля определенному уровню грамматики английского языка.

Шаг 2. Рассмотрение вариантов повторного использования существующих онтологий

Возможно, что онтология лингвистики уже создана. Следует тщательно проверить доступные источники на существование подобной онтологии. Анализ доступных источников показал, что подобных работ ранее сделано не было.

Шаг 3. Перечисление важных терминов в онтологии.

В качестве основных терминов в работе будут использоваться термины, предложенные экспертами: nominalization, noun, noun with abstract suffix, noun with -or suffix, personal stance, verb, desemanticised verbs, verbs of broad abstract semantics, tenses, future, past, present, voice, passive voice, adverb, intensifying adverbs, attributes, prepositive attributes, postpositive attributes, cohesiveness, complex, archaisms, complex prepositions, logic connectors.

За основу для создания онтологии была взяты выделенные основные с точки зрения экспертов аспекты академического английского и остальные важные термины, перечисленные в предыдущем шаге.

Наиболее важные аспекты:

1. ***Nominalization***
2. ***Personal Stance***
3. ***Verb***
4. ***Adverb***
5. ***Attributes***
6. ***Cohesiveness***

Шаг 4. Определение классов и иерархии классов.

Ниже представлена иерархия классов, основанная на имеющихся терминах.

**Noun (Существительное):**

* noun with abstract suffix (существительное с абстрактным суффиксом);
* noun with *-or* suffix (существительное с суффикcом -or).

**Personal Stance (Личное отношение):**

* personal pronouns (личные местоимения);
* demonstrative pronouns (указательные местоимения).

**Verb (Глагол):**

* desemanticised verbs (смысловой глагол);
* verbs of broad abstract semantics (глагол широкой абстракции) ;
* tenses (времена);
* voice (залог).

**Adverb (Наречие):**

* intensifying (усилительные наречия).

**Attributes (Атрибуты):**

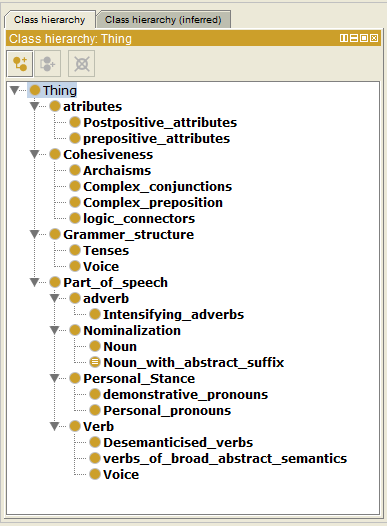
* prepositive attributes (препозитивные атрибуты);
* postpositive attributes (постпозитивные атрибуты).

**Cohesiveness (Логичность):**

* complex conjunctions (составные союзы);
* archaisms (архаизмы);
* complex prepositions (сложные6 предлоги);
* logic connectors (логические связки).

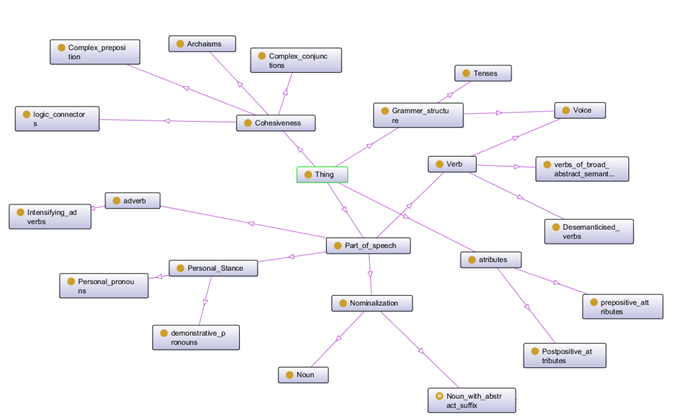
Таким образом, получилась иерархия, где каждый маркированный элемент является подклассом вышестоящего выделенного понятия (рис.1). Самым распространенным типом отношений, использующимся во всех онтологиях, является отношение категоризации, то есть отнесение к определенной категории, что мы и можем наблюдать.

Понятия, описанные выше, являются наиболее заметными и важными в данной онтологии, однако это не значит, что они описывают всю предметную область. Так, например, глядя на классы иерархии Noun (Существительное), Verb (Глагол), Demonstrative pronouns (Указательные местоимения), Personal pronouns (Личные местоимения) и Adverb (Наречие) можно сказать, что над ними стоит класс под названием Part of speech (Часть речи), который не отображен в иерархии понятий на данный момент. Именно тут можно начать говорить об онтологии нижнего уровня, онтологии лингвистики.



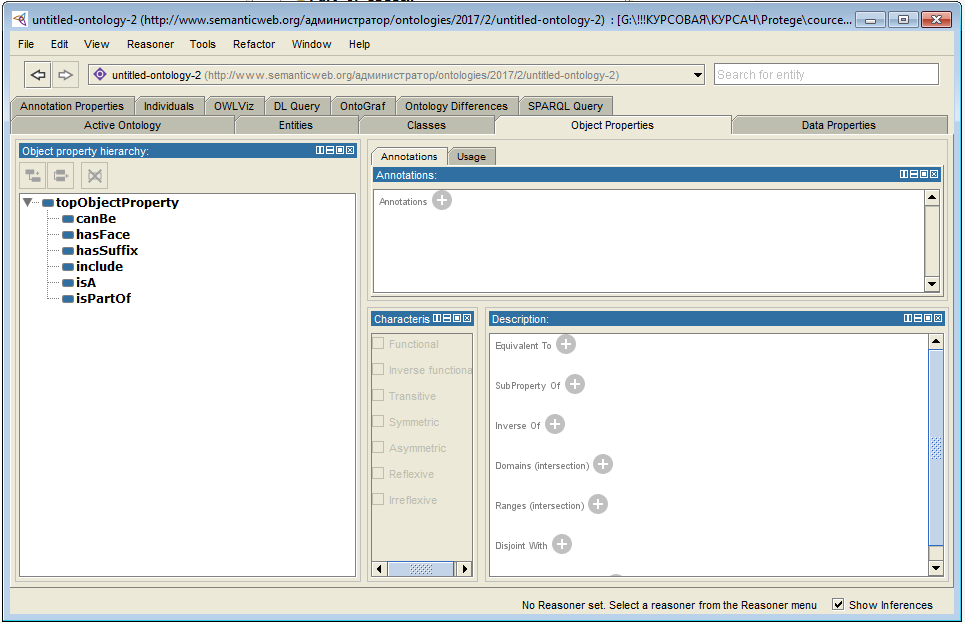
**Рисунок 1. Иерархия классов**

После анализа данных было решено добавить дополнительные классы, лежащие в основе описанных выше понятий. Tenses (Времени) и Voice (Залог) были представлены в качестве подкласса Grammar Structure (Грамматическая структура), поскольку они непосредственно имеют отношение к этому уровню лингвистики. При необходимости можно будет подняться на более высокий уровень категоризации, выделив дополнительно такие классы как Part of word (Часть слова), чтобы включить туда класс Noun with abstract suffix (Существительное с абстрактным суффиксом). Граф, на основе созданных классов, представлен на рисунке 2.



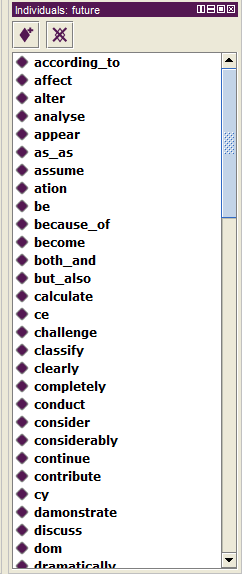
**Рисунок 2. Онтологический граф**

На данный момент создана основная иерархия классов, нужно создать свойства и выделить экземпляры.

Для обозначения отношений между понятиями были введены следующие свойства объектов (рис 3):

**Рисунок 3. Свойства объектов**

На вкладке Individuals были созданы экземпляры классов и привязаны соответствующие свойства (рис. 4).



**Рисунок 4. Экземпляры класса**

Результатом работы является онтология маркеров стиля, позволившая сделать из маркеров стиля упорядоченную систему, которой можно управлять. Построение онтологии не является конечной целью, она является частью большего проекта, выполняя в нем функцию генерации правил поиска маркеров стиля при аннотировании текстов.

В процессе анализа существующих данных было выявлено, что онтология может изменяться в процессе работы с ней, могут добавляться некоторые понятия, свойства, экземпляры, поскольку эксперты не закончили анализ данной предметной области.

При сравнении программных средств для реализации поставленной задачи был выбран редактор онтологий Protégé, поскольку он обладает открытым исходным кодом и широким функционалом, способным осуществить поставленные задачи. Помимо этого, он предоставляет возможности по экспорту и импорту созданных онтологий в различные форматы.

На стадии проектирования были проанализированы данные, представленные экспертами, после чего построена иерархия классов с последующей категоризацией и типизацией. Помимо выделенных терминов предметной области, которые были связаны лишь с маркерами стиля, были выделены дополнительные классы и свойства, связанные непосредственно с областью лингвистики, чтобы онтология маркеров стиля существовала в конкретной предметной области, а не отдельно от нее.

Так как проектирование онтологии является итеративным процессом, возможны добавления связей между экземплярами и классами, с целью получения лучшего эффекта при аннотировании текстов корпуса.

Список литературы

1. General Architecture for text engineering// Система обработки естественного языка 1995. URL: https://gate.ac.uk/download/ (дата обращения: 25.03.2017).

2. Соснина Е.П. Введение в прикладную лингвистику: учебное пособие. Ульяновск: УГТУ, 2012. 110 с.

3. Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. СПб: НИУ ИТМО, 2013. 72 с.

4. Захаров В.П. Корпусная лингвистика: учебное пособие. СПб., 2005. 48 с.

5. Кузнецов О.П., Суховеров В.С., Шипилина Л.Б. Онтология как систематизация научных знаний: структура, семантика, задачи: учебное пособие.

6. Константинова Н.С., Митрофанова О.А. Онтологии как системы хранения знаний: учебное пособие. Спб: СпбГУ, 2007. 54 с.

7. Абдулраб Х. Онтологический инжиниринг / Х. Абдулраб, Э.А. Бабкин // Вопросы образования. – 2007. -№2. – С.3-14.

8. Protégé // Редактор онтологий. URL: http://protege.stanford.edu/ (дата обращения: 25.04.2017).

11. Luyckx K., Daelemans, W. Shallow text analysis and machine learning for authorship attribution // Computational Linguistics in the Netherlands 2004: selected papers from the Fifteenth CLIN Meeting / van der Wouden T. [Ed.], e.a., Utrecht, LOT, 2005, – C. 149-160.

12. Scholz T., Conrad S. Style Analysis of Academic Writing // Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems, Proceedings. NLDB 2011, Alicante, Spain, June 28-30, 2011. – C. 246-249.

13. Kent D. Bimson , Richard D. Hull Unnatural Language Processing: Characterizing the Challenges in Translating Natural Language Semantics into Ontology Semantics// Semantic Web 2015, – C. 119-135.

14. Strinyuk S. A., Shuchalova Y., Lanin V. Academic Papers Evaluation Software, in: Application of Information and Communication Technologies (AICT), 2015 9th International Conference on, 14-16 Oct. 2015. Rostov-on-Don: IEEE, 2015. doi P. 506-510.

15. Загорулько Ю.А. Построение порталов научных знаний на основе онтологий // Вычислительные технологии. – 2007. – Т. 12.– Спецвып. 2. – С. 169–177.