# **Представление онтологического ресурса для классификации выделенных атрибутов научного текста**

Корпусная и компьютерная лингвистика с недавних пор стали очень популярными направления прикладной лингвистики, поскольку способны решить проблемы лингвистики с помощью вычислительной техники. Различные программы анализируют корпуса текстов и помогают экспертам делать те или иные заключения, выводы, выдвигать гипотезы. В рамках данной работы именно корпусная лингвистика используется в качестве основного подхода. На основе собранного корпуса идут исследования, проводимые при помощи специального программного обеспечения GATE [1]. Функционал данной программы обширен: программа выполняет функцию поиска маркеров в тексте, разделяет текст на токены, каждому из которых предписывается его тип, осуществляет семантическую аннотацию и т.д.

В предыдущем докладе была представлена онтология, однако, так как создание онтологии итеративный процесс, она была изменена.

Данная онтология позволит решить спектр задач. 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. Возможность осуществлять интеллектуальный поиск также имеет большое значение в изучении предмета, поскольку при необходимости ознакомиться с одним из аспектов будет получена полная информация, хранящаяся в онтологии, со всеми связями и отношениями. Онтология – отличный инструмент для человека, который исследует академический английский. Она может быть использована для анализа собственной статьи: насколько статья удовлетворяет требованиям академического английского, какие маркеры в ней присутствуют и т.д. Онтология будет связана с понятиями лингвистики, что в дальнейшем позволит автоматически генерировать правила поиска.

В ходе поиска все-таки была найдена онтология лингвистики GOLD, с которой интегрировалась онтология маркеров стиля. GOLD (General Ontology for Linguistic Description) – онтология для описательной лингвистики, которая предоставляет наиболее основные категории и взаимоотношения, использующиеся в научном описании естественного языка, в формализованном виде. GOLD нацелена на сбор знаний наиболее квалифицированных лингвистом, и может рассматриваться как попытка кодифицировать общие знания данной области.

Таким образом, получилась иерархия, где каждый маркированный элемент является подклассом вышестоящего выделенного понятия. Эти понятия являются наиболее заметными и важными в данной онтологии, однако это не значит, что они описывают всю предметную область. Теперь вернемся к онтологии нижнего уровня – онтологии Лингвистики. Основываясь на анализе, проведенном по онтологии GOLD было решено строить онтологию маркеров на основе части онтологии GOLD, поскольку она наиболее полно описывает предметную область лингвистики. Было решено не использовать всю онтологию GOLD, потому что онтология описывают данную область максимально полно, пытаясь осветить все понятия, включая устный язык, разговорный, знаковый, что не является предметом данной работы. Поэтому чтобы избежать избыточности онтология маркеров стиля будет основываться только на особенностях письменного английского.

Теперь вернемся к иерархии классов онтологии лингвистики (рис.2.3). Как было описано ранее, наиболее интересным классом, с точки зрения онтологии маркеров стиля, является класс Abstract. Все аспекты, выделенные экспертами выше, принадлежат подклассам LinguisticProperty и LinguisticUnit. Именно этот кусок иерархии классов ляжет в основу онтологии маркеров стиля в качестве онтологии лингвистики.

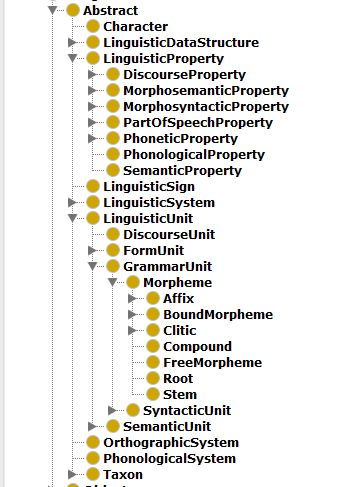
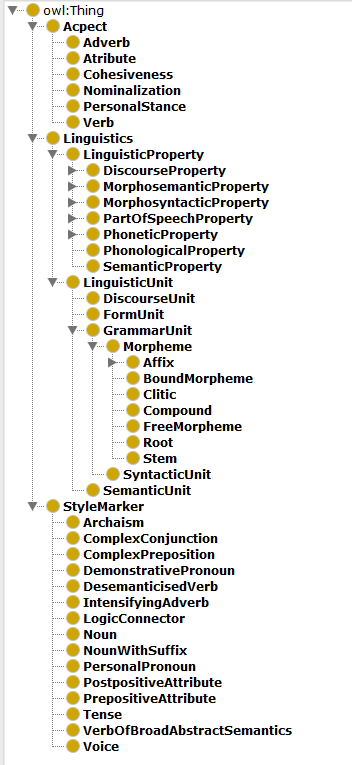
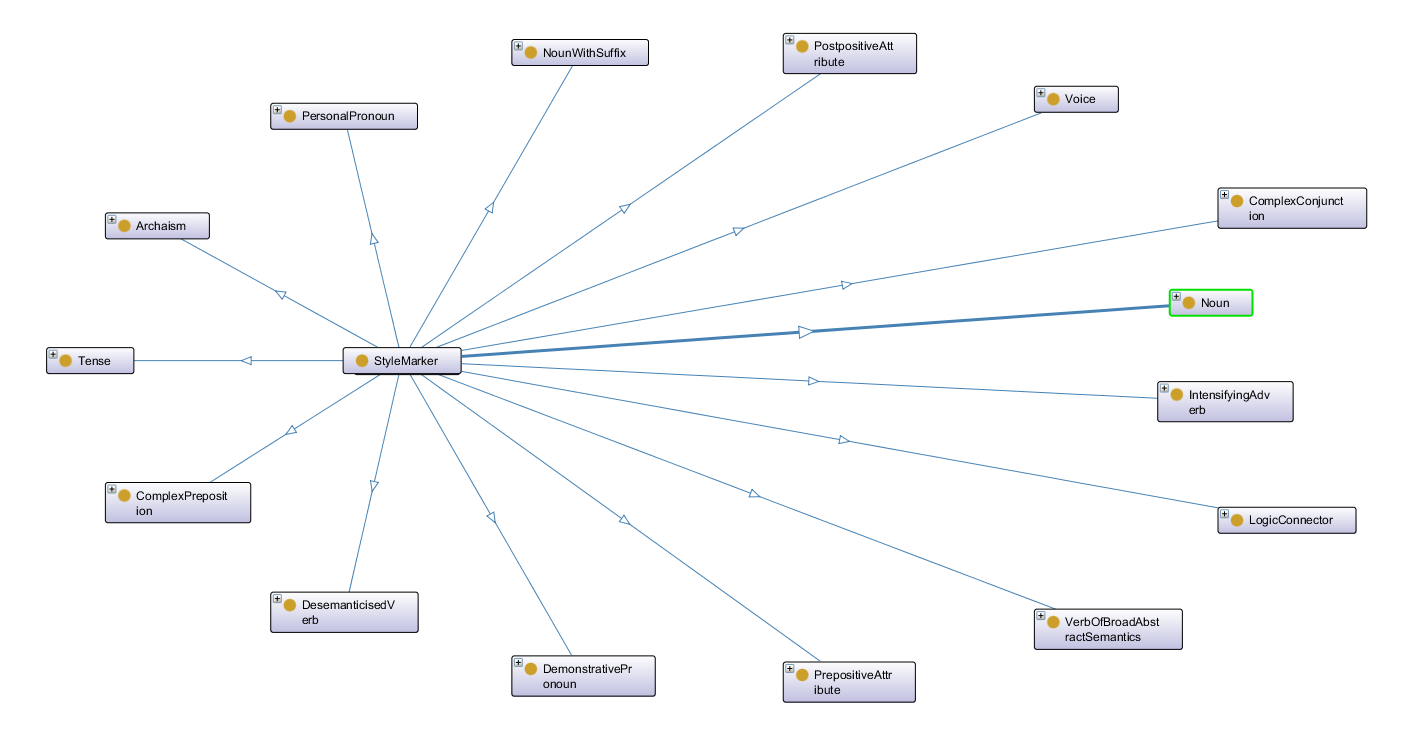


Рисунок 1. Наиболее важные классы для онтологии маркеров стиля

В ходе анализа было решено создать три основные класса: Linguistics – термины лингвистики, Aspect – аспекты, выделенные экспертами и StyleMarker – маркеры стиля. Иерархия классов представлена рис. 2.



**Рисунок 2. Получившаяся иерархия классов**

Графы, на основе созданных классов, представлены на рисунках ниже. 

**Рисунок 3. Онтологический граф класса StyleMarker**

Отношения типа «класс-подкласс» выглядят следующим образом:

*SubClassOf (: StyleMarker : Noun)*

*SubClassOf (:StyleMarker: Tense)*

Данный класс StyleMarker отражает конкретные маркеры стиля, которые являются чертами письменного академического английского. Он состоит из понятий описанных выше. Подклассы содержат в себе экземпляры, которые представляют собой маркер стиля. Маркеры не связаны никакими связями между собой, но имеют связи с выражающими их аспектами. Рисунок с отображением этих связей будет представлен на шаге создания отношений.

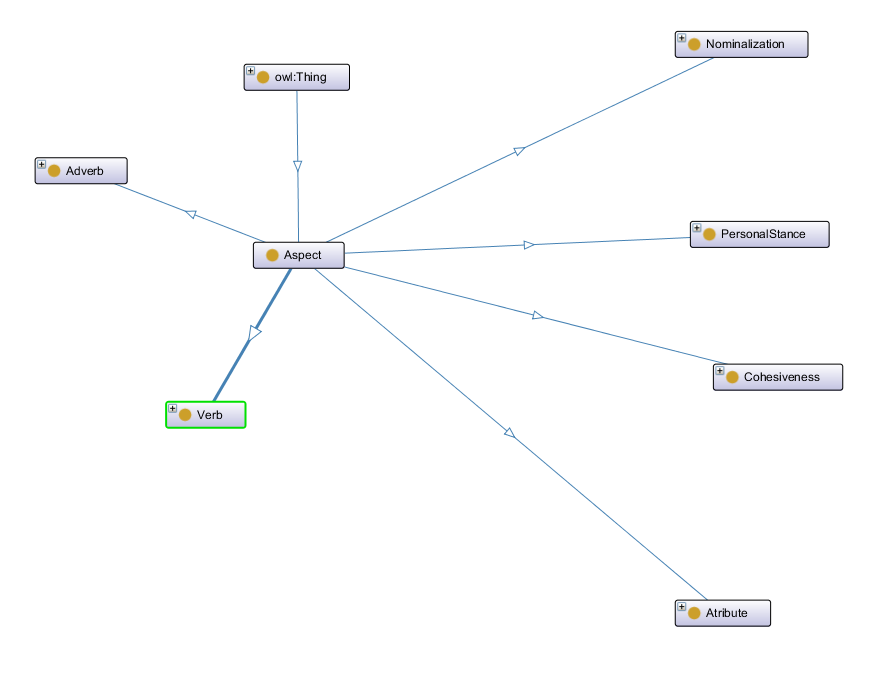


Рисунок 4. Онтологический граф класса Aspect

Отношения типа «класс-подкласс» выглядят следующим образом:

*SubClassOf (: Aspect : Verb)*

*SubClassOf (:Aspect: Nominalization)*

Эксперты выделили шесть наиболее важных по их мнению аспектов академического английского. Как и маркеры, аспекты не связаны свойствами между собой, но будут связаны как с понятиями класса StyleMarker, так и с Linguistics. На данный момент отображена лишь иерархическая зависимость классов и подклассов.

На рис. 5 отображена иерархия классов, взятая из онтологии лингвистики GOLD. Как было описано ранее после анализа имеющихся понятий и их возможном месте в имеющейся онтологии, было решено взять классы LinguisticProperty и LinguisticUnits, в качестве лингвистической основы.

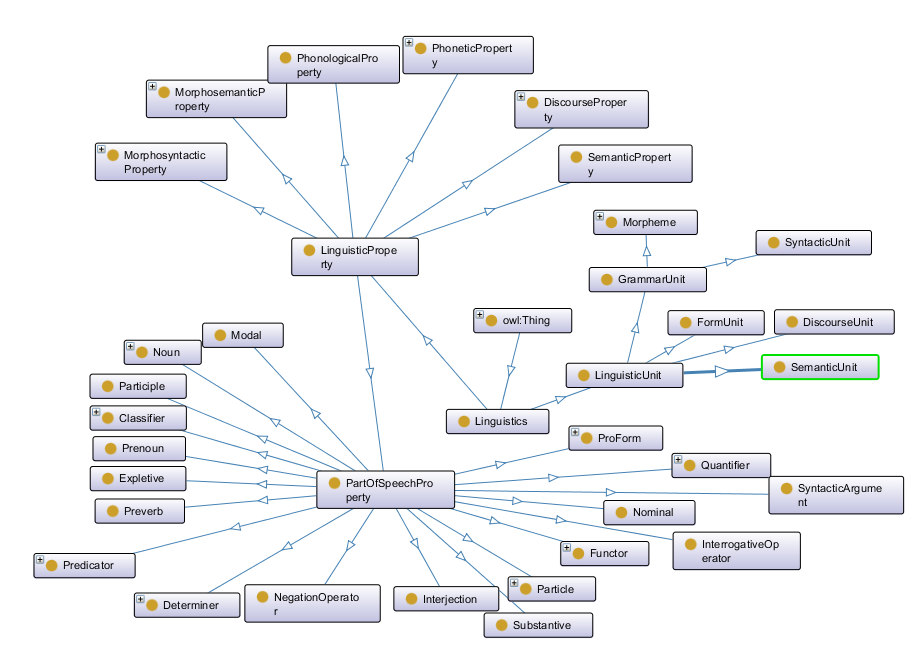
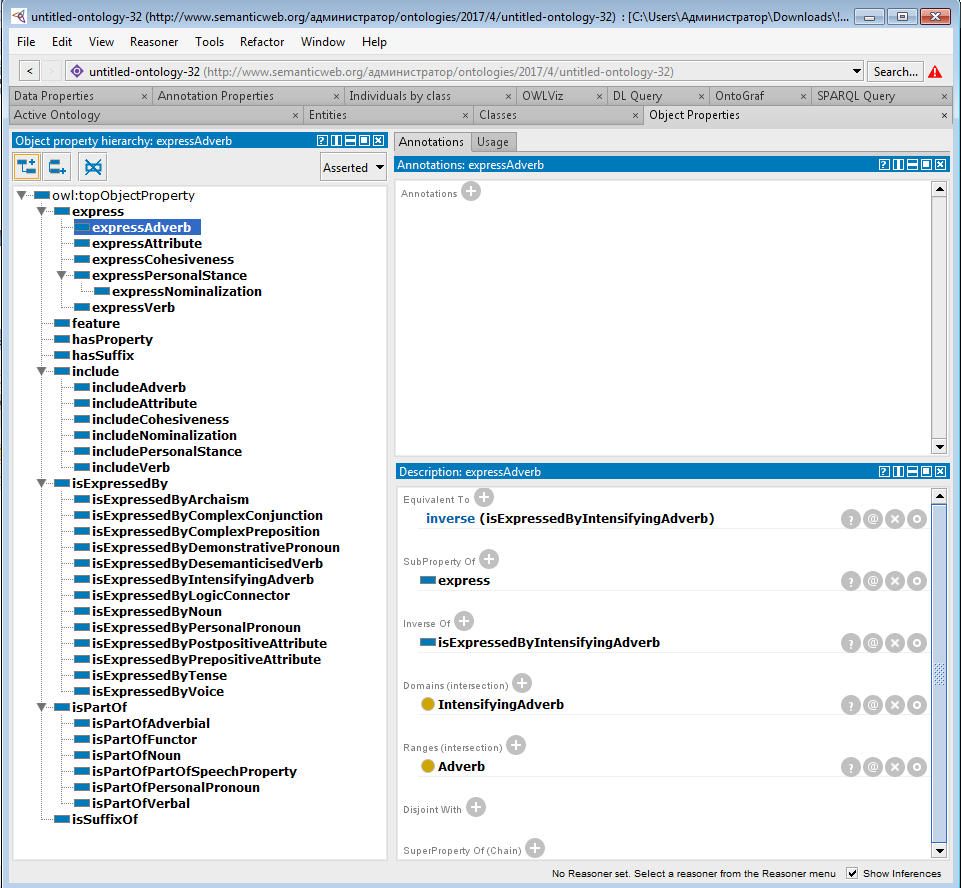


Рисунок 5. Онтологический граф класса Linguistics

Шаг 5. Теперь определим связи между созданными понятиями (классами), помимо связи «класс – подкласс». Было решено соединить свойствами *express* и обратным ему свойством *isExpressedBy* понятия класса Aspect и StyleMarker. Логика такова: каждый класс аспектов выражается тем или иным набором классов/классом с соответствующими маркерами стиля. И наоборот, каждый класс маркеров стиля выражает определенный аспект письменного академического языка. На рис. 6 изображены созданные свойства.

 **Рисунок 6. Свойства объектов**

На вкладке Individuals были созданы экземпляры классов и отнесены определенным классам. Эти экземпляры являются частными проявлениями маркеров стиля в письменном академическом английском. Так как созданные экземпляры представляли собой общий массив экземпляров без принадлежности, необходимо было каждый из экземпляров отнести к экземплярам конкретного класса (рис. 7).

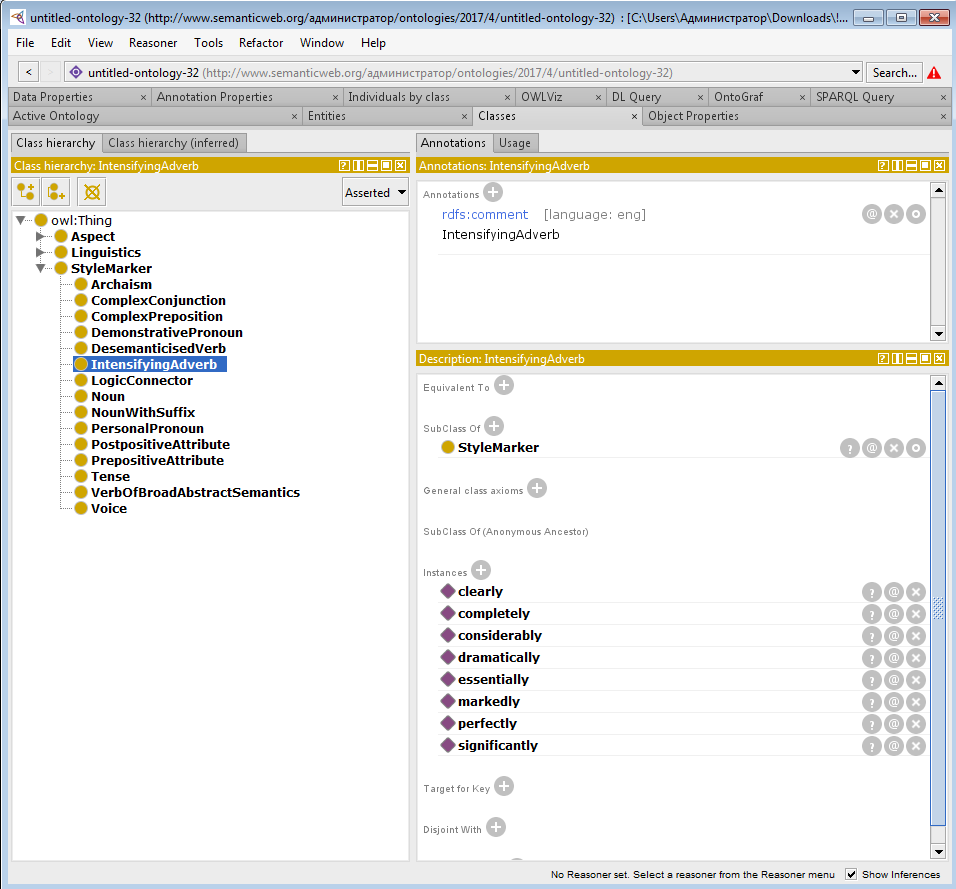


Рисунок 7. Массив экземпляров классов

Не все классы имеют экземпляры. Ниже представлена таблица с названием класса и существующими в нем экземплярами (табл.1).

**Таблица 2.1. Экземпляры классов**

| Название класса | Экземпляры |
| --- | --- |
| PersonalPronoun | I, we, he, she |
| DesemanticisedVerb | be, become, seem, remain, grow, consider |
| VerbOfBroadAbstractSemantics | be, exist, have, appear, occur, alter, continue, contribute, discuss, involve, investigate, conduct, consider, illustrate, assume, find, calculate, demonstrate, identify, analyse, support, challenge, examine, affect, provide, include, classify, establish |
| Tense | Future, past, present |
| IntensifyingAdverb | clearly, dramatically, completely, considerably, essentially, significantly, markedly, perfectly |
| AbstractSuffix | -ment, -ion, -ation, -ition, -tion, -sion, -f, -ness, -ce, -cy, -ity, -dom, -th, -ery, -ry, -ise, -ice, -hood, -ics, -ship |
| ComplexConjunction | not\_merely, but\_also, both\_and, as\_ as, neither \_nor, the\_the, not\_so\_as |
| Archaism | thereby, therewith, hereby |
| ComplexPreposition | throughout, within, in\_accordance\_with, instead\_of, according\_to, because\_of, due\_to, regardless\_of |
| LogicConnector | since, therefore, follow, so, thus, lead\_to, result\_ in |

Теперь следует рассмотреть свойства, созданные для связи между классами Aspect, StyleMarker и Linguistic.

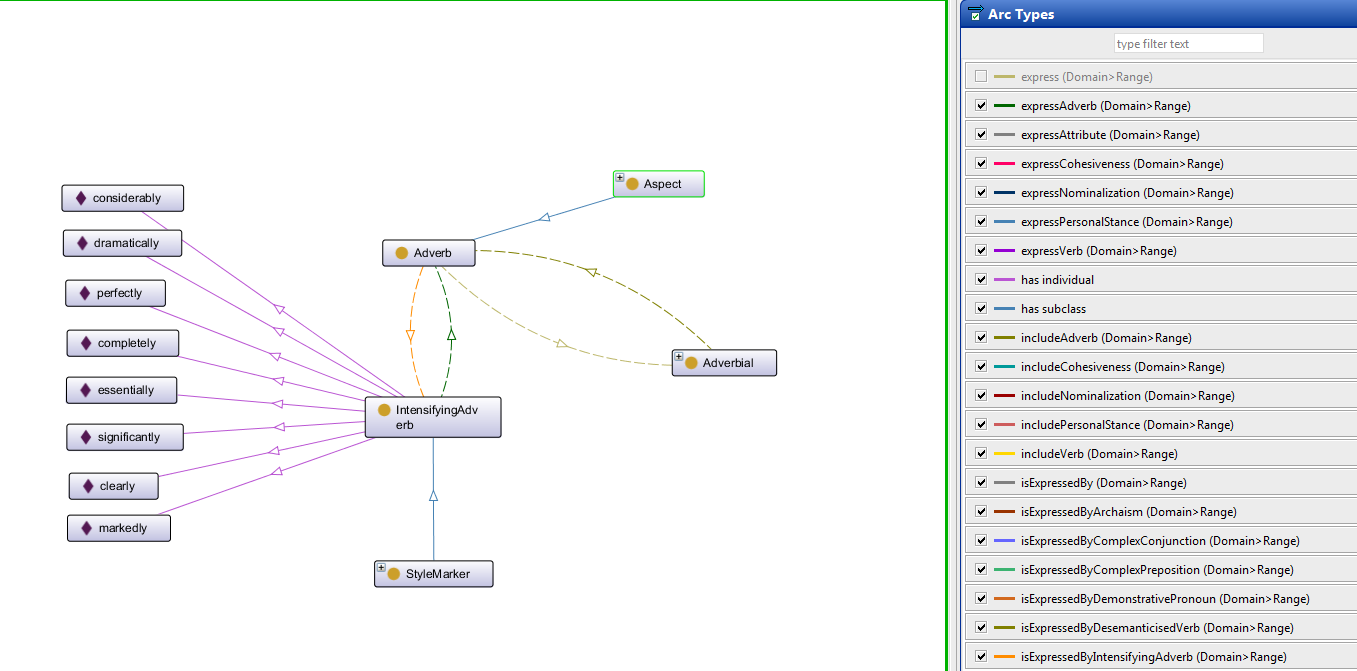


Рисунок 8. Онтологический граф с отношениями и экземплярами аспекта Adverb

Граф, отображенный на рис. 8 отображает связи аспекта Adverb. Из графа видно, что данный класс является подклассом Aspect, он связан с подклассом IntensifyingAdverb класса StyleMarket. Свойства получились следующие: Adverb *isExpressedBy* IntensifyingAdverb, что означает, что аспект Наречие *выражается* маркером стиля Усиливающее наречие*,* иобратное свойство *isExpressedBy – express*. Помимо уже описанных свойств можно увидеть свойство *hasIndividual* и соответствующие данному классу маркеров стиля экземпляры. Что касается связи онтологии лингвистики и аспектов, тут создано свойство *include* и *isPartOf* , представляющее собой: Adverb *isPartOf* Adverbial и обратное предложение со свойством *include*.

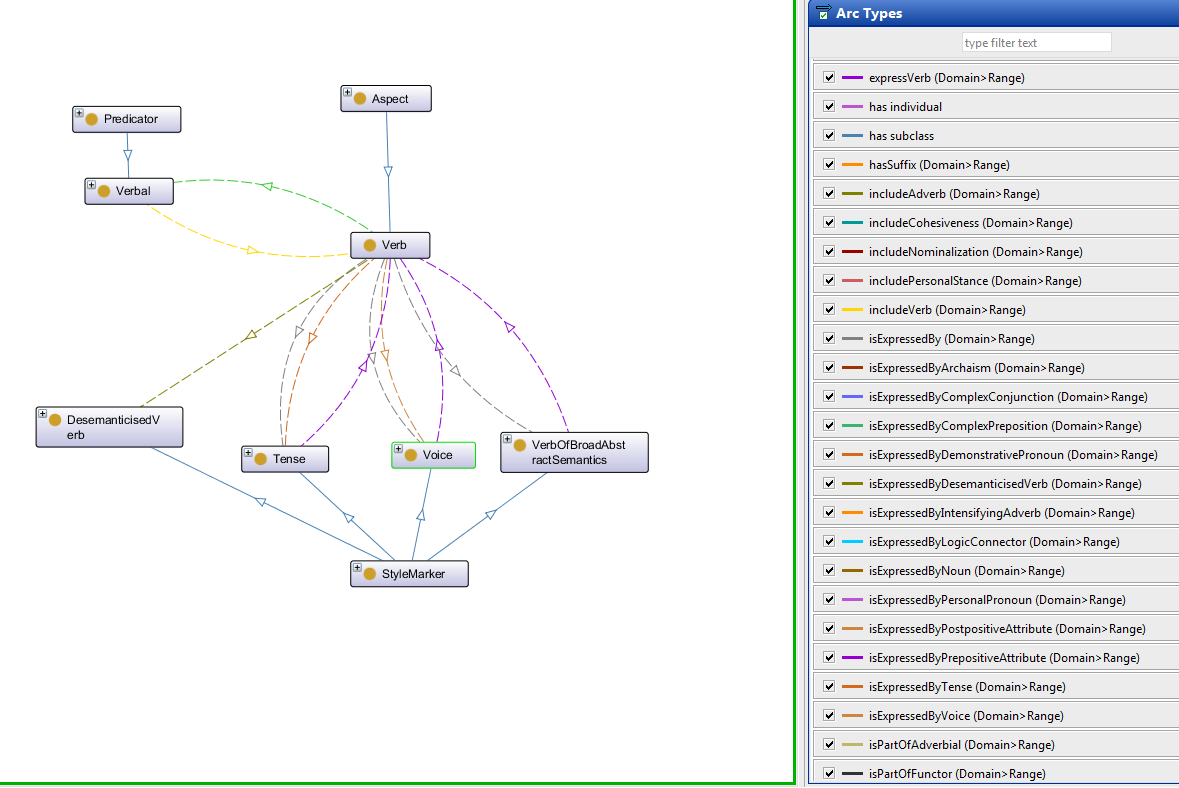


Рисунок 9. Онтологический граф с отношениями аспекта Verb

Граф, отображенный на рис. 9 отображает связи аспекта Verb. Из графа видно, что данный класс является подклассом Aspect, он связан с подклассами Voice, Tense, DesimanticisedVerb класса StyleMarket. Свойства получились следующие: Verb *isExpressedBy* Voice, Tense, DesimanticisedVerb что означает, что аспект *Глагол выражается* маркерами стиляВремя, Залог и Сысловым глаголом иобратное свойство *express*. На данном графе не отображено свойство *hasIndividual* и соответствующие данному классу маркеров стиля экземпляры, посколько это бы сильно нагрузило граф, но имеющиеся экземпляры можно увидеть в табл. 1. Что касается связи онтологии лингвистики и аспектов, тут создано свойство *include* и *isPartOf*, представляющее собой: Verb *isPartOf* Verbal и обратное предложение со свойством *include*. Такое свойство действует для связи всех аспектов с понятиями лингвистики.

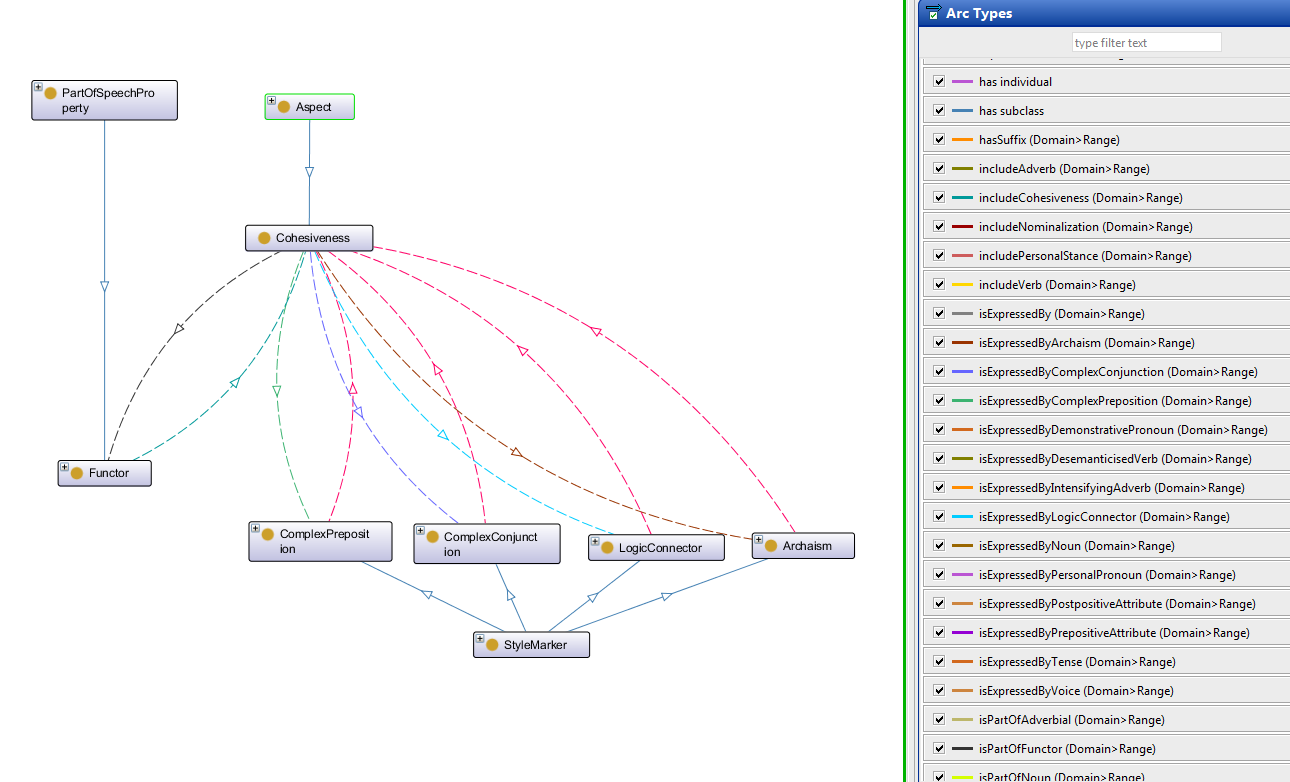


Рисунок 10. Онтологический граф для аспекта Cohesiveness

На рис. 10 онтологический граф аспекта Cohesiveness, отражающий связи между маркерами стиля и понятиями лингвистики. Данный аспект выражается подклассами класса StyleMarker ComplexСonjunction, ComplexPreposition, LogicConnector и Archaism. Данный аспект является частью понятия лингвистики Functor.

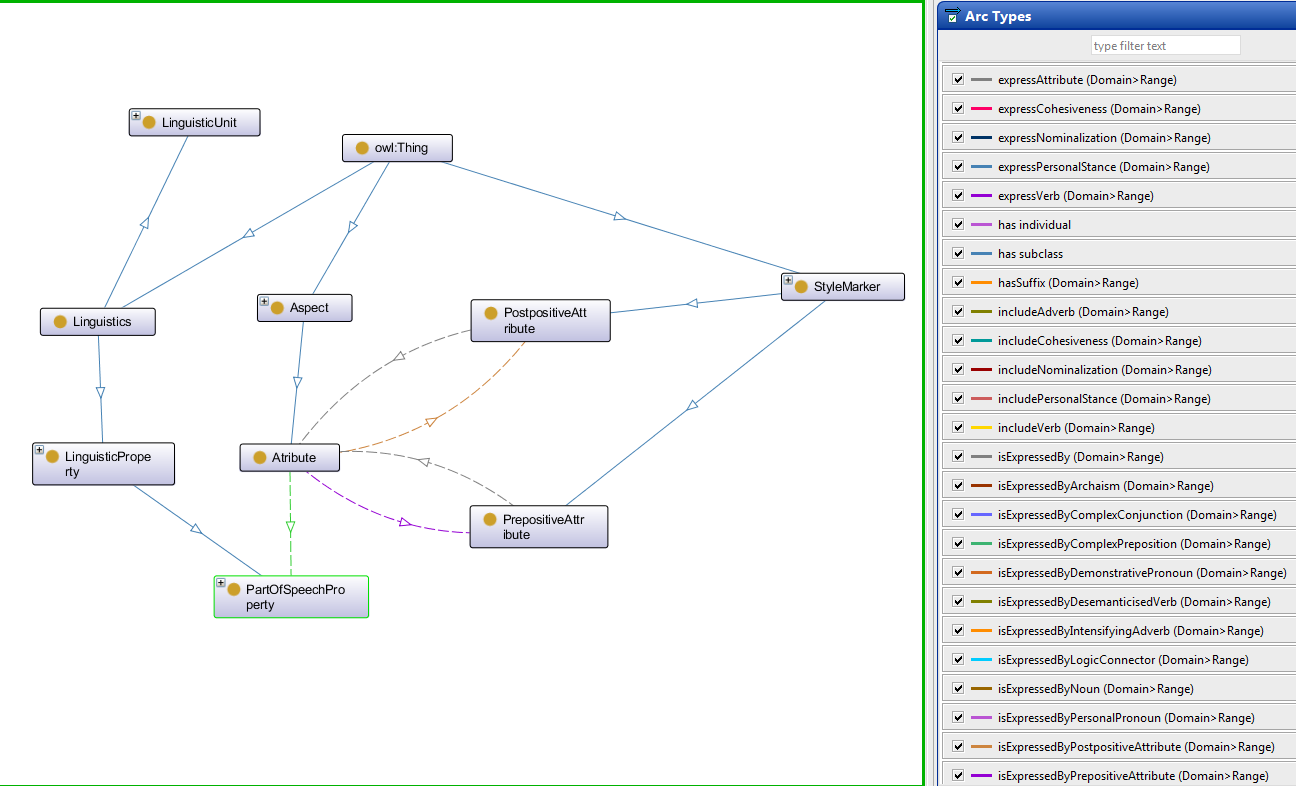


Рисунок 11. Онтологический граф с отношениями аспекта Attribute

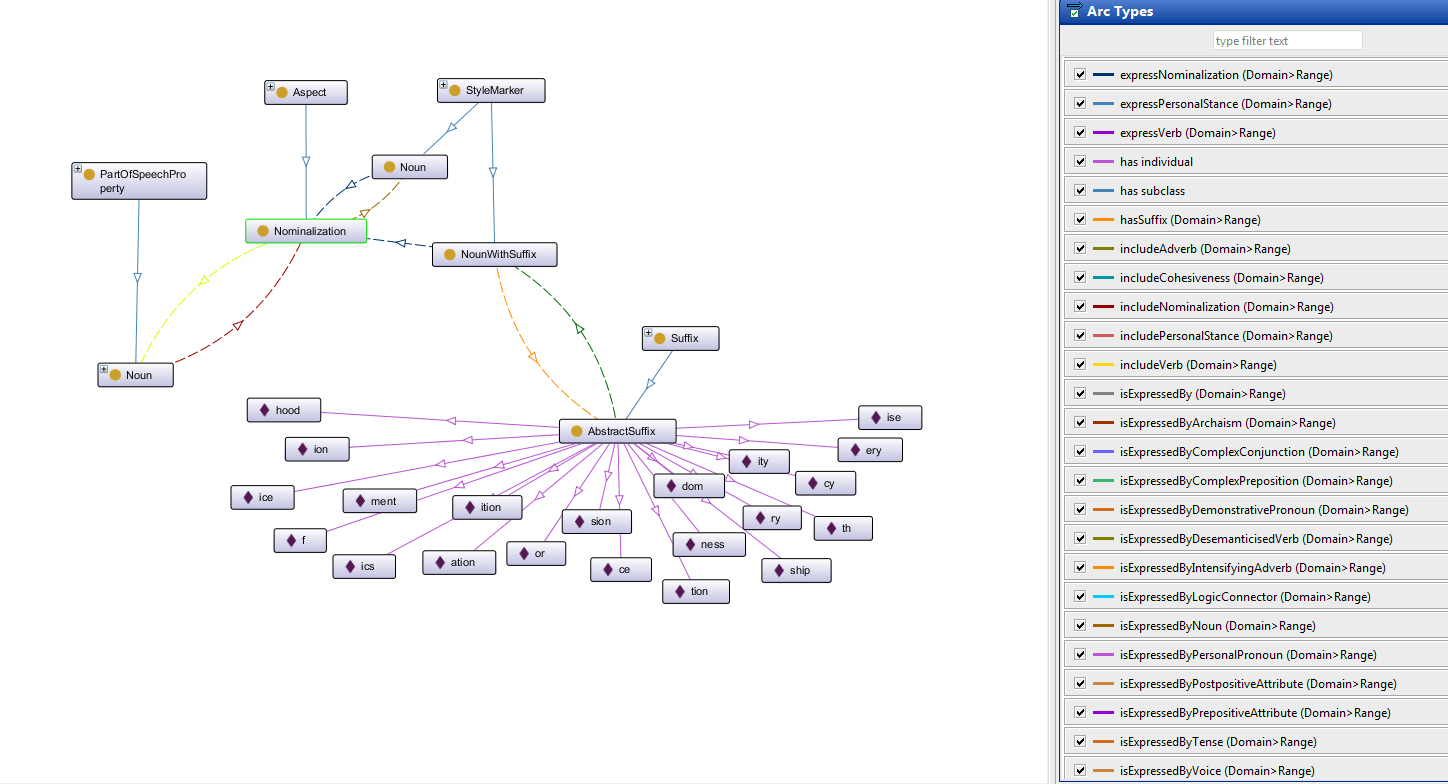


Рисунок 12. Онтологический граф с отношениями и экземплярами аспекта Nominalization

Аспект Nominalization выражаются маркерами Noun и NounWithAbstractSuffix. NounWithAbstractSuffix в свою очередь имеет свойство hasSuffix, ссылающееся на абстрактные суффиксы, которые отображены в виде экземпляров класса AbstractSuffix на графе. Аспект Nominalization является частью лингвистического понятия Noun.

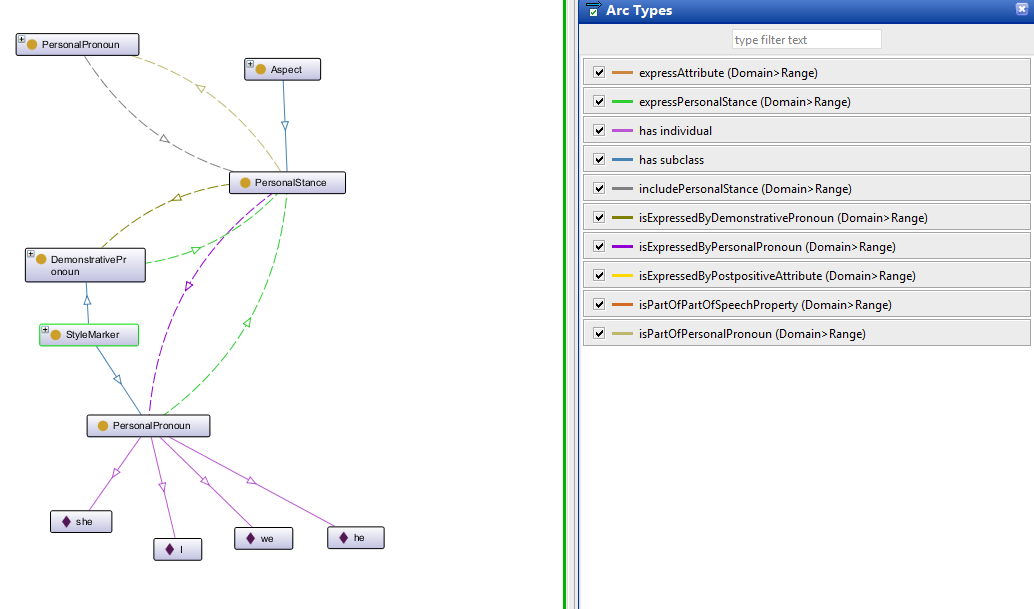
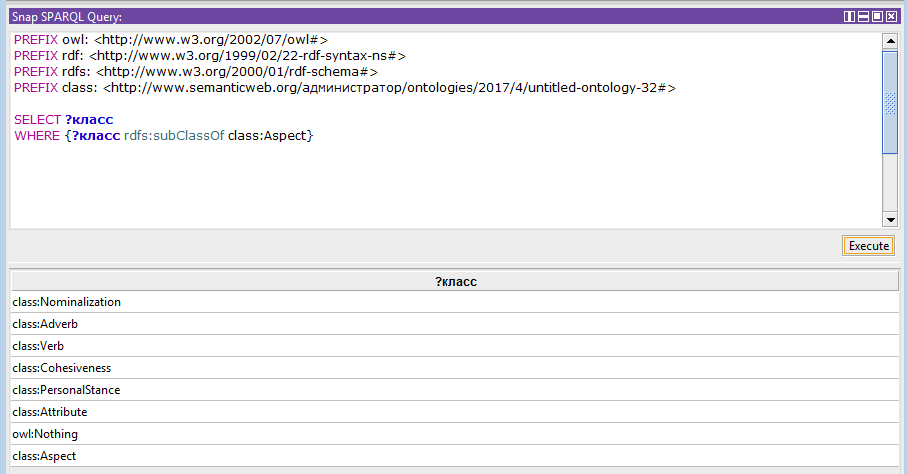


Рисунок 13. Онтологический граф с отношениями и экземплярами аспекта PersonalStance

Аспект PersonalStance выражается маркерами стиля PersonalNoun и DemonstrativePronoun. Аспект связан с понятием лингвистики PersonalPronoun. Маркеры стиля имеют свойство hasIndividual, что можно наблюдать на графе.

После того как онтология создана, заданы свойства, определены связи между классами и созданы экземпляры классов можно приступать разработке типовых запросов. Запрос по выводу всех классов онтологии и соответствующих им подклассов представлен на рис. 14. Другой запрос выводит только подклассы, принадлежащие конкретному классу. В данном случае представлены подклассы класса Aspect (рис. 15) и StyleMarker (рис. 16).

**Рисунок 14. Результаты выполнения запроса для выведения классов и подклассов**

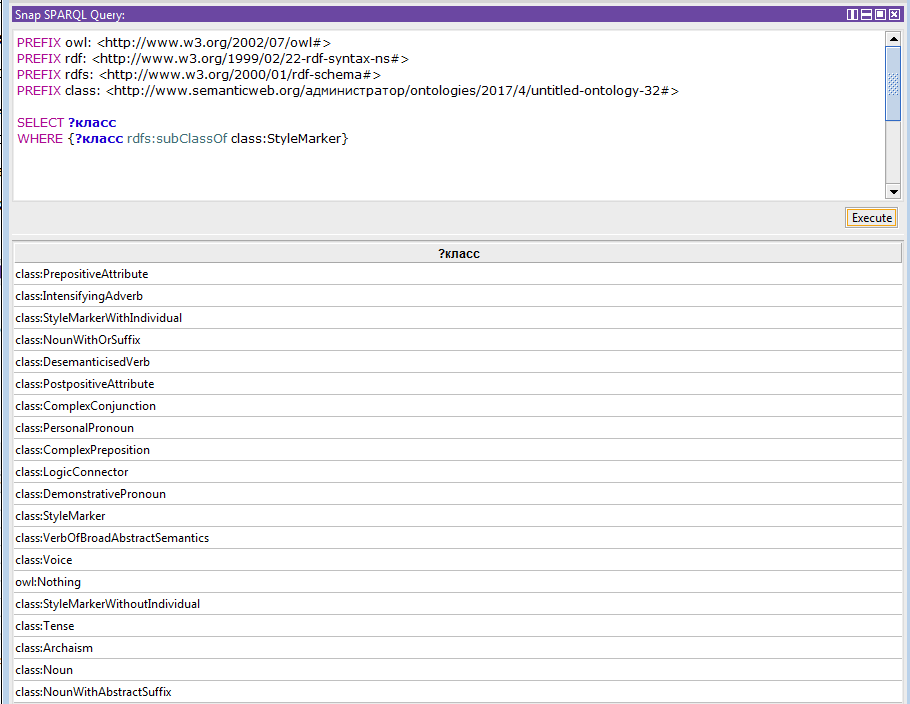


Рисунок 15. Результаты выполнения запроса для вывода аспектов

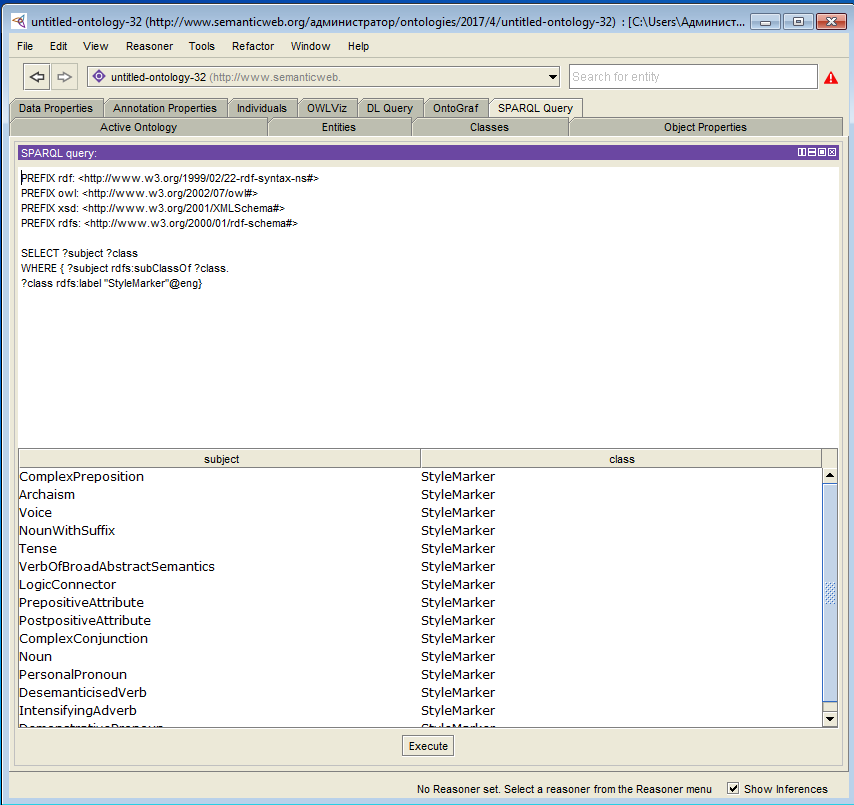


Рисунок 16. Результаты выполнения запроса для вывода маркеров стиля

Помимо выводов классов и соответствующих им подклассов было решено осуществить вывод всех имеющихся экземпляров и соответствующих им классов, а так же осуществить вывод экземпляров только одного класса, показанный на примере вывода всех экземпляров класса AbstractSuffix. Результаты запросов представлены на рис. 17 и рис. 18.

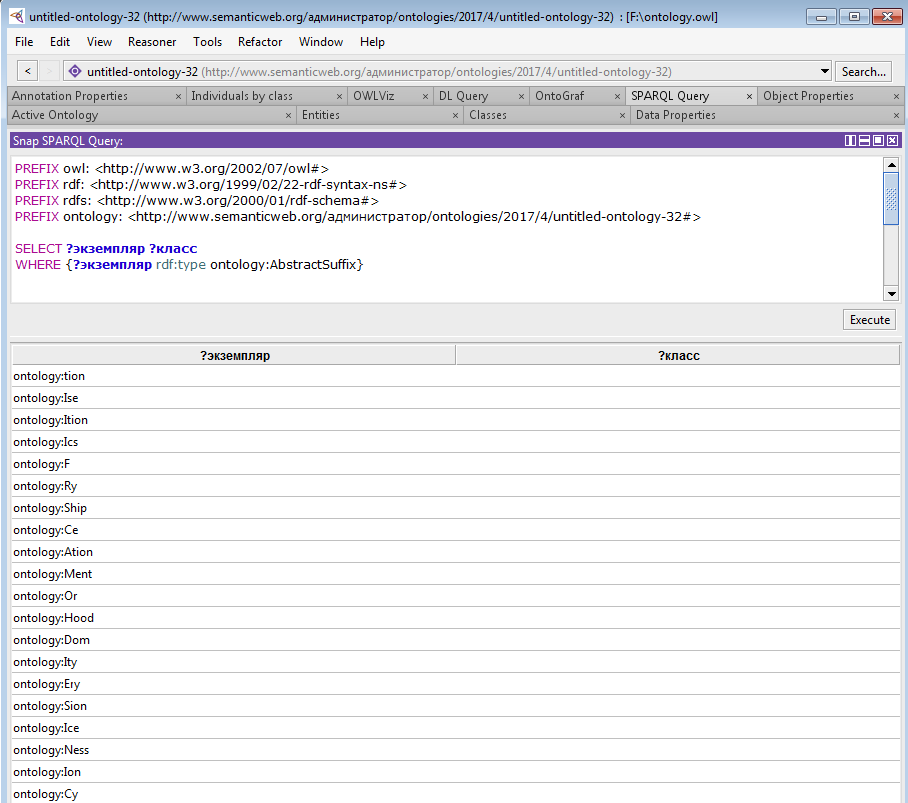


Рисунок 2.17. Результаты выполнения запроса для вывода всех экземпляров и соответствующих им классов

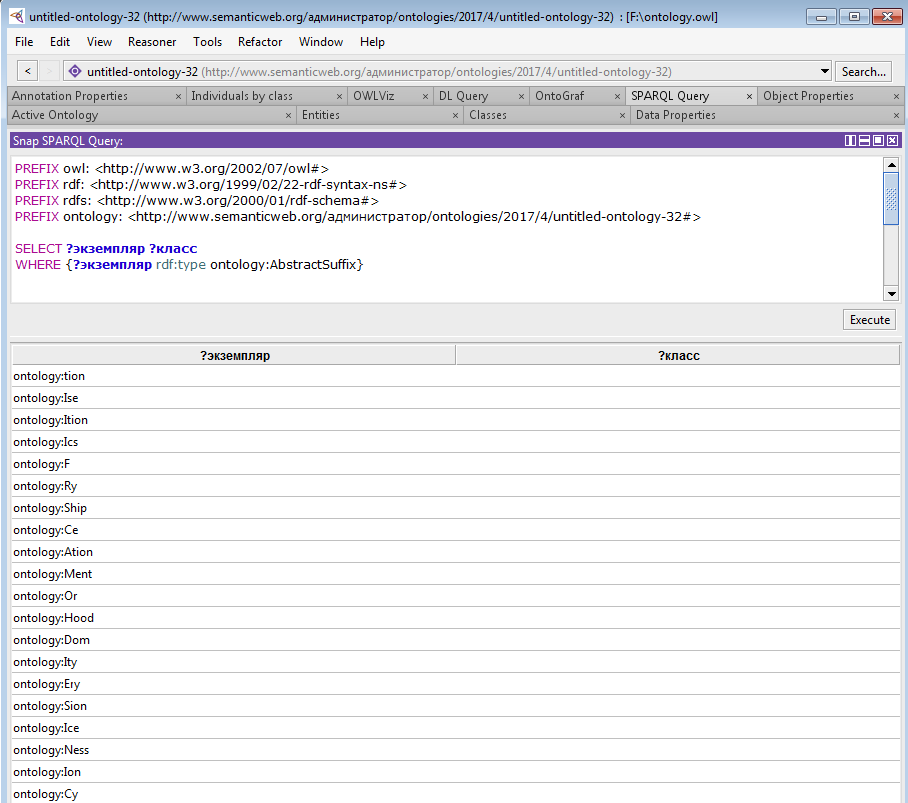


Рисунок 18. Результаты выполнения запроса для вывода всех экземпляров класса AbstractSuffix

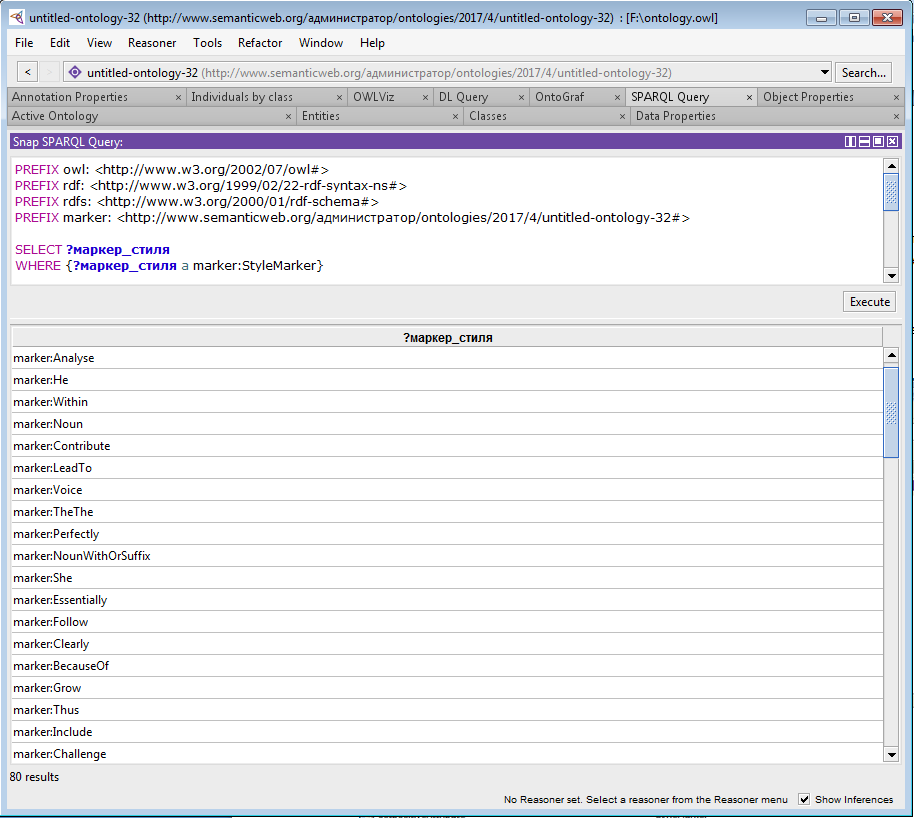
После того, как была создана онтология маркеров стиля, можно перейти к алгоритму генерации правил поиска маркеров. Для этого нужен будет плагин, который на основе созданной онтологии будет выполнять следующий порядок действий:

1. Для всех классов, у которых базовый класс StyleMarker, получаем экземпляры.
2. У экземпляров берем свойство template, в котором записано JAPE -выражение по поиску конкретного маркера.
3. Извлекаем JAPE -выражение или получаем имя файла для поиска маркера.
4. Соединяем выражения по поиску конкретных маркеров с помощью оператора или в общее JAPE -выражение, переходим по имени к файлу, содержащему реализацию поиска маркера.

JAPE – Java – обработчик для аннотационных шаблонов. JAPE предоставляет конечное множество преобразований над аннотациями, основанные на регулярных выражениях.

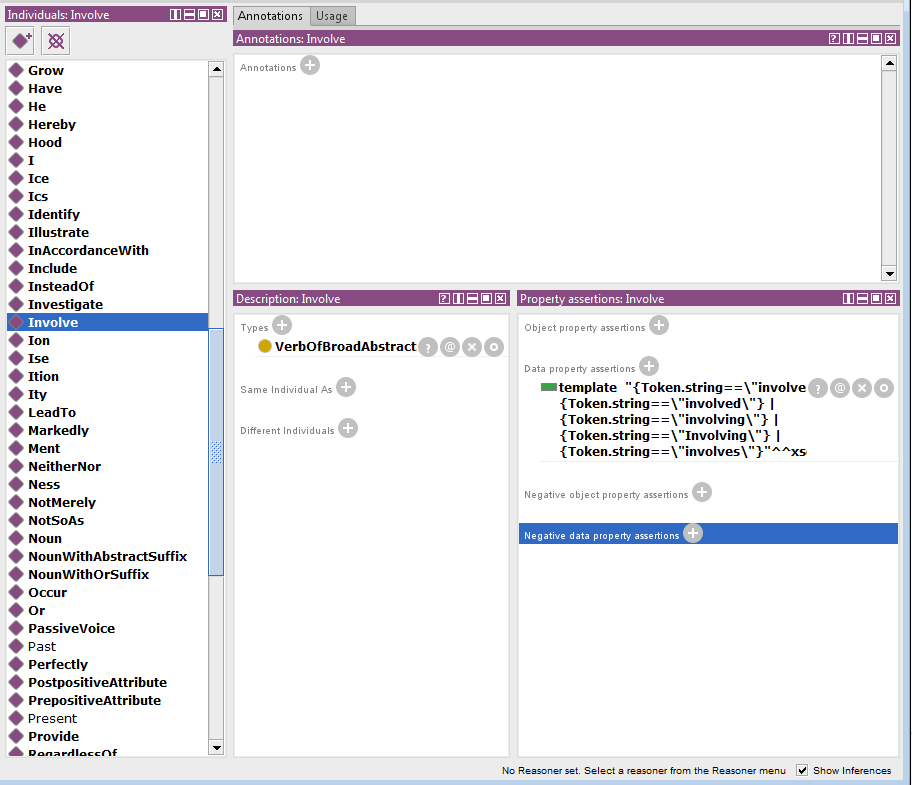
JAPE – правило состоят из двух частей: образец, называемый LHS (left hand side), и действие - RHS (right hand side). Часть LHS представляет собой регулярное выражение, в качестве атомов которого выступают аннотации, выделенные на предыдущих фазах. Если последовательность аннотаций во входном документе соответствует регулярному выражению из LHS, выполняется действие, описанное в RHS.

Для выполнения данных действий должны быть написаны SPARQL – запросы. Первый запрос по получению всех экземпляров класса StyleMarker представлен на рис. 19.

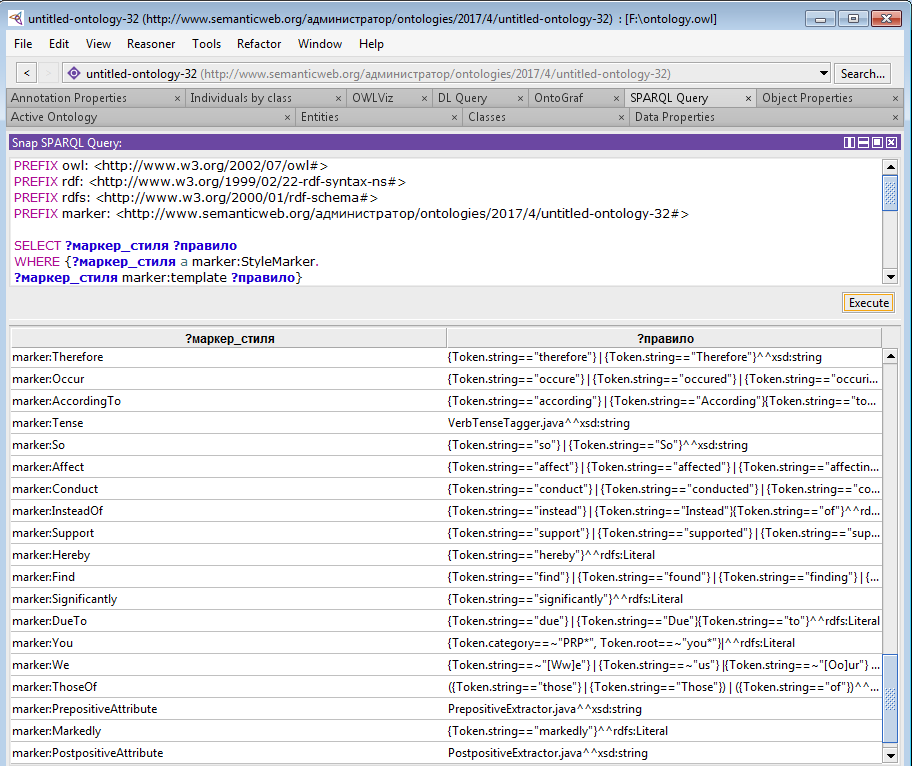


*Рисунок 19. Результаты выполнения SPARQL – запроса на получение всех экземпляров класса StyleMarker*

Второй пункт является наиболее основным, так как в нем говорится о свойстве Template, в котором содержится JAPE – выражение или имя файла, в котором реализован поиск маркера. В ходе выполнения данного пункта было решено разделить класс StyleMarker на два подкласса SyleMarkerWithIndividual (маркер стиля с экземпляром) и SyleMarkerWithoutIndividual (маркер стиля без экземпляра). В случае если экземпляры у выделенного маркера имеются, было решено добавить свойство Template, в котором содержалась бы левая сторона JAPE – выражения по поиску маркера. Если у класса нет экземпляров, то был создан единственный экземпляр с таким же свойством Template, содержащим в большинстве случаев имя Java – файла. На рис. 20 показано заполнение свойства Template правилом. Результат запроса на вывод всех экземпляров класса StyleMarker и их свойства Template представлен на рис. 21.

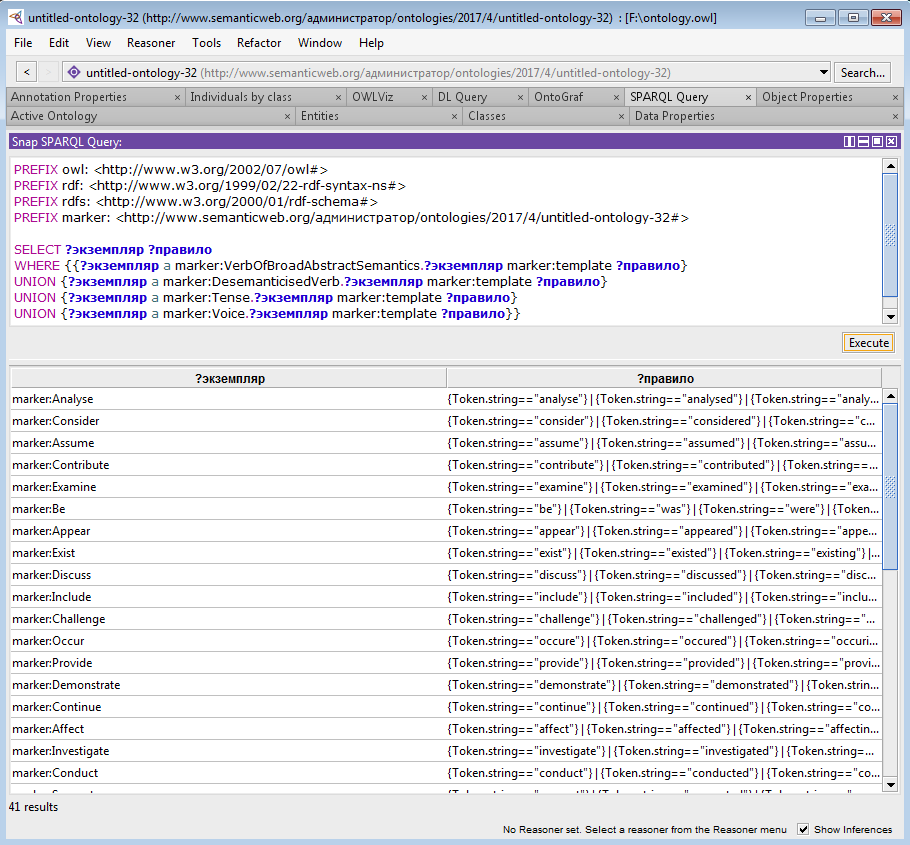


*Рисунок 20. Заполнение свойства Template JAPE – выражением*

**

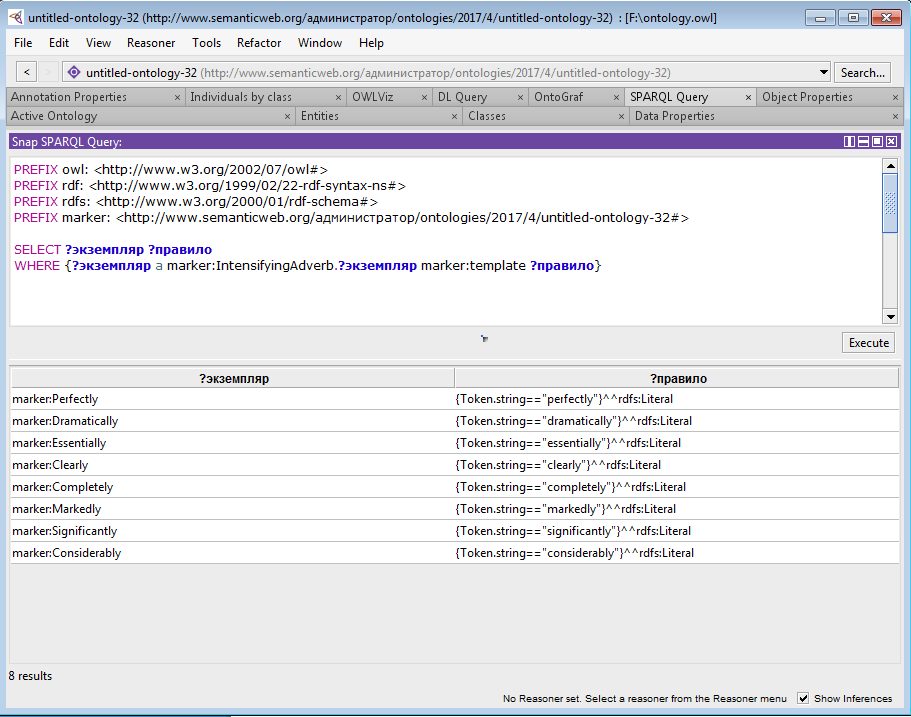
*Рисунок 21. Результаты выполнения SPARQL – запроса на вывод всех экземпляров класса StyleMarker и свойства Template*

Результаты запросов на поиск маркеров стиля, принадлежащих только к одному из выделенных аспектов, представлены на рис. 22, 23, 24, 25, 26, 27.



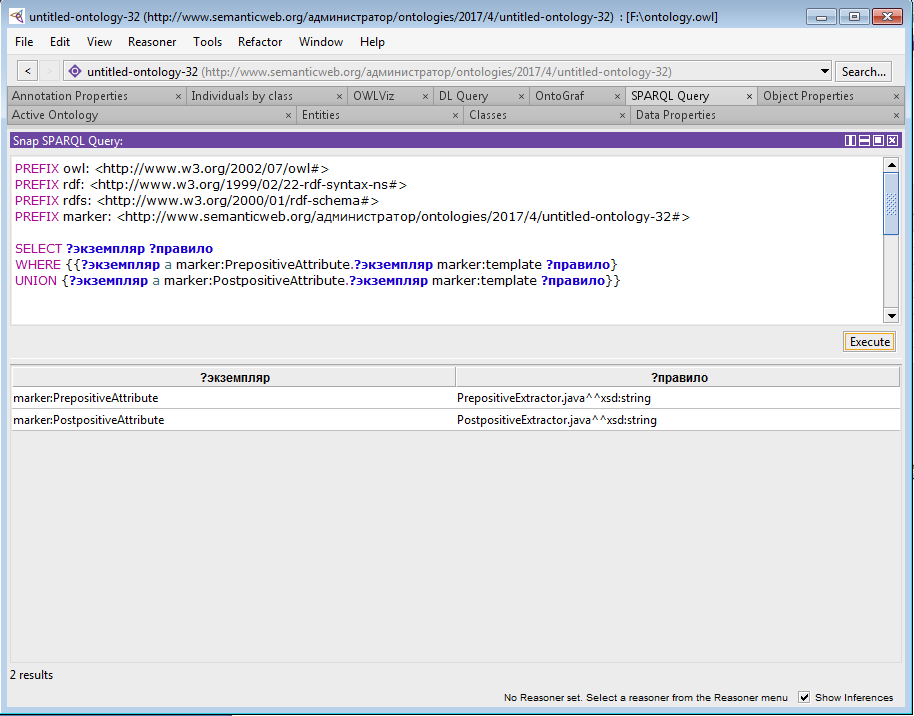
*Рисунок 22. Результаты выполнения SPARQL – запроса на вывод всех экземпляров классов, выражающих аспект Verb, и их свойств Template*

Аспект Verb выражается через четыре подкласса класса StyleMarker: VerbOfBroadAbstractSemantics, DesemanticisedVerb, Tense и Voice. Два из этих классов (VerbOfBroadAbstractSemantics, DesemanticisedVerb) имеют экземпляры, в которые были вписаны JAPE – выражения. Для классов Tense и Voice были созданы экземпляры и аналогично в свойстве Template указаны имена файлов для поиска данных маркеров.



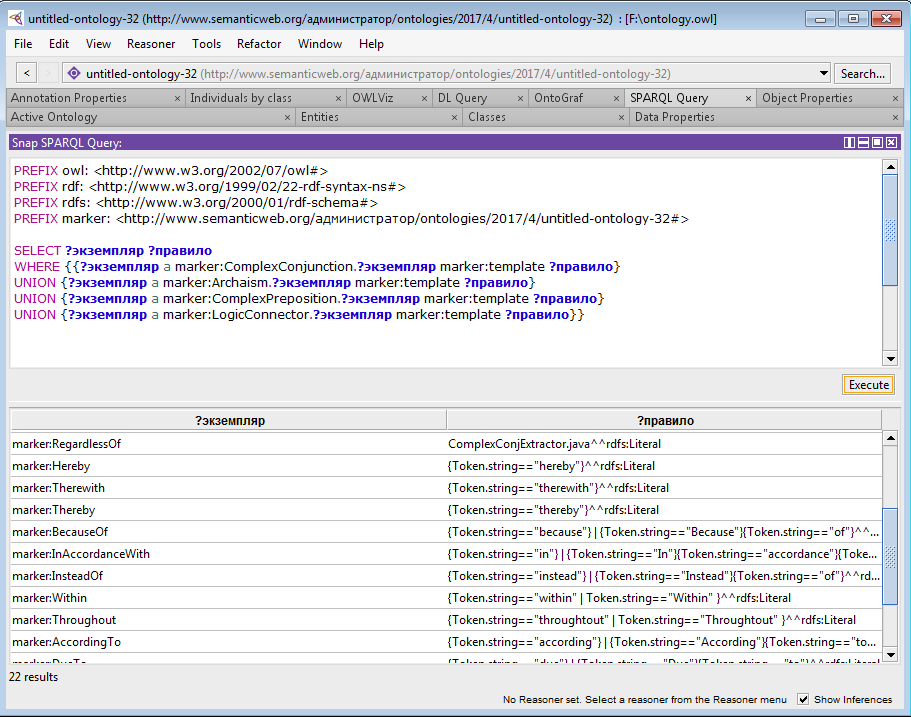
*Рисунок 23. Результаты выполнения SPARQL – запроса на вывод всех экземпляров классов, выражающих аспект Adverb, и их свойств Template*

Аспект Adverb выражается лишь одним классом маркеров стиля IntensifyingAdverb, каждый экземпляр которого имеет свойство с JAPE – выражением по поиску выделенных в качестве конкретных экземпляров наречий.



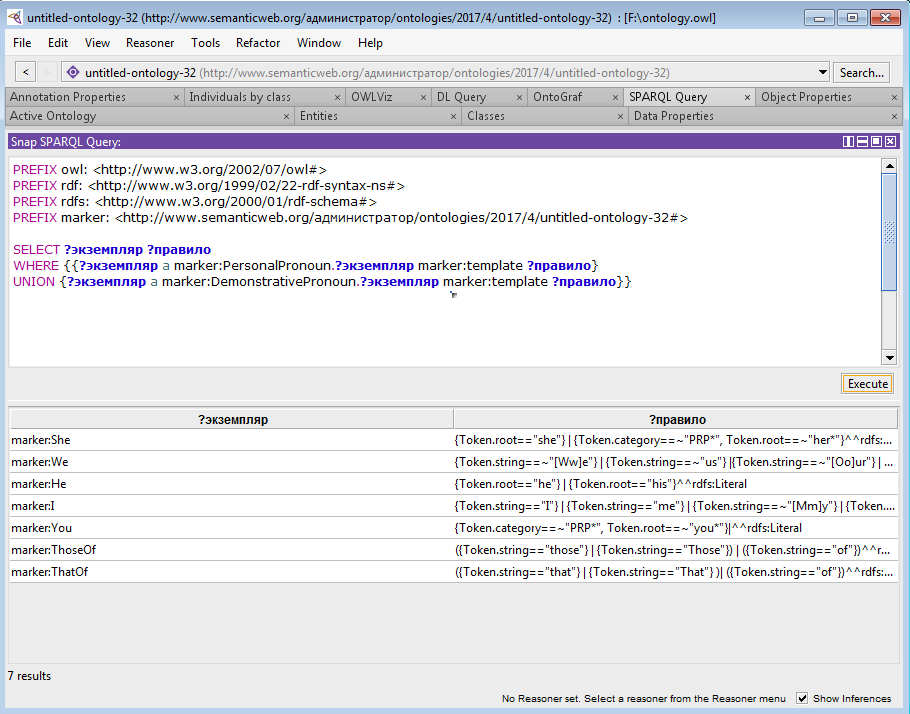
*Рисунок 24. Результаты выполнения SPARQL – запроса на вывод всех экземпляров классов, выражающих аспект Attribute, и их свойств Template*

Аспект Attribute выражается классами PostpositiveAttribute и PrepositiveAttribute, которые не имеют экземпляров и JAPE – правил для их поиска в тексте, но которые могут быть проаннотированы в тексте с помощью Java – файла, имя которого записано в свойство Template.

**

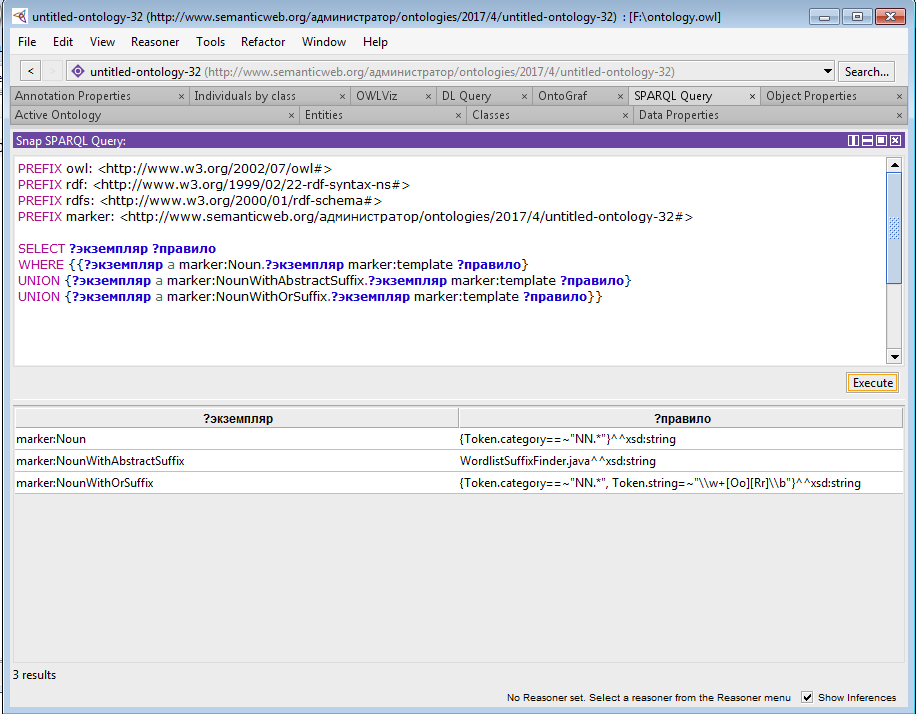
*Рисунок 25. Результаты выполнения SPARQL – запроса на вывод всех экземпляров классов, выражающих аспект Cohesiveness, и их свойств Template*

Аспект Cohesiveness выражается через четыре подкласса класса StyleMarker: ComplexConiunction, Archaism, ComplexPreposition и LogicConnector. Три из этих классов (Archaism, ComplexPreposition и LogicConnector) имеют экземпляры, в которые были вписаны JAPE – выражения. Для класса ComplexConiunction был создан экземпляр и аналогично в свойстве Template указано имя файла для поиска маркера стиля.



*Рисунок 26. Результаты выполнения SPARQL – запроса на вывод всех экземпляров классов, выражающих аспект PersonalStance, и их свойств Template*

Аспект PersonalStance выражается через два класса маркеров, экземпляр каждого из которых содержит Jape – выражение.



*Рисунок 27. Результаты выполнения SPARQL – запроса на вывод всех экземпляров классов, выражающих аспект Nominalization, и их свойств Template*

Аспект Nominalization выражается через три подкласса класса StyleMarker: Noun, NounWithAbstractSuffix и NounWithOrSuffix. Два класса Noun и NounWithOrSuffix не имеют экземпляров. Для них были созданы экземпляры и в свойстве Template указаны JAPE – выражения для поиска маркера стиля. Для экземпляров маркера NounWithAbstractSuffix было указано имя файла по их поиску.

Пункты 3 и 4 алгоритма выполняются непосредственно плагином, так как все данные, которые можно было предоставить с помощью онтологии, были получены с помощью типовых запросов. Таким образом, при появлении новых экземпляров маркеров стиля их достаточно будет добавить в нужный класс маркеров и прописать в свойство Template левую часть правила, JAPE – выражения, по его поиску в тексте, после чего плагин сможет его использовать для автоматического поиска. На данный момент поиск выделенных маркеров производится вручную, путем прописывания JAPE – правил для поиска каждого маркера, что занимает много времени. С помощью онтологии достаточно будет прописать левую часть правила поиска или имя файла, реализующего поиск маркера в свойство Template, после чего плагин сможет начать работу с ними, автоматизировав тем самым поиск маркеров стиля в тексте.

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 была проанализирована онтология описательной лингвистики GOLD, которая использовалась в качестве лингвистического уровня разрабатываемой онтологии. Помимо лингвистического уровня были выделены уровень апсекта и уровень маркера стиля, выделенные путем анализа данных, полученных от экспертов. Таким образом, 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.

Финальной стадией работы было составление алгоритма по генерации правил поиска маркеров стиля в тексте с помощью разработанной онтологии. С помощью онтологии можно будет генерировать правила автоматически, а не искать маркеры стиля в тексте, прописывая их каждый раз вручную с помощью JAPE – выражений и кода. Более того, при добавлении новых маркеров, достаточно будет только прописать правило поиска, а дальше плагин использует его для аннотирования текста согласно запросу. Так как проектирование онтологии является итеративным процессом, возможны добавления связей между экземплярами и классами, с целью получения лучшего эффекта при аннотировании текстов корпуса.

**Библиографический список**

1. FrameNet // лингвистический ресурс. URL: https://framenet.icsi.berkeley.edu/fndrupal/ (дата обращения: 17.05.2017).
2. General Architecture for text engineering// Система обработки естественного языка 1995. URL: <https://gate.ac.uk/download/> (дата обращения: 16.11.2017).
3. 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.
4. 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.
5. Protégé // Редактор онтологий. URL: <http://protege.stanford.edu/> (дата обращения: 16.11.2017).
6. 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.
7. [Strinyuk S. A.](https://www.hse.ru/en/org/persons/202208), Shuchalova Y., Lanin V. [Academic Papers Evaluation Software](https://publications.hse.ru/view/175342696" \t "_blank), in: Application of Information and Communication Technologies (AICT), 2015 9th International Conference on, 14-16 Oct. 2015. Rostov-on-Don: IEEE, 2015. [doi](http://doi.org/10.1109/ICAICT.2015.7338611) P. 506-510.
8. Абдулраб Х. Онтологический инжиниринг / Х. Абдулраб, Э.А. Бабкин // Вопросы образования. – 2007. -№2. – С.3-14.
9. Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. СПб: НИУ ИТМО, 2013. 72 с.
10. Загорулько Ю.А. Построение порталов научных знаний на основе онтологий // Вычислительные технологии. – 2007. – Т. 12.– Спецвып. 2. – С. 169–177.
11. Захаров В.П. Корпусная лингвистика: учебное пособие. СПб., 2005. 48 с.
12. Константинова Н.С., Митрофанова О.А. Онтологии как системы хранения знаний: учебное пособие. Спб: СпбГУ, 2007. 54 с.
13. Кузнецов О.П., Суховеров В.С., Шипилина Л.Б. Онтология как систематизация научных знаний: структура, семантика, задачи: учебное пособие.
14. Лукашевич Н.В. Тезаурусы в задачах информационного поиска – М., 2010. – С. 92-93.
15. Соснина Е.П. Введение в прикладную лингвистику: учебное пособие. Ульяновск: УГТУ, 2012. 110 с.