Годгильдиева Мария Михайловна

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
mariagodg@gmail.com

Захаров Виктор Павлович

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия

СОЗДАНИЕ СЛОВАРЯ ВАЛЕНТНОСТИ РУССКОГО ЯЗЫКА НА ОСНОВЕ КОМПЬЮТЕРНОГО СЛОВАРЯ ВАЛЕНТНОСТИ ЧЕШСКОГО ЯЗЫКА VERBALEX

Ключевые слова: валентность, словарь валентности, русский язык, семантические роли.

Целью работы является создание словаря валентностей русского языка с помощью методов проекта Verbalex. Модель описания, разработанная авторами Verbalex, позволяет максимально полно описать как морфосинтаксические, так и семантические характеристики аргументов, принимаемых глаголом. На данный момент в качестве эксперимента нами были вручную описаны 8 русских глаголов, чтобы проверить, сочетается ли модель описания Verbalex с аргументной структурой русского глагола. В результате модель была адаптирована для системы русского языка. В дальнейшем мы планируем создание словаря валентностей для глаголов русского языка полуавтоматическим образом с использованием методов составления, используемых в проекте Verbalex.

Godgildieva Maria

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

Zakharov Viktor

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

CREATION OF A RUSSIAN VALENCY DICTIONARY BASED ON DICTIONARY VERBALEX

Keywords: verbalex; valency, valency dictionary, Russian, semantic roles, Verbalex.

The goal of our project is to create a valency dictionary for Russian language using the methods of Verbalex project. The description model that was developed by the creators of Verbalex, allows to comprehensively describe both morphosyntactic and semantic properties of the verb arguments. At this moment as an experiment we have described 8 Russian verbs to see if Verbalex description model matches the argument structure of Russian verbs. As a result the model was adapted for

Russian language. In future we are planning to compile a valency dictionary for Russian verbs semiautomatically using the Verbalex methods.

1. Введение

Знание валентностей глагола необходимо для правильного понимания и составления текста на естественном языке как для человека, так и при автоматической обработке. Поскольку создание словаря валентностей вручную является долгим и трудоемким занятием, необходимо разработать способ автоматической обработки языкового материала и создания словаря. Также важно составить словарь так, чтобы им могли пользоваться и автоматические системы, и человек.

В нашей работе мы решили воспользоваться опытом чешского словаря валентности Verbalex. С учетом родственных связей чешского и русского языков была выдвинута гипотеза о том, что именно методы и модели описания, разработанные для чешского языка, легче всего адаптировать для русского языка и использовать в дальнейшей работе.

2. Структура создаваемого словаря и словаря Verbalex

2.1. Словарь Verbalex

Проект Verbalex [Verbalex] был начат в 2006 году в университете им. Т. Г. Масарика г. Брно и продолжает развиваться до сих пор.

Важным свойством Verbalex является его тесная связь с семантической сетью WordNet [Hlaváčková, Horák, 2006]. Именно по примеру WordNet основной единицей словаря является синсет — синонимический ряд.

Кроме того, приводится общее определение и семантический класс. Для указания класса была использована классификация М. Палмер (Palmer) [Nevěřilová, 2010].

Второй частью словарной статьи является описание рамок валентности [Horák, Pala, Hlaváčková, 2013], характерных для всего синонимического ряда. Для каждого актанта указывается падеж, в котором он может употребляться в данной конструкции. Для бо́льшей точности (и указания на одушевленность) приводится вопрос, который можно задать к актанту. Если один из актантов факультативен, ставится помета opt.

Семантические роли актантов разделяются на два уровня. На первом уровне содержатся основные семантические роли, их опи-

сание основывается на сущностях первого (1stOrderEntity) и второго порядка (2ndOrderEntity) по EuroWordNet Top Ontology и Base Concepts. На первом уровне для обозначения ролей отбирались понятия с номером значения 1 или 2, т. е. самые общие. Всего используется 32 семантические роли первого уровня. На втором уровне приводится более детальное описание семантических ролей, большее внимание уделяется семантическим ограничениям. Используются прямые гипонимы из WordNet [Hlaváčková, 2007], которые служат для того, чтобы показать наиболее ожидаемое значение актанта. Роли второго порядка формируют открытый список, который можно расширить по необходимости. На 2013 г. список содержал 811 семантических ролей.

Подобный подход позволяет сузить разнообразие лексико-семантических групп, элементы которых могут занять данную позицию в рамке валентности. Так, к примеру, в бо́льшей части случаев актанту в позиции подлежащего приписывается роль AG (agens, агенс), которая, на самом деле, является очень общей и обозначает просто того, кто выполняет данное действие. Однако с помощью семантических ролей второго уровня можно уточнить возможное значение данного актанта: человек, животное, организация и т. д. Иногда это сужение является существенным. К примеру, подлежащим глагола *родить* в прямом значении может быть только женщина, поэтому роль первого порядка AG логично сузить до роли второго порядка woman:1.

2.2. Описание эксперимента

Наша цель — исследовать систему описания глаголов, используемую в Verbalex, и проверить её на русском языке. На первом этапе работы было решено перевести словарные статьи Verbalex на русский язык и посмотреть, как эти рамки валентностей соотносятся с соответствующими глаголами русского языка. Для этого было выбрано 8 глаголов: *říct*, *rodit*, *žít*, *zářit*, *vlastnit*, *šetřit*, *jet*, *chránit*, в переводе — *сказать*, *родить*, *жить*, *сиять*, *владеть*, *беречь*, *ехать*, *защищать*. Для двух глаголов количество рамок сократилось сразу же при переводе. Это связано с идиоматическими употреблениями глаголов, которые не повторялись в русском языке.

На следующем этапе из корпуса Araneum Russicum Minus для каждого русского глагола было выбрано по 100 контекстов (предложений). Оказалось, что полученные рамки не могут полностью отразить употребления русского языка. Возникли две проблемы: в контекстах из корпуса не было примера, который можно было бы

отнести к одной из рамок, и, наоборот, находились предложения, которые нельзя было описать одной из имеющихся рамок. Первую проблему можно было бы списать на небольшой размер выборки, но вторую — нет. Поэтому было решено пойти по другому пути и попытаться самостоятельно описать валентности тех же глаголов по уже имеющимся контекстам.

Разметка предложений производилась вручную в несколько этапов. Сначала предложения разбивались на группы по синтаксическим характеристикам (отсутствие/наличие прямого дополнения, косвенного дополнения и т. п.). Далее эти группы размечались с помощью семантических ролей первого порядка. Использовался список ролей, из [Hlaváčková, 2007], и их описание, приведенное в данной работе. Затем группы первого порядка размечались по семантическим ролям второго порядка. Мы не переводили на русских язык семантические роли ни первого, ни второго порядка, поскольку в Verbalex также используются англоязычные обозначения. Морфосинтаксические сведения были переведены с указанием соответствующего вопросительного слова и падежа в русской падежной системе.

Пример полученной рамки валентности для глагола *сиять*:

AG (object:1, что1) /VERB/ ATTR (color:1, attribute:2, чем6)
Пример: *Фрески сияли яркими красками.*

3. Заключение

В ходе проведенной работы были исследованы методы и принципы описания словаря Verbalex. Была подтверждена гипотеза о возможности использования чешской системы описания глаголов для русского языка. Более того, данная модель была адаптирована для системы русского языка. Таким образом, составление словаря валентностей русского языка на основе Verbalex представляется возможным.

В дальнейшем мы видим несколько вариантов развить данное исследование: самым логичным путем будет собственно создание словаря валентностей глаголов русского языка, аналогичного Verbalex. Главной целью является автоматизация данного процесса.

Кроме того, также можно исследовать валентность других частей речи и адаптировать модель описания Verbalex, к примеру, для существительных или прилагательных. Следующим этапом будет создание словаря для выбранной части речи.

ЛИТЕРАТУРА

Hlaváčková D. Databáze slovesných valenčních rámců VerbaLex. 2007.

Hlaváčková D. The Relations between Semantic Roles and Semantic Classes in VerbaLex. In Recent Advances in Slavonic Natural Language Processing RASLAN 2007. Brno, 2007.

Hlaváčková D., Horák A. VerbaLex — New Comprehensive Lexicon of Verb Valencies for Czech. In Computer Treatment of Slavic and East European Languages. Bratislava, 2006. P. 107–115.

Horák A., Pala K., Hlaváčková D. Preparing VerbaLex Printed Edition. In Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013. Brno, 2013. P. 3–11.

Nevěřilová Z. Semantic Role Patterns and Verb Classes in Verb Valency Lexicon. In Proceedings of the 13th International Conference on Text, Speech and Dialog TSD 2010. Heidelberg, 2010.

Verbalex. URL: https://nlp.fi.muni.cz/cs/VerbaLex (дата обращения: 14.05.2016).