Создание словаря валентностей русского языка на основе компьютерного словаря валентностей чешского языка Verbalex

М.М. Годгильдиева, В.П. Захаров

Санкт-Петербургский государственный университет
mariagodg@gmail.com v.zakharov@spbu.ru

Аннотация

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.

Ключевые слова: валентность, словарь валентностей, русский язык, Verbalex.

1. Введение

Знание валентностей глагола необходимо для правильного понимания и составления текстов на естественном языке, как для человека, так и при автоматической обработке. Для русского языка создано множество словарей, отражающих разные аспекты лексики, и во многих рамки валентностей глагола определённым образом представлены. Тем не менее, для русского языка пока не было создано словаря, главной задачей которого было бы максимально полно отразить синтаксическую и семантическую валентность лексики. Более того, поскольку создание словаря вручную является достаточно долгим и трудоемким занятием, необходимо разработать способ автоматической обработки языкового материала для создания словаря. Также важно составить систему описания синтаксических и семантических связей глагола таким образом, чтобы ею могли пользоваться и автоматизированные системы, и человек.

Для ряда других языков подобная работа была проведена (и проводится до сих пор). Широко известны различные словари английского языка, к примеру, A Valency Dictionary of English, Framenet, Pattern Dictionary of English verbs. Словарей подобного масштаба для русского языка пока не существует. Частично информация о валентности глаголов содержится в WordNet-подобных словарях, но и полного WordNet-словаря для русского языка также не существует.

В нашей работе мы решили воспользоваться опытом создания чешского словаря валентностей Verbalex. История развития компьютерных словарей валентностей в чешской лингвистике насчитывает более 20 лет. Учитывая родственные связи чешского и русского языков, нами 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. Теоретические вопросы

Понятие «валентность» появилось в работах по исследованию сочетаемости языковых единиц [1]. Основоположником теории валентности является Л. Теньер [2]. Он первым предложил разделять члены предложения, зависимые от глагола, на актанты и сирконстанты. Актантами он называл субъекты и объекты, принимающие участие в действии, обозначаемом глаголом, сирконстантами — остальные зависимые единицы, выражающие, в основном, различные обстоятельства. На самом деле, деление на актанты и сирконстанты намного сложнее и туманнее. К примеру, для глагола *проживать* место действия будет актантом, а не сирконстантом. Грубо говоря, актантом глагола является член предложения, соответствующий важному, неотъемлемому участнику ситуации [3].

Более того, Теньер разделил глаголы на группы в зависимости от того, сколько актантов они могут присоединять (авалентные, одноактантные, двухактантные, трехактантные). По Теньеру не существует глаголов, способных присоединять более 3 актантов. Однако Ю.Д. Апресян [4] приводит подобные примеры: *арендовать* (5 актантов) — кто? что? у кого? на сколько? за сколько? и *командировать* (6 актантов) — кто? кого? куда? откуда? на сколько? с какой целью? Очевидно, что подобные глаголы редко употребляются в предложении со всеми актантами. Тогда можно говорить об обязательной и факультативной валентности. Для русского языка характерна факультативность валентности с восполнением недостающих актантов из лингвистического и экстралингвистического контекста.

В «Лингвистическом энциклопедическом словаре» валентность определяется как «способность слова вступать в синтаксические связи с другими элементами» [5]. Таким образом, понятие валентности распространяется и на другие части речи, помимо глагола. Тем не менее, в данной работе мы планируем создание словаря валентностей только для глаголов.

Следует также разделять семантическую и синтаксическую валентность глагола. «Семантической валентностью лексемы называется любая переменная X, входящая в толкование данной лексемы», а «синтаксической валентностью лексемы L называется селективный признак, который указывает, что данная лексема L может иметь в качестве вершины или в качестве зависимого слова слово W» [3]. Другими словами, семантическая валентность обуславливается значением глагола, синтаксическая — его грамматическими характеристиками. Обычно семантическая и синтаксическая валентности согласованы. Если в ситуации, обозначаемой лексемой, есть определенный участник, то, скорее всего, у лексемы будет и синтаксический актант, соответствующий этому участнику. Однако семантические валентности лексемы, соответствующие синтаксическим актантам, могут не отражаться в предложении. Такая валентность называется нереализуемой. Ю.Д. Апресян приводит в качестве примера глагол *промахнуться*, обладающий 4 семантическими валентностями (кто? в кого? чем? посредством чего?), но реализующий обычно только актант подлежащее [4].

Для теории валентности также важно понятие семантической роли. Если мы представляем себе семантику глагола как описание некой ситуации, то его актанты воплощают собой участников ситуации. В схожих ситуациях участники будут играть схожие роли, и эти закономерности можно описывать в терминах семантических ролей. Одним из первых идею о семантических ролях или, как он их называл, «глубинных падежах» выдвинул Чарльз Филлмор [6] [7]. По его теории, падеж является глубинным синтактико-семантическим отношением, которое выражается в языке каким-либо способом (аффиксация, частицы, порядок слов и т. д.). Вместе с формальной структурой предложения семантические роли влияют на модель управления глагола. Изначально Филлмор предложил список из 6 глубинных падежей: агентив — инициатор действия, инструменталис — средство, датив — тот, кого затрагивает действие, фактитив — результат действия, локатив — место действия, объектив — нейтральный падеж, значение определяется глаголом. Лингвисты, пытавшиеся создать перечень универсальных семантических ролей, столкнулись с двумя проблемами: семантические роли не должны были быть слишком общими или слишком специальными. В зависимости от стратегии составителя список может включать от 6 до более чем 100 семантических ролей.

2.2. Словарь Verbalex

Проект Verbalex [8] был начат в 2006 году в университете им. Т.Г Масарика (Чехия, Брно) и продолжает развиваться до сих пор. Целью создателей было составить словарь, которой бы был полезен и понятен для лингвистов, исследователей чешского языка, и который также можно было бы использовать в задачах автоматизированной обработки текста (информационный поиск, разрешение неоднозначности, извлечение информации из текста и т. п.).

На данный момент Verbalex описывает более 10 500 глагольных лексем [9]. Лексемы выбираются по источникам Slovník spisovné češtiny (Словарь литературного чешского языка) и Slovník spisovného jazyka českého (Академический Словарь чешского литературного языка). При этом учитывается частотность лексем, которая проверяется по корпусным данным. В

VerbaLex вошли только лексемы литературного языка, не учитывались диалектные, разговорные, книжные, устаревшие или редко употребляемые слова. Тем не менее, приоритет при выборе, включать ли лексему, отдавался частоте в корпусе. Лексемой всегда считается одно слово, исключение составляют глаголы reflexive tantum, где возвратная частица se, si пишется отдельно. Отрицание не учитывается, допускается, что оно не влияет на рамку валентности глагола. Большая часть аннотирования была проведена вручную с использованием специальных программ.

Важным свойством Verbalex является его тесная связь с семантической сетью WordNet. В частности, именно по аналогии с WordNet основной единицей словаря является синсет – синонимический ряд, а не отдельная лемма.

В словарной статье словаря Verbalex описывается синонимические ряды глаголов и их составная рамка валентности (complex valency frame). Заголовком статьи, как уже упоминалось, является не отдельный глагол, а синсет (синонимический ряд). Более того, его элементами являются не леммы как таковые, а их отдельные значения. Синонимия понимается не в традиционном, а в более широком смысле, элементы синонима близки по значению, но не всегда взаимозаменяемы в контексте. Для синсета уточняется значение (по WordNet), вид (совершенный, несовершенный или оба) и то, могут ли глаголы синсета образовывать пассивный залог. Если глагол входит в данный синсет в обоих видах, то номер значения приписывается сразу обеим формам.

Кроме того, приводится общее определение и семантический класс. В качестве основы была использована классификация Бет Левин (Beth Levin), созданная для английских глаголов. Затем в проекте Verbnet Марты Палмер (Martha Palmer) количество базовых семантических классов увеличилось с 48 до 82 классов. В рамках проекта Verbalex эта классификация была переведена и адаптирована для чешского языка. В оригинале классификация основывалась на аргументной структуре глаголов английского языка, после адаптации стали использоваться и семантические критерии, и классы стали лучше сочетаться с чешской языковой системой [10; 11]. Главным семантическим критерием, который был предложен авторами проекта, является идея, что глаголы, получающие в качестве аргумента одну и ту же семантическую роль второго порядка (см. ниже), относятся к одному и тому же семантическому классу. Естественно, подобный критерий применяется только к семантическим ролям с достаточно большой частотой в VerbaLex (от 30 до 1000 раз). У семантических ролей с большей частотой обычно слишком общее значение, чтобы их можно было использовать таким образом.

Второй частью словарной статьи является описание простых (базовых) рамок валентностей (basic valency frames), характерных для всего синонимического ряда (см. рис. 1). Тем не менее, каждая рамка валентностей относится к субсинсету (т.е. подгруппе основного синонимического ряда). При описании учитываются как морфосинтаксические, так и семантические характеристики актантов. Для каждого актанта указывается падеж(и), в котором он может употребляться в данной конструкции. Для большей точности приводится вопрос, который можно задать к актанту. Более того, таким способом учитывается одушевленность/неодушевленность существительных

(вопросы кто? что?). В случае если один из актантов факультативен, ставится помета opt.

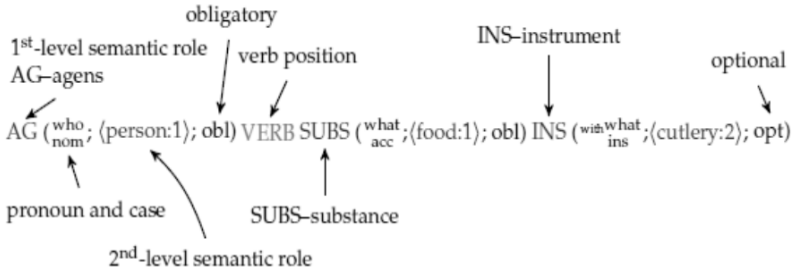

Рис. 1. Пример рамки валентностей в Verbalex (на английском) [9]

Место глагола в конструкции отмечается знаком VERB. Рамка представлена в так называемом стандартном порядке: актанты слева от глагола (обычно подлежащее) — глагол — актанты справа от глагола. Это не так существенно для чешского и русского языка, в которых достаточно свободный порядок слов, но создатели словаря предполагают, что эта спецификация может пригодиться не только для анализа предложения, но для генерирования. Особенным образом описываются безличные конструкции (с нулевой валентностью), используется специальная семантическая роль ISUB (inexplicit subject) в позиции подлежащего.

Семантические роли актантов разделяются на два уровня (см. рис. 2). На первом уровне содержатся основные семантические роли, их описание основывается на сущностях первого (1stOrderEntity) и второго порядка (2ndOrderEntity) по EuroWordNet Top Ontology и Base Concepts. Оба списка описывают ядро межъязыкового лексикона, понятия, встречающиеся во многих языках. 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 [12], которые показывают наиболее ожидаемое значение актанта, заполняющего данную валентность. Роли второго порядка формируют открытый список, который можно расширить по необходимости. По состоянию на 2013 г. список содержал 811 семантических ролей. Более абстрактные значения, почти повторяющие значения ролей первого порядка (к примеру, beneficiary:1, patient:2), заменяются соответствующими, более конкретными значениями (обычно person:1, или другие варианты в зависимости от конструкции).

AG	**agens** – *ten, kdo něco aktivně vykonává* entity:1, person:1, woman:1, man:1, animal:1, organization:1, institution:1, group:1, social group:1, caregiver:1, entertainer:1, policeman:1, artist:1, adventurer:1, player:1, ...
ART	**výrobek** – artifact:1, creation:2, garment:1, textile:1, goods:1, publication:1, line:18, excavation:3, knot:2, cord:1, building material:1, brick:1, covering:2, cigaret:1, roofing material:1, footwear:1, millstone:3, pipe:1, cigarette holder:1, emblem:1, way:6, glassware:1, ceramic ware:1, puppet:1, ...

Рис. 2. Пример семантических ролей первого и второго порядка [13]

Подобный подход позволяет сузить разнообразие лексико-семантических групп, элементы которых могут занять данную позицию в рамке валентности. Так, к примеру, в большей 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. В некоторых случаях это сужение может и не иметь большого смысла, но иногда является очень значимым. Для примера, подлежащим при глаголе *родить* в прямом значении может быть только женщина, поэтому роль первого порядка AG логично сузить до роли второго порядка woman:1.

Для каждой рамки валентности приводятся несколько примеров употребления, возможные синонимы, возвратность глагола и значение, в котором в ней употребляется глагол в данном контексте. Для последней части используются три пометы: prim — в прямом значении, posun — в переносном значении, idiom — в идиоматическом или фразеологическом.

2.3. Описание эксперимента

Цель практической части нашей работы состояла в том, чтобы исследовать систему описания глаголов, используемую в словаре Verbalex, и проверить, применима ли она к глаголам русского языка. На первом этапе работы мы попробовали перевести словарные статьи Verbalex на русский язык и посмотреть, как эти рамки валентностей соотносятся с соответствующими глаголами русского языка (переводными эквивалентами глаголов в заголовке статьи). Для этого было выбрано 8 глаголов: *říct, rodit, žít, zářit, vlastnit, šetřit, jet, chránit*, в переводе — *сказать, родить, жить, сиять, владеть, беречь, ехать, защищать*. Как можно заметить, для чистоты эксперимента мы подбирали глаголы из разных лексико-семантических групп, с разным количеством актантов, переходные и непереходные. Количество рамок валентности сильно отличалось в зависимости от глагола (см. табл. 1). Для двух глаголов количество рамок сократилось сразу же при переводе. Это связано с идиоматическими употреблениями глаголов, которые очевидным образом не повторялись в русском языке.

Таблица 1. Соотношение количества рамок валентности на первом этапе

Чешский глагол	Количество рамок валентности	Русский эквивалент	Количество рамок валентности при переводе
Říct	7	Сказать	7
Rodit	3	Родить	3
Žít	6	Жить	5
Zářit	3	Сиять	3
Vlastnit	2	Владеть	2
Šetřit	3	Беречь	3
Jet	7	Ехать	7
Chránit	21	Защищать	19

На следующем этапе полученные рамки валентности были проверены по корпусу. Для этого из корпуса Araneum Russicum Minus (http://ucts.uniba.sk/) для каждого русского глагола было выбрано по 100 контекстов (предложений). В результате оказалось, что полученные путем перевода рамки не могут полностью отразить употребления русского языка. Возникали две проблемы: в контекстах из корпуса не находилось примера, который можно было бы отнести к одной из рамок, и, наоборот, встречались предложения, которые нельзя было описать ни одной из имеющихся рамок. Если первую проблему можно было бы списать на небольшой размер выборки, то со второй проблемой нельзя было поступить таким же образом. Поэтому было решено пойти другим путем и попытаться самостоятельно описать валентности тех же глаголов по уже имеющимся контекстам.

Таблица 2. Количество рамок валентности на втором этапе

Глаголы	Количество рамок валентности
Сказать	5
Родить	5
Жить	4
Сиять	4
Владеть	8
Беречь	7
Ехать	5
Защищать	9

Разметка предложений производилась вручную в несколько этапов. Сначала предложения разбивались на группы по чисто синтаксическим характеристикам (отсутствие/наличие прямого дополнения, косвенного дополнения и т. п.). Далее эти группы размечались с помощью семантических ролей первого порядка. Использовался список ролей, приведенный в приложении В к [13], и их описание, приведенное в данной работе. Затем группы первого порядка размечались по семантическим ролям второго порядка. Как и в Verbalex, для этого использовались базовые понятия из WordNet с указанием соответствующего номера. Мы не переводили на русский язык семантические

роли ни первого, ни второго порядка, поскольку в Verbalex также используются англоязычные обозначения. Тем не менее, морфосинтаксические сведения (данные о форме, в которой может употребляться аргумент) были переведены с указанием соответствующего вопросительного слова и падежа в русской падежной системе.

Пример полученных рамок валентности можно увидеть на следующем рисунке.

2. РОДИТЬ (НСВ)

Определение: дать жизнь ребенку, произвести на свет

Class: engender-27

Passive: yes

1) AG (woman:1, кто1) /VERB/ REC (man:1, кому3, opt) STATE (marriage:1, state:4, в чем6, opt)

Пример: она должна родить в замужестве.

Употребление: прямое

2) AG (animal:1, кто1) /VERB/ ENT(animal:1, что4, opt)

Пример: собака родила одного щенка

Употребление: прямое

Рис.3. Пример рамок валентности для глагола *родить*

3. Заключение

В ходе проведенной работы были исследованы методы и принципы описания словаря Verbalex. Они же использовались для создания собственного словаря. В результате подтвердилась гипотеза, что использование системы описания глаголов чешского языка в целом возможно и для русского языка. Однако 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 развития данного исследования:

Самым логичным путем будет собственно создание словаря валентностей глаголов русского языка, аналогичного Verbalex. Естественно, мы заинтересованы в автоматизации данного процесса. Для этого необходимо освоить и адаптировать инструменты, используемые в оригинальном проекте, или при необходимости создать собственные.

Кроме того, также можно исследовать валентность других частей речи и адаптировать модель описания Verbalex, к примеру, для существительных или

прилагательных. Аналогичным образом, следующим этапом будет создание словаря для выбранной части речи.

Благодарности

Исследование поддержано грантом РГНФ № 16-04-12019 «Интеграция тезаурусов RussNet и YARN» и частично грантом РГНФ № 15-04-12029 «Программная разработка электронного ресурса с онлайн-версией русскоязычной вопросно-ответной системы».

Литература

[1] Гак В.Г. Сочетаемость // Лингвистический энциклопедический словарь. URL: http://tapemark.narod.ru/les/483a.html (дата обращения: 14.05.2016).
[2] Теньер Л. Основы структурного синтаксиса. / Пер. с франц. Вступ. ст. и общ. ред. В. Г. Гака. М., 1988.
[3] Тестелец Я.Г. Введение в общий синтаксис. М., 2001.
[4] Апресян Ю.Д. Избранные Труды. Том 1. Лексическая семантика. Синонимические средства языка. М., 1995.
[5] Гак В.Г. Валентность // Лингвистический энциклопедический словарь. URL: http://tapemark.narod.ru/les/079c.html (дата обращения: 14.05.2016).
[6] Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. Вып. 10. М., 1981.
[7] Филлмор Ч. Дело о падеже открывается вновь// Новое в зарубежной лингвистике. Вып. 10. М., 1981.
[8] Verbalex. URL: https://nlp.fi.muni.cz/cs/VerbaLex (дата доступа: 14.05.2016).
[9] Horák A., Pala K., Hlaváčková D. Preparing VerbaLex Printed Edition // Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013. Brno, 2013. P. 3 – 11.
[10] Hlaváčková D. The Relations between Semantic Roles and Semantic Classes in VerbaLex // Recent Advances in Slavonic Natural Language Processing RASLAN 2007. Brno, 2007.
[11] Nevěřilová Z. Semantic Role Patterns and Verb Classes in Verb Valency Lexicon // Proceedings of the 13th International Conference on Text, Speech and Dialog TSD 2010. Heidelberg, 2010. P. 150–156.
[12] Hlaváčková D., Horák A. VerbaLex — New Comprehensive Lexicon of Verb Valencies for Czech // Computer Treatment of Slavic and East European Languages. Bratislava, 2006. P. 107–115.
[13] Hlaváčková D. Databáze slovesných valenčních rámců VerbaLex. 2007.

Creation of a Russian valency dictionary based on NLTK Czech computer valency dictionary Verbalex

M. Godgildieva, V. Zakharov
Saint Petersburg State University

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.

Keywords: valency, valency dictionary, Russian, Verbalex.