

Святошик Марина Игоревна
аспирант кафедры теории и практики
китайского языка
Минский государственный
лингвистический университет
г. Минск, Беларусь

Marina Svyatoshchik
PhD Student of the Department
of Theory and Practice
of the Chinese Language
Minsk State Linguistic University
Minsk, Belarus
svyatoshchikm@mail.ru

О КОНВЕРТАЦИИ ЧАСТЕЙ РЕЧИ В ЧАСТИ ЯЗЫКА

ABOUT CONVERTING PARTS OF SPEECH INTO PARTS OF LANGUAGE

Статья посвящена конвертации частей речи в части языка (категории китайско-японской лингвистической традиции). В статье демонстрируется семантическое преобразование русских лексических единиц предметной области историко-культурного наследия Республики Беларусь в соответствии с предложенной системой (тайгенов и ёгенов) в ручном режиме. Рассматриваются конкретные приемы и методы конвертации. Подтверждается перспективность электронного словаря частей языка, демонстрируются его фрагменты.

К л ю ч е в ы е с л о в а: *конвертация; части речи; части языка; лексическая разметка; предметная область; историко-культурное наследие; тайген; ёген.*

The article is devoted to the conversion of parts of speech into parts of language (categories of the Sino-Japanese linguistic tradition). It demonstrates the semantic transformation of Russian lexical units of the subject area of the historical and cultural heritage of the Republic of Belarus in accordance with the proposed system (taigens and yogens) in manual mode. Specific conversion techniques and methods are considered. The prospects of the electronic dictionary of parts of the language are confirmed, its fragments are demonstrated.

K e y w o r d s: *conversion; parts of speech; parts of language; lexical markup; subject area; historical and cultural heritage; taigen; yogen.*

Идеи возникновения концепций частей речи уходят своими корнями в глубокую древность. Ученые Александрийской школы (например, Дионисий Фракийский) зафиксировали наличие восьми частей речи для древнегреческого языка: имя, глагол, артикль, причастие, наречие, местоимение, предлог, союз. Учение о восьми частях речи содержалось в первых грамматиках церковнославянского языка XVI–XVII вв. В работах Ломоносова также рассматривались восемь частей речи, что соответствовало латинскому варианту, но в своей «Грамматике» Ломоносов, в отличие от ученых древности, разграничил знаменательные и служебные части речи на две категории. Знаменательные части речи указывают на предмет, явления, факты реального мира, то есть выполняют номинативную роль. Служебные части речи включают слова, которые не называют ни предметов, ни признаков, а обозначают лишь отношения между ними. В русском языкознании известный отход от античных схем грамматических воззрений начинается в XIX веке с грамматики А. Х. Востокова, в которой причастие как

часть речи было заменено прилагательным. Г. П. Павский и Ф. И. Буслаев ввели имя числительное. Последними, уже в XIX веке, в структуре частей речи были введены частицы [1].

«Частями речи называются слова, выделяемые в некоторые классы по форме, значению и функции» [1]. Особая роль в этом определении отводится форме... Если следовать рассматриваемому определению, то можно сказать, что в китайском языке части речи отсутствуют, так как в нем нет ни склонения существительных, ни спряжения глаголов и т. п. Не удивительно поэтому, что «проблема частей речи обсуждалась 50 лет и это обсуждение не принесло никаких результатов...» [1].

Если ремонт – существительное, тогда как объяснить синтаксис сочетания ремонт автомашин? Известно, что в русском языке имена согласуются в роде, числе и падеже, здесь же одно имя ремонт, словно глагол, управляет другим именем автомашин. С точки зрения частей речи это досадное исключение, «языковой парадокс», а с точки зрения частей языка – стандартный синтаксис, ибо ёгенам вне зависимости от их морфологии свойственно сочетаться с тайгенами. Новая теория тогда чего-либо стоит, если она выясняет все, что выясняла старая, а также то, что старая выяснить не могла. Со времен атомистов парадигма в физике менялась много раз, лингвистика же никак не справится с противоречиями в греко-римских частях речи. Если последовательно применять семантический критерий в разграничении элементов языковой системы и строго следовать правилам вывода, то неизбежно возникает понимание частей языка как подмножеств языковой системы, элементами которых являются знаки с общим предельно абстрактным значением, а именно: обозначающие индивидов в модели мира тайгены и признаки индивидов ёгены. Здесь мы не открываем ничего нового, лишь, выражаясь словами Конфуция, “возрождаем утраченное” [2, с. 17].

В традиционном учении о частях речи присутствует ряд противоречий, которые вызывают у ученых, в течении всей истории развития лингвистики, сомнения в его научной достоверности. «Например: наряду с существительными, прилагательными, глаголами и наречиями, обозначающими, соответственно, предметы, признаки предметов, процессы и признаки процессов, в самостоятельные части речи выделены местоимения и числительные, которые по содержанию являются разновидностями существительных (*он, это, восемь*) или прилагательных (*его, этот, восьмой*). Слово *столовая* (место приема пищи) вопреки морфологии считается существительным, в то же время среди существительных оказывается и слово *бег* с глагольной семантикой. К междометиям относят слова, выражающие, но не называющие эмоции и волеизъявления. Вместе с тем, во фразе: «*А девица хи-хи-хи да ха-ха-ха!*» слова *хи-хи-хи* и *ха-ха-ха* рассматриваются как междометия, хотя они обозначают процесс и играют роль сказуемого» [3, с. 174]. К сожалению, избежать противоречий в рамках греко-латинской классификации невозможно, поскольку между синтаксисом, морфологией и семантикой не однозначного соответствия. Решить данную проблему призвана парадигма частей языка [4, с. 44], восходящая к китайско-японской лингвистической

традиции [5]. С исходными понятиями и описанием тайгенов и ёгенов можно ознакомиться подробнее в работе А. Н. Гордея: «Парадигма частей языка» [3, с. 176].



Рис. 1. Графическая репрезентация тайгенов.



Рис. 2. Графическая репрезентация ёгенов.

Целью нашей работы является преобразование частей речи в части языка и создание словаря частей языка. Нами были взяты фрагменты текстов из истории культурного наследия Республики Беларусь, которые мы разбили на сегменты (слова) и представили в виде словаря. В словаре разграничиваются категориальные признаки в соответствии с парадигмой частей языка А. Н. Гордея, представленной на рисунке 1 и 2.

Например, слово «*богатое*» в терминах частей речи рассматривается как – прилагательное, в терминах частей языка как – ёген (свернутый – сокращенный – физический – постоянный – 2-й положительной степени – многоместный – интенсивный – нарицательный – качественный – произвольный – относительный – повествовательный).

Рассмотрим слово «*археологическое*». В терминах частей речи рассматривается как – прилагательное, в терминах частей языка как – ёген (развернутый – сокращенный – физический – постоянный – 1-й степени – многоместный – интенсивный – нарицательный – качественный – произвольный – относительный – повествовательный).

В терминах частей языка слово: «*культурный*» – ёген (свернутый – сокращенный – информационный – постоянный – 2-й положительной степени – многоместный – интенсивный – повествовательный – нарицательный – качественный – произвольный – относительный).

Например, также слово «*историко-культурный*» – ёген (развернутый – составной – информационный – постоянный – 1 степени – интенсивный – повествовательный – нарицательный – качественный – произвольный – относительный).

В терминах частей речи «*бег*» – существительное, в терминах частей языка – ёген (свернутый – сокращенный – физический – постоянный – 1-й степени – повествовательный – нарицательный – качественный – произвольный – относительный).

	A	B	C	D	E	F	G	H	I	J	K	L
1	Лексема	Ошибка	ЗАС	Тайген	Ёген	Развернутый	Свернутый	Составной	Сложный	Сокращенный	Сжатый	Информационный
2	культурный				1		1			1		1
3	культурного				1		1			1		1
4	культурному				1		1			1		1
5	культурным				1		1			1		1
6	культурном				1		1			1		1
7	культурная				1		1			1		1
8	культурной				1		1			1		1
9	культурную				1		1			1		1
10	культурною				1		1			1		1
11	культурное				1		1			1		1
12	культурные				1		1			1		1
13	культурных				1		1			1		1
14	культурными				1		1			1		1
15	культурен				1		1			1		1
16	культурна				1		1			1		1
17	культурно				1		1			1		1
18	культурны				1		1			1		1
19	культурные				1		1			1		1

Рис. 3. Представление ёгенов в словаре частей языка.

Тайгены в словаре представлены аналогичным образом. Например, в терминах частей речи «*родословная*» – существительное, в терминах частей языка – тайген (развернутый – сложный – информационный – постоянный – нарицательный – качественный – многоместный – интенсивный).

Например, также слово «злотый» – существительное, в терминах частей языка – тайген (свернутый – сокращенный – сжатый – физический – постоянный – 1 степени – нарицательный – качественный – повествовательный).

В терминах частей языка слово: «подлежащее» – тайген (свернутый – сокращенный – физический – переменный – одностепенный – интенсивный – повествовательный), а в терминах частей речи – существительное.

1	Лексема	Ошибка	ЗАС	Тайген	Еген	Развернутый	Свернутый	Составной	Сложный	Сокращенный	Сжатый	Информационный
18926	биографические				1	1			1			1
18927	биографических				1	1			1			1
18928	биографическими				1	1			1			1
18929												
18930	родословная			1		1			1			1
18931	родословной			1		1			1			1
18932	родословную			1		1			1			1
18933	родословною			1		1			1			1
18934	родословные			1		1			1			1
18935	родословных			1		1			1			1
18936	родословным			1		1			1			1
18937	родословными			1		1			1			1

Рис. 4. Представление тайгенов в словаре частей языка.

«Парадигмы частей языка достаточно для переписывания основ китайского синтаксиса [6, с. 22]. Перспективность ее для английского и русского языков также не вызывает сомнений, потому что путем бесконечного членения множества тайгенов и множества ёгенов на подмножества и подмножества подмножеств достигается любой уровень дискретности в репрезентации языковых единиц» [3, с. 175].

Данный словарь является контекстно-свободным и применим для любого нетерминала, т.е. объекта, обозначающего какую-либо сущность языка (например, – формулы арифметического выражения, команды), причем, независимо от контекста этого нетерминала. Любой контекст проходит стадию лемматизации и теггирования.

Теггирование – маркировка, нанесение тегов на поверхность предложения. Тег – метка, которая размечает информацию и позволяет облегчить процесс поиска необходимой информации. Например, слову *лектор* присваивается тег: тайген, свернутый, сокращенный, переменный, множественный, интенсивный, качественный, относительный. Одному знаку также могут принадлежать разные теги: *коса* (девичья), *коса* (инструмент для покоса травы и т.д.); *он* – для обозначения как живых индивидов, так и неживых и т.д.

Решить проблему омонимии на синтаксическом уровне признан парсер. Парсер – это программное обеспечение (ПО), выделяющее определенные части информации из массива данных по определенным правилам.

После теггирования, с опорой на словарь А. А. Зализняка, нами составляются парадигмы для каждой части языка и выделяются леммы. Лемма – первоначальная словарная форма слова [7]. Лемматизация (англ. lemmatization) – метод морфологического анализа, который сводится к приведению словоформы к ее первоначальной словарной форме (лемме).

В результате лемматизации от словоформы отбрасываются флексивные окончания и возвращается основная или словарная форма слова. Например, в русском языке словарной формой для существительного считается именительный падеж, единственное число (*мечами* – *меч*); для глагола – инфинитивная форма (*читали* – *читать*); для прилагательного – единственное число, именительный падеж, мужской род (*заснеженными* – *заснеженный*) [8].

Тайгены в именительном падеже в единственном и множественном числе рассматриваются как разные леммы, а ёгены во множественном числе – варианты одной леммы.

«Необходимо отметить, что семантическая категория количества не имеет своим прямым коррелятом морфологическую категорию числа. Во-первых, не все греко-латинские части речи имеют парадигму единственного и множественного числа. Во-вторых, для некоторых частей речи категория числа оказывается не связанной с их семантикой (знак *белые* имеет форму множественного числа, однако обозначает только одно свойство индивидов, знак *пестрый* имеет форму единственного числа, но обозначает некоторую совокупность свойств; *умерли* – единократный процесс, несмотря на морфологическую форму множественного числа, *стучал* – многократный процесс, несмотря на единичность участвующего в нем индивида), и для того, чтобы описать количественность обозначаемых данными частями речи явлений, требуется введение дополнительных категорий: итеративность или фреквентативность для глаголов [9, с. 450–451]. В-третьих, даже для частей языка, которые обозначают индивидов, а не их признаки, морфологическая категория числа не всегда верно отражает количественный характер этих индивидов: знак *Афины* имеет форму множественного числа, однако обозначает один индивид, в то время как знак *толпа*, имеющий форму единственного числа, служит для обозначения множества, некоторой совокупности индивидов. Таким образом, применение морфологической категории числа в анализе языка весьма ограничено и даже в рамках этой ограниченной области не всегда дает однозначные результаты относительно отражения идеи единичности/множественности в языке.

«Для обозначения категорий языка, соответствующих категории количества в модели мира, мы будем придерживаться терминологии А. Н. Гордея [3, с. 174] и вслед за ним выделять одноместные, то есть обозначающие один индивид или один признак индивида, и множественные, то есть обозначающие множество индивидов или множество признаков индивидов, части языка» [10, с. 96–97]. Тогда *Афины* – одноместный тайген, *толпа* – множественный тайген; *белые*, *стукнули* – одноместные ёгены, а *пестрый*, *стучал* – множественные ёгены.

«Для одноместных и множественных, интенсивных и экстенсивных частей языка пока были приведены лишь декларативные и интуитивные определения, которых недостаточно для полного «количественного» анализа языка. Необходимо дать процедуральные определения, которые, во-первых, помогут формализовать количественный аспект языка, а во-вторых, позволят

с помощью установленных процедур однозначно разграничить части языка в сложных случаях». Эти определения и процедуры были даны и установлены А. Н. Гордеем. Прежде всего сформулируем вышеприведенные определения, используя математическую терминологию. Итак, одноместные части языка обозначают множество с числом индивидов равным одному, или мощностью один; многоместные части языка обозначают множество с числом индивидов больше одного, или мощностью больше единицы. Тогда процедуральные определения будут звучать следующим образом. Одноместные – части языка, обозначающие целое множество, многоместные – части языка, обозначающие целое множество и его некоторую часть. Интенсивные тайгены и ёгены обозначают целое множество и его некоторую часть, представленную собственным подмножеством с числом индивидов больше одного, экстенсивные же тайгены и ёгены называют целое множество и его некоторую часть, представленную группой подмножеств с числом индивидов больше либо равно одному [3, с. 178–179].

Процедура конвертации частей речи в части языка помогает устанавливать одно-однозначное соответствие между синтаксическим и семантическим анализом предложения.

Создание электронного словаря и ручная лексическая разметка тестов предметной области историко-культурного наследия Беларуси в семантических категориях частей языка (субстантивах-тайгенах и предикативах-ёгенах) являются обязательным предварительным этапом для построения алгоритмов машинного обучения.

ЛИТЕРАТУРА

1. Методика изучения частей речи [Электронный ресурс]. – 2016. – Режим доступа : <https://ik-ptz.ru/dictations-on-the-russian-language-grade-5/metodika-izucheniya-chastei-rechi-referat-istoriya-izucheniya-chastei.html>. – Дата доступа : 22.04.22.
2. Гордей, А. Н. Semantic Categories in the Sino-Japanese Linguistic Tradition / А. Н. Гордей // Иностранные языки в высшей школе : научный журнал. – 2018. – № 4 . – С. 7–22.
3. Гордей, А. Н. Парадигма частей языка / А. Н. Гордей // Словообразование и номинативная деривация в славянских языках : материалы VIII Междунар. науч. конф., Гродно, 15–16 апр. 2003 г. / Гродн. гос. ун-т ; отв. ред.: С. А. Емельянова [и др.]. – Гродно, 2003. – С. 173–179.
4. Заде, Л. Понятие лингвистической переменной и его применение к принятию решений. – М. : Мир, 1976. – С. 167.
5. Цзин, Чжаоцзы. Исследование грамматики национального языка. – Шанхай : Языки и языкознание, 1955. – С. 199.

6. *Задоевко, Т. П., Хуан Шуин.* Учебник китайского языка. М. : Цитадель, 2002. – С. 79.
7. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Е. И. Большакова [и др.]. – М.: НИУ ВШЭ, 2017.
8. Лемматизация. [Электронный ресурс]. – Режим доступа : <https://promopult.ru/library/> / Лемматизация. – Дата доступа : 22.04.2022.
9. *Теньер Л.* Основы структурного синтаксиса. М., 1988.
10. *Руденко, Н. И.* Языковые категории интенсивности и экстенсивности в процедуральном представлении / Н. И. Руденко // Весн. Бел. дзярж. ун-та. Сер. 4, Філ. Журн. Пед. – 2006. – № 2. – С. 94–100

Поступила в редакцию 09.12.2022