МРНТИ 20.23.25, 16.31.21, 28.23.39
УДК 004.82; 004.912; 81.322.2

https://doi.org/10.51889/2020-3.1728-7901.30

Л. Жеткенбай[1], Б.Ш. Разахова[1], Б. Ергеш[1], А.С. Муканова[1]

[1]Евразийский национальный университет им. Л.Н.Гумилева, г. Нур-Султан, Казахстан

РАЗРАБОТКА ОНТОЛОГИЧЕСКОЙ МОДЕЛИ ПРОСТОГО ПРЕДЛОЖЕНИЯ ТУРЕЦКОГО ЯЗЫКА

Аннотация

В статье описаны синтаксические правила предложений турецкого языка и показаны их деревья составляющих посредством формальной грамматики Хомского. Вместе с тем, построена онтологическая модель синтаксических правил простых предложений турецкого языка с учетом его семантики. Для обозначения синтаксических категорий и понятий в предлагаемых онтологических моделях используются термины из унифицированной метаязыка UniTurk. Результат этих работ могут быть использованы для решения задач NLP, например, в системах извлечения знаний, системах информационного поиска, вопросно-ответных системах, машинного перевода, автореферирования турецкого текста, а также в информационно-справочных и обучающих системах, а также в дальнейшем могут быть использованы и планируются к использованию при построении онтологических моделей синтаксических правил других тюркских языков.

Ключевые слова: обработка естественного языка, синтаксические правила, лингвистические разметки, база знаний, онтологическое моделирование.

Аңдатпа

Л. Жеткенбай[1], Б.Ш. Разахова[1], Б. Ергеш[1], А.С. Муканова[1]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан қ., Қазақстан

ТҮРІК ТІЛІ ЖАЙ СӨЙЛЕМІНІҢ ОНТОЛОГИЯЛЫҚ МОДЕЛІН ӘЗІРЛЕУ

Бұл мақалада Хомскийдің формалды грамматикасының көмегімен түрік тіліндегі жай сөйлемдердің синтаксистік ережелері сипатталған және құраушы ағаштары көрсетілген. Сонымен қатар түрік тілінің семантикасын ескере отырып, жай сөйлемдердің синтаксистік ережелерінің онтологиялық моделі тұрғызылған. Ұсынылатын онтологиялық моделдерде синтаксистік категориялар мен түсініктерді белгілеу үшін UniTurk метатілі терминдері қолданылды. Бұл жұмыстың нәтижелерін NLP есептерін шешу үшін қолдануға болады, мысалы, білім алу жүйелерінде, ақпараттық іздеу жүйелерінде, сұрақ-жауап жүйелерінде, машиналық аудармада, түрік мәтінін автоматты түрде рефераттауда, сонымен қатар ақпараттық-анықтамалық және оқыту жүйелерінде, бұдан басқа түркі тілдерінің синтаксистік ережелерінің онтологиялық онтологиялық модельдерін құру кезінде пайдаланылуы мүмкін және де пайдаланылу жоспарлануда.

Түйін сөздер: табиғи тілді өңдеу, синтаксистік ережелер, лингвистикалық белгілер, білімдер базасы, онтологиялық модельдеу.

Abstract

DEVELOPMENT OF AN ONTOLOGICAL MODEL OF SYNTACTIC RULES OF SIMPLE SENTENCES OF TURKISH LANGUAGE

Zhetkenbay L.[1], Razakhova B.[1], Yergesh B.[1], Mukanova A.[1]

[1]L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

This article describes the syntactic rules of sentences in Turkish language and presented its tree components as well by means of formal grammars Chomsky. At the same time, an ontological model of syntactic rules of simple sentences of the Turkish language is constructed, taking into account its semantics. The proposed ontological models use terms from the unified metalanguage UniTurk to denote syntactic categories and concepts. The results of this work can be used to solve NLP tasks, for example, in the systems of knowledge, information retrieval, question and answering systems, in machine translation, automatic summarization of Turkish texts, as well as in the reference and training systems, moreover, in building the ontological model of syntax rules of Turkic languages and is planned to use.

Keywords: natural language processing, syntactic rules, linguistic markup, knowledge base, ontological modeling.

1. Введение

В настоящее время в связи с резким увеличением объема информации на естественных языках в интернете и социальных сетях исследование и разработки в области компьютерной лингвистики становятся чрезвычайно актуальными. Как известно, компьютерная лингвистика является новым научным направлением и входит в состав искусственного интеллекта, который также является новым

направлением информатики (вычислительной науки). Компьютерная лингвистика включает в себе компьютерную обработку естественных языков (ОЕЯ) – Natural Language Proccesing (NLP).

Для компьютерной обработкой любых естественных языков требуются, во-первых, формализация их грамматических (морфологических и синтаксических) правил, во-вторых, разработка алгоритмов анализв и синтеза слов и предложении по этим правилам, в-третьих, программная реализация всех этих алгоритмов, в-четвертых, построение текстовых корпусов (база данных размеченных текстов) и аудиокорпусов (база данных размеченных аудиозаписей) и других программ для анализа и обработки текстов, например, сентимент анализ.

Онтологическая модель синтаксических правил турецкого языка была создана в соответствии с целью проекта, которая является разработкой единого многоязычного электронного тезауруса тюркских языков для многоязычного поиска и извлечения знаний.

В качестве инструмента моделирования предметной области выбран редактор онтологий Protégé. Protégé – это свободный, открытый редактор онтологий и фреймворк для построения баз знаний. Платформа Protégé поддерживает два основных способа моделирования онтологий посредством редакторов Protégé-Frames и Protégé-OWL. Онтологии, построенные в Protégé, могут быть экспортированы во множество форматов, включая RDF (RDF Schema), OWL и XML Schema.

Protégé имеет открытую, легко расширяемую архитектуру за счѐт поддержки модулей расширения функциональности [1].

Прикладная онтология «Синтаксис турецкого языка» была реализована с соответствии с синтаксическими описаниями, которые указаны выше и состоит из отдельных индивидов, свойств и классов, а также функций интерпретации, заданных на концептах или отношениях онтологии.

2. Разработка онтологической модели простого предложения турецкого языка

Известно, что формальная грамматика Хомского позволяет описать синтаксис заданного языка [2], а онтологические модели не только его синтаксис, но и семантику. Можно также сказать, что онтология – это база знаний, потому что если добавить интерпретирующие функции к структурно-семантической модели, то она станет базой знаний [3-5].

В данном разделе с помощью формальной грамматики Хомского описаны синтаксические правила предложений турецкого языка и показаны их деревья составляющих, а также построены онтологии синтаксических правил простых предложений турецкого языка с учетом их семантики.

Для формализации повествовательных простых предложении турецкого языка прежде всего необходимо ввести специальные лингвистические разметки – тэги. Тэги для описания структуры предложении турецкого языка в системе UniTurk [6] представлены в таблице 1.

Контекстно-свободная грамматика Хомского (CFG) наиболее широко используемая формальная система для моделирования составных структур в естественных языках.

Контекстно-свободная грамматика состоит из набора правил или правил выводов, каждое из которых выражает способы, которыми символы языка могут быть сгруппированы и упорядочены вместе. Список сокращении и их значения описаны в 1-таблице.

Таблица 1. Тэги для описания структуры предложении турецкого языка

Тэг	*Английский*	*Русский*	*Казахский*	*Турецкий*
S	*Simple sentence*	*Простое предложение*	*Жай сөйлем*	*Yalın cümle*
Sub	*Subject*	*Подлежащее*	*Бастауыш*	*Özne*
Obj	*Object*	*Дополнение*	*Толықтауыш*	*Tümleç*
Obj 1	*Object*	*Дополнение*		*Nesine*
Abr	*Abreviatura*	*Определение*	*Анықтауыш*	
Adl	*Adverbial*	*Обстоятельство*	*Пысықтауыш*	*Zarf tümleci*
Pre	*Predicate*	*Сказуемое*	*Баяндауыш*	*Yüklem*

КС-грамматика общего вида G определяется следующими параметрами [7]:

$$G = < N_s, T_s, R, S > \quad (1)$$

где:

N_s– множество нетерминальных символов (переменных);

T_s– множество терминальных символов (констант): при этом $N_s \cap T_s = \emptyset;$

R – множество правил вывода вида $A \to \alpha$, где A - нетерминальный символ, α - строка символов в

алфавите $Ns \cup Ts$ т.е. $\alpha \in (Ns \cup Ts)^*$;

S – начальный нетерминальный символ

Структуру предложении можно представить из двух частей: именное, глагольное. Синтаксис повествовательных простых предложений турецкого языка можно описывать с помощью конкретной КС-грамматики.

Например, пусть заданы следующие простые предложения турецкого языка:

1. “Samat kitap okuyor” - “ Самат читает книгу ”;
2. “ Samat kütüphanede kitap okuyor ” - “ Самат читает книги в библиотеке ”;
3. “Samat annesiyle dün geldi” - “ Самат пришел вчера со своей матерью”;
4. “En yakın arkadaşım Samat okuyor ” - “ Мой лучший друг Самат читает”;

Чтобы описать структуры этих предложений для параметров КС грамматики присвоим следующие значения:

$N_s = \{S, NP, VP, Adj, Adv\}$

$T_s = \{S, a, d, e, g, h, ı, i, k, l, m, n, o, p, r, t, s, ş, u, ü, y\}$

$R = \{S \rightarrow NP \mid VP, NP \rightarrow N \mid N \mid Adj \mid Adv, VP \rightarrow N \mid V \mid Adv \mid NP \mid VP\}$

Используя правила этой грамматики деревья составляющих вышеуказанных предложений турецкого языка представлены на рисунках 1-4:

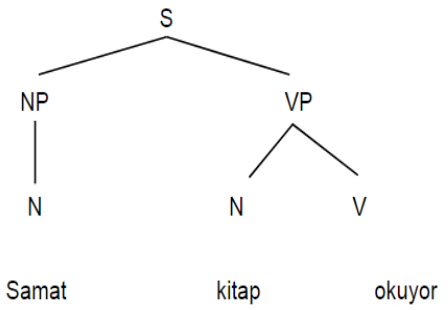

Рисунок 1. Дерево составляющих S(NP(N),VP(N,V))

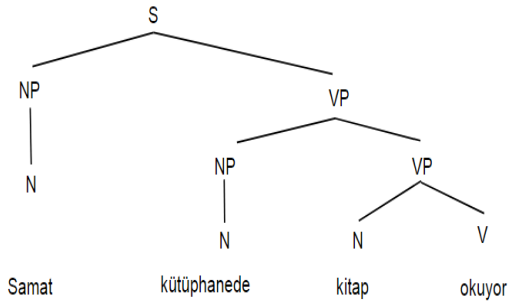

Рисунок 2. Дерево составляющих S(NP(N),NP(N,N)),VP(N,V)

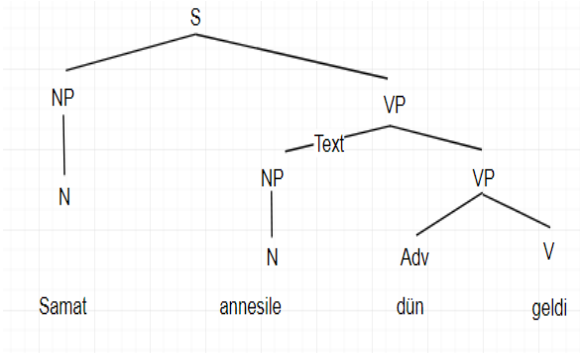

Рисунок 3. Дерево составляющих S(NP(N), VP(NP(N), VP(Adv,V)))

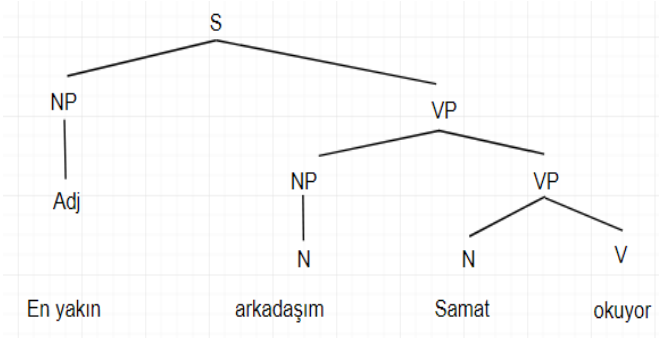

Рисунок 4. Дерево составляющих S(NP(Adj), VP(NP(N), VP(N,V)))

На рисунке 5 представлен тип предложения SubObjPre (Özine, Tümleç, Yüklem). Для этого типа предложения должны быть выполнены следующие необходимые и достаточные условия:

$$\begin{aligned} \text{SubObjPre} \equiv\ & \exists hasNP(PL\ or\ Pers) \\ & \sqcap \exists hasVP\ \left(DirVP1\right. \\ & \left. \sqcap\ (\exists hasHead\ \left(V \sqcap \left(\exists hasRoot\ \left(root\ \sqcap\ (\forall isSpace\ noSpace)\right)\right)\right)\right)\right) \end{aligned} \tag{2}$$

где:

- SubObjPre – Типы предложения состоящие из подлежащего, дополнения и сказуемого (Özine, Tümleç, Yüklem)
- $hasNP$ – имеет именное словосочетание;
- PL – слова во множественном числе
- $Pers$- личные местоимения
- $hasVP$ – имеет глагольное словосочетание;
- $DirVP1$ – глагольное управление (имя существительное в направительном падеже + глагол)
- $hasHead$ – имеется главное слово;
- V− глагол
- $hasRoot$ –имеет корень
- $root$ – корень
- $isSpace$ –является ли словом с пространственным значением
- $noSpace$–*непространственная семантика*

Если все условия соблюдены тогда при запуске резонера в среде Protégé, предложение *«Çocuklar yemeğe doydu»* который является индивидом концепта Sentence (предложение) определяется как тип предложения $SubObjPr$, так как глагол «doy» является непространственным глаголом, а также выполнены необходимые и достаточные условия этого типа предложения.

На рисунке 6 представлен тип предложения «Онтологическая модель предложения турецокого языка типа $SubObjPre$» [8, 9].

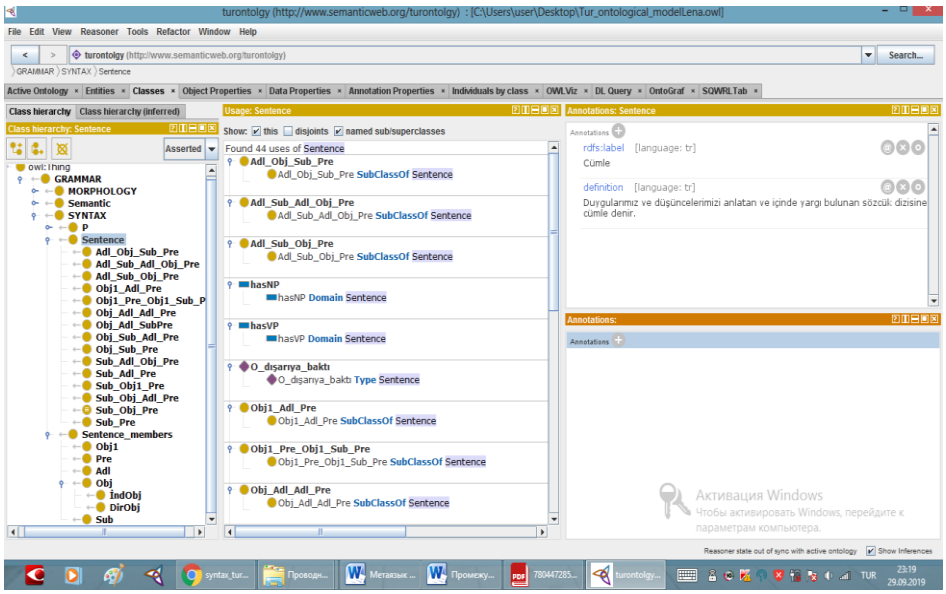

Рисунок 5. Фрагмент структуры онтологической модели предложения турецкого языка

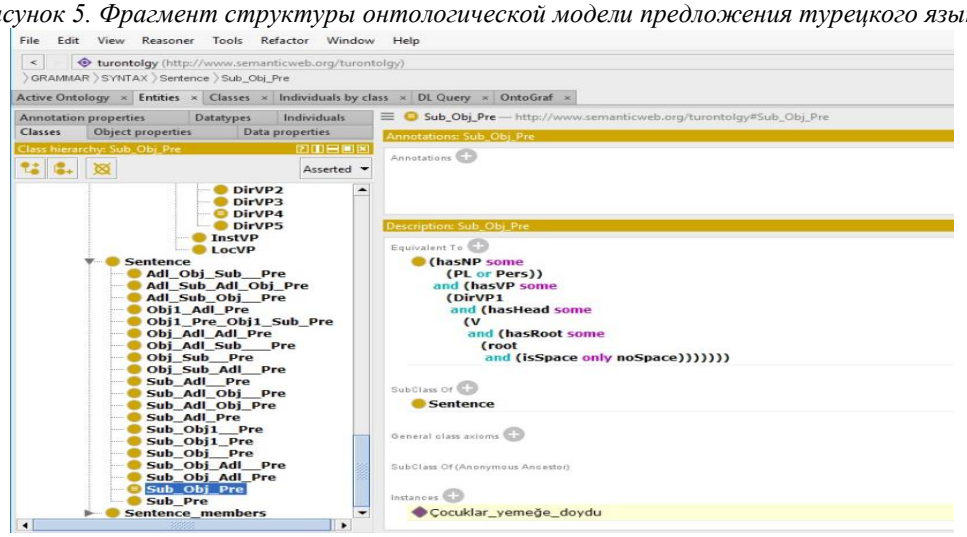

Рисунок 6. Онтологическая модель предложения турецокого языка типа $SubObjPre$

Заключение

Полученная онтологическая модель предметной области позволяет формализовать ее знания и представить онтологии как ее спецификацию. Практическая ценность полученных результатов состоит в том, что разработанный метаязык и построенные онтологические модели синтаксических правил турецкого языка будет использованы в проведении сравнительного анализа этих языков с указанием качественных и количественных показателей, в разработке онлайн обучения этим языкам, создания систем многоязычного поиска и извлечения знаний, а также машинного перевода между этими языками.

Работа выполнена при поддержке грантового финансирования научно-технических программ и проектов Министерством науки и образования Республики Казахстан (грант № AP05132249, 2018-2020 годы).

Список использованной литературы:

1 https://ru.wikipedia.org/wiki/Prot%C3%A9g%C3%A9

2 Chomsky N. Syntactic Structures. — The Hague: Mouton, 1957. (Переиздание: Chomsky N. Syntactic Structures. – De Gruyter Mouton, 2002. – ISBN 3-11-017279-8.)

3 Цуканова Н. И. Онтологическая модель представления и организации знаний. – Москва: Горячая линия – Телеком, 2015. – с. 272.

4 Лапшин В.А. Онтологии в информационных системах. – Москва, 2009. – с.247.

5 Bekmanova G., Sharipbay A., Altnbek G., Adalı E., Zhetkenbay L., Kamanur U., Zulkhazhav A. A uniform morphological analyzer for the Kazakh and Turkish languages / Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017) July 2017, Moscow, Russia. –P. 20-30.

6 Горшков С. Введение в онтологическое моделирование. – ООО «ТриниДата», 2016. – с.165.

7 Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.). - Prentice Hall PTR, Upper Saddle River, NJ, USA, 2009.

8 Щека Ю.В. Практическая грамматика турецкого языка. М.: АСТ: Восток–Запад, 2007. – 666 [6] с. – ISY 978-5-17-043016-1; ISY 978-5-478-00529-0.

9 Hengirmen, M.: Turkish Grammar (in Turkish) Ankara, (2007)

References

1 https://ru.wikipedia.org/wiki/Prot%C3%A9g%C3%A9

2 Chomsky N. (2002) Syntactis Structures. — The Hague: Mouton, 1957. (Pereizdanie: Chomsky N. Syntactis Structures. – De Gruyter Mouton. – ISBN 3-11-017279-8.) (In English)

3 Cukanova N. I. (2015) Ontologicheskaja model' predstavlenija i organizacii znanij [Ontological model of knowledge representation and organization]. Moskva: Gorjachaja linija, Telekom. 272. (In Russian)

4 Lapshin V.A. (2009) Ontologii v informacionnyh sistemah [Ontologies in information systems]. Moskva. 247. (In Russian)

5 Bekmanova G., Sharipbay A., Altnbek G., Adalı E., Zhetkenbay L., Kamanur U., Zulkhazhav A. (2017) A uniform morphological analyzer for the Kazakh and Turkish languages. Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017), Moscow, Russia. 20-30. (In English)

6 Gorshkov S. (2016) Vvedenie v ontologicheskoe modelirovanie [Introduction to ontological modeling]. OOO «TriniData». 165. (In Russian)

7 Jurafsky D., Martin J. H. (2009) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.). - Prentice Hall PTR, Upper Saddle River, NJ, USA. (In English)

8 Shheka Ju.V. (2007) Prakticheskaja grammatika tureckogo jazyka [Practical grammar of the Turkish language]. ACT: Vostok–Zapad. 666. ISY 978-5-17-043016-1; ISY 978-5-478-00529-0. (In Russian)

9 Hengirmen, M. (2007) Turkish Grammar. Ankara. (In Turkish)