СЕДЬМАЯ МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ «TURKLANG 2019»

Труды конференции

КАЗАНЬ
2019

УДК 004.8+81’32
ББК 81.1

Организаторы:

Крымский федеральный университет имени В. И. Вернадского
Институт иностранной филологии

Крымский инженерно-педагогический университет
Факультет истории, искусств и крымскотатарского языка и литературы

Академия наук Республики Татарстан
Институт прикладной семиотики

Евразийский национальный университет имени Л. Н. Гумилёва Министерства образования и науки Республики Казахстан
НИИ «Искусственный интеллект»

Стамбульский технический университет

Российская ассоциация искусственного интеллекта

Евразийский институт развития им. Исмаила Гаспринского

Крымская республиканская универсальная научная библиотека имени И. Я. Франко

Научные редакторы:
к. филол. н. Кубединова Л. Ш.

Седьмая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2019»: Труды конференции. – Казань: Издательство Академии наук Республики Татарстан, 2019. – 351 с.
ISBN 978-5-9690-0548-8

Сборник содержит материалы Седьмой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2019» (Симферополь, Крым, Россия, 3–5 октября 2019 г.)

Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной лингвистики и ее приложений.

УДК 004.8+81’32
ББК 81.1

ISBN 978-5-9690-0548-8

ПРЕДИСЛОВИЕ

В 2019 году на базе Крымского федерального университета имени В. И. Вернадского, и Крымского инженерно-педагогического университета имени Февзи Якубова, а также с участием Евразийского института развития имени Исмаила Гаспринского прошла седьмая международная конференция по компьютерной обработке тюркских языков TurkLang-2019.

Предыдущие конференции прошли в Астане (2013), Стамбуле (2014), Казани (2015, 2017), Бишкеке (2016), Ташкенте (2018). Расширение географии проведения и состава участников конференции подтверждают, что тематика конференции остается весьма актуальной. Целью серии международных конференций TurkLang является создание пространства совместных компьютерных лингвистических исследований для тюркских языков. На конференции представляются новые результаты, связанные с разработкой компьютерных лингвистических приложений для тюркских языков. Это связано со сложной и самобытной грамматической и семантической системами тюркских языков, для которых простой перенос решений, полученных на материале других языков не всегда возможен. Организаторы конференции надеются, что участие в ее проведении целого ряда ведущих научно-образовательных учреждений Крыма послужит очередным толчком для развития данного направления компьютерной лингвистики в Республике Крым.

В сборник трудов включены статьи участников VII Международной конференции по компьютерной обработке тюркских языков «TurkLang-2019» (Симферополь, Крым, Россия, 3–5 октября 2019 г.). Участниками конференции были ученые и специалисты из Узбекистана, Казахстана, Кыргызстана, Турции, Азербайджана, России (Татарстан, Башкортостан, Москва, Саха (Якутия), Чувашия, Тува, Крым, Алтай и др.). Они представили доклады, посвященные актуальным проблемам компьютерной лингвистики в плане их разрешения в контексте тюркских языков. В ходе конференции активно и плодотворно обсуждались вопросы разработки концептуальных и формальных лингвистических моделей, лингвопроцессоров, систем машинного перевода, электронных корпусов, речевых технологий, а также проблемы функционирования тюркских языков в Интернет-технологиях. Также в рамках конфе-

ренции была проведена демонстрация программных продуктов, на которой татарские и крымскотатарские IT-разработчики представили различные электронные крымскотатарские словари, мобильные приложения и интерактивный самоучитель для изучения крымскотатарского языка, русско-татарский машинный перевод, синтезатор татарской речи, морфологические анализаторы татарского языка и др. В этом году участники конференции продолжили обсуждение реализации совместного проекта по созданию компьютерной онтологии тюркской грамматики. Как показывает это обсуждение, унификация систем понятий и терминов не является тривиальной практической задачей и требует большой и кропотливой работы в пересмотре многих традиционных грамматических описаний для тюркских языков.

Организаторы конференции выражают благодарность директору Евразийского института развития имени Исмаила Гаспринского, Сулейманову М., директору Института иностранной филологии КФУ им. В. И. Вернадского, Петренко А. Д., декану Факультета истории, искусств и крымскотатарского языка и литературы, Ганиевой Э. С., за их вклад и успешное проведение конференции «TurkLang-2019».

ПРОГРАММНЫЙ КОМИТЕТ

1. Сулейманов Джавдет Шевкетович (Казань, Татарстан, РФ) – сопредседатель
2. Шарипбаев Алтынбек Амирович (Нур-Султан, Казахстан) – сопредседатель
3. Ешреф Адалы (Стамбул, Турция) – сопредседатель
4. Петренко Александр Демьянович (Симферополь, Крым, РФ)
5. Абдурахмонова Нилуфар (Ташкент, Узбекистан)
6. Алтынбек Гулила (Урумчи, Китай)
7. Гатиатуллин Айрат Рафизович (Казань, Татарстан, РФ)
8. Дыбо Анна Владимировна (Москва, РФ)
9. Желтов Валериан Павлович (Чебоксары. Чувашия, РФ)
10. Исраилова Нелла Амантаевна (Бишкек, Кыргызстан)
11. Кубединова Ленара Шакировна (Симферополь, Крым, РФ)
12. Мамедова Масума Гусейновна (Баку, Азербайджан)
13. Офлазер Кемаль (Доха, Катар)
14. Садыков Ташполот (Бишкек, Кыргызстан)
15. Салчак Аэлита Яковлевна (Кызыл, Тыва, РФ)
16. Сиразитдинов Зиннур Амирович (Уфа, Башкортостан, РФ)
17. Татевосов Сергей Георгиевич (Москва, РФ)
18. Торотоев Гаврил Григорьевич (Якутск, Саха, РФ)
19. Тукеев Уалишер Ануарбекович (Алматы, Казахстан)
20. Якубов Чингиз Февзиевич (Симферополь, Крым, РФ

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

Председатель: Петренко А. Д. д. филол. н., профессор, ФГАОУ ВО «КФУ им. В. И. Вернадского» (Симферополь)

Члены организационного комитета:

1. Кубединова Ленара Шакировна
2. Гатиатуллин Айрат Рафизович
3. Храбскова Данута Михайловна
4. Полховская Елена Васильевна
5. Беловенцева Мария Владимировна
6. Ганиева Эмине Сулеймановна
7. Гордиенко Татьяна Петровна
8. Саттарова Зера Мамбетовна
9. Заатов Исмет Аблятифович
10. Ясинова Елена Валентиновна
11. Сулайманов Мухаммед Али
12. Гафарова Ранетта Исметовна

УДК 81’33

DEVELOPMENT TECHNIQUES OF ANALYZING HOMONYMS IN THE LANGUAGES NOT POSSESSING NATIONAL CORPUS

Manzura Abjalova
Navoi State Mining Institute, Navoi, Uzbekistan
manzura_ok@mail.ru

Homonym forms have always been regarded as one of the events in the linguistic centre. For this reason, the texts in the context are studied separately in the Russian and foreign linguistics, depending on the linguistic corpus. But how can this problem be solved in the native languages of which there is no corpus? This article deals with the solution to this problem.

Keywords: corpus; homonym; model; tag; ID.

СОЗДАНИЕ ТЕХНИК АНАЛИЗА ОМОНИМОВ В ЯЗЫКАХ, КОТОРЫЕ НЕ ОБЛАДАЮТ НАЦИОНАЛЬНЫМ КОРПУСОМ

Манзура Абжалова
Навоийский государственный горный институт,
Навои, Узбекистан
manzura_ok@mail.ru

Омонимичные формы всегда рассматривались как одна из главных проблем в лингвистике. Этот вопрос в русском и зарубежном языкознании изучается с использованием электронных языковых корпусов. Но как решить эту проблему в тех языках, у которых нет электронных корпусов? Эта статья посвящена решению данной проблемы.

Ключевые слова: корпус; омоним; модель; тег; идентификатор.

It is well known that each of two or more lexical units which have the same spelling or pronunciation is called a homonym [1]. Homonymy includes a derivative lexeme, a root, and the equivalent of what happens when you add affixes expressing grammatical meanings. For example, the word «*terim»* is a combination consisting of adding a word-formation affix «*-im»* to the verb «*ter-«*, which is the result of the word form singular in the nominative case. To be more precise, several types of homonyms can be distinguished.

1) the part of the speech or the form of the dictionary (lemma);

2) some morphological features, such as case, affiliation or number (in rich morphological languages);

3) they are different from each other only in content (this occurs in semantic distinctions).

Because of the lack of a universal system (corpus) that covers the morphology of the language, it is addressed to the morphological dictionaries included in the linguistic supply in the combination of words in various words within the word sequences with affixes. Consequently, the concept of ambiguity can vary slightly for different dictionaries.

In addition to the morphological dictionaries, there is a complete collection of texts called corpuses [2]. Depending on the formation of the base, morphological, semantic and syntactic features of the shaped forms are also evident. As a result of the easy access to information, corpus becomes the perfect source of language information, for example, the most common type of static forms can be found.

Today, linguistic corpuses have become an indispensable tool of modern linguistics to deal with linguistic research and practical tasks, as the corpus serves to solve various linguistic tasks.

The corpus provides a convenient way to use in linguistic research, because it is processed on the basis of lexical, morphological, grammatical, semantic characters. According to Sh. Khamroeva, the linguistic corpus, in contrast to the electronic library, suggests the collection of necessary, useful and interesting texts for studying, learning and teaching a particular language.

Dependence on different bodies varies. It is also related to the use of not only different models of morphology, but also the style of texts. The textbooks incorporated into the base of the Russian language have a lot of good words. Therefore, the issue of elimination of ambiguity in the foreign and Russian linguistics («paperwork») is studied separately [6,7,8,9,10,11].

From grammatical sources, it is known that practically all methods of eradication are divided into two groups:

1. Grammar-based methods. In turn, they are grouped as follows:

– manual methods without interference from technology;

– automatic creation of rules.

2. Statistical methods.

Each of these groups has advantages and disadvantages. It is assumed that the resulting method of combination of methods in

both groups may show better results. It is important to note that in the studied sources, the phenomenon of survival was investigated by corpus linguistics. However, the corpus of many languages is still not created. Therefore, methods such as Brill method, the hidden Markov model, and modification of models cannot be used to analyse the concepts in the texts of the non-national languages.

Due to the fact that the Uzbek national corpus has not been created, a robust analysis technology was developed depending on N gram to analyze the authentic forms in the Uzbek language. For this purpose, the Uzbek language has been introduced into the linguistic supply by ID, which gives an opportunity to analyze the grammatical formations of the context. In order to improve the accuracy of the analysis, the species of ammonia has been identified and given a symbolic value. Then, the List of homonym (Models of Comp Words) model. The result is the Q – affix corresponding to the M – Deposit ID number on the LT, depending on the N – word that comes with M – Deposit. That is, M [id] + Q [id].

Here's how to analyze this method: *yeng*1 → n / a (part of garment covering), *yeng*2 → V / verb (win, head), *yeng*3 → V / verb (imperative form) is a grammatical tool that analyses the word syntax with the word combination (bigger, two-character sequence), in line with the symbols of the word combination in a syntactic module. Hence, here we refer to N and V related models.

In general, the method of deposit analysis is explained by the fact that a certain word is connected to the word through a certain grammatical index, which increases the probability and the accuracy of the analysis.

In conclusion, when studying the methods of elimination of ammonia in the world of computer linguistics, using this experience, the appropriate formulation of verbal expressions in Uzbek texts was created. To eliminate ammonia, each word must be «classified», which can be compared to a set of lemma and morphological features that are added to a tag for convenience. To find out all the possible tags, find morphological references to words in the morphological dictionary, or use a morphological analyser such as My Stem [12], which will help you find word tags. Then you only need to select the appropriate tag among multiple tags.

The optimal linguistic method used to analyse homonyms is an important factor in editing and analysing texts, machine translation, and text processing.

REFERENCED LITERATURE:

1. Раҳматуллаев Ш. Ўзбек тили омонимларининг изоҳли луғати (Rakhmatullaev Sh. Explanatory Dictionary of the homonyms of the Uzbek language). – Tashkent: Учитель, 1984. – P. 5.

2. Захаров В.П. Корпуса русского языка. www.dialog-21.ru/media/2138/zakharov.pdf.

3. Open the national site of the Russian corpus – http://opencorpora.org.

4. Недошивина Е. В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров. Учебно-методическое пособие. – СПб, 2006. – С. 26.

5. Khamraeva Sh. Linguistic foundations of the creation of the author corpus of the Uzbek language: Author's abstract. Dr. Philosophy (PhD) in phil. – Karshi, 2018.

6. Кобрицов Б. П. Модели многозначности русской предметной лексики: глобальные и локальные правила разрешения омонимии. Автореф… канд. филол. наук. Москва: РГГУ, 2004.

7. Зеленков Ю. Г., Сегалович И. В., Титов В. А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов (Электрон ресурс). http://www.dialog-21.ru/media/2444/zelenkov_segalovich.pdf.

8. Кобрицов Б. П. Методы снятия семантической неоднозначности. НТИ, Сер. 2, Вып. 3, 2004.

9. Кобрицов Б. П., Ляшевская О. Н., Шеманаева О. Ю. Снятие лексико-семантической омонимии в новостных и газетножурнальных текстах: поверхностные фильтры и статистическая оценка (Электрон ресурс). http://elar.urfu.ru/bitstream/10995/1388/1/IMAT_2005_03.pdf.

10. Hearst M. A. Noun homograph disambiguation using local context in large text corpora // Proceedings of the 7th conference, UW Centre for the New OED & Text Research Using Corpora, 1991.

11. Рысаков С. В. Методы борьбы с омонимией. http://samag.ru/archive/article/3059.

12. http://tech.yandex.ru/mystem – website morphological analysis MyStem.

УДК 81’33

ANALYSIS AND VISUALIZATION OF A SENTIMENT CLASSIFIER BEHAVIOR ON TURKISH WORDS, SENTENCES, AND PARAGRAPHS

Muhammed Enes ALMAHDI, Abdullah AL NAHAS, Cengiz DURNA, Burcu YILMAZ, Yusuf Sinan AKGUL
Gebze Technical University, Gebze, Kocaeli, 41400, Turkey
almahdi@gtu.edu.tr, aalnahas@gtu.edu.tr, cengizdurna@gtu.edu.tr, byilmaz@gtu.edu.tr, akgul@gtu.edu.tr

The widespread usage of social media produces massive user-generated data. Such data conveys users’ opinions on politicians, services, and products. Politicians, business owners, and other interested parties can make use of this data for opinions mining. Moreover, some sentiment analysis models are lexicon-based, which highlights the need for sentiment lexicons. In this paper, we propose a novel automatic sentiment lexicon generation algorithm. Our method is semi-supervised, and at the same time, it distills the classifier’s behavior. Furthermore, we analyze and visualize the sentimental sentences in movie reviews. We apply our method on a Turkish movie review dataset and, as a comparison, on the English IMDB movie review dataset.

Keywords: Natural Language Processing; Sentiment Analysis; Sentiment Lexicon Generation.

АНАЛИЗ И ВИЗУАЛИЗАЦИЯ ПОВЕДЕНИЯ СЕНТИМЕНТ КЛАССИФИКАТОРА НА ОСНОВЕ ТУРЕЦКИХ СЛОВ, ПРЕДЛОЖЕНИЙ И ПАРАГРАФОВ

Мухаммед Энес Альмахди, Абдуллах Аль Нахас, Дженгиз Дурна, Бурджу Йылмаз, Юсуф Синан Акгуль
Технический университет Гебзе, Гебзе, Турция
almahdi@gtu.edu.tr, aalnahas@gtu.edu.tr, cengizdurna@gtu.edu.tr, byilmaz@gtu.edu.tr, akgul@gtu.edu.tr

Широкое использование социальных сетей генерирует массивы пользовательских данных. Такие данные передают мнения пользователей о политиках, услугах и товарах. Политики, бизнесмены и другие заинтересованные стороны могут использовать эти данные для анализа мнений. Более этого, некоторые модели сентимент анализа основаны на лексиконе, что подчеркивает необходимость в сентимент словаре. В данной статье мы предлагаем новый алгоритм автоматической генерации сентимент словаря. Наш метод частично обучаемый, и в то

же время извлекает сущность поведения классификатора. Кроме того, мы анализируем и визуализируем сентимент предложения в обзорах фильмов. Мы аппробируем наш метод на турецкой базе данных обзора фильмов и, для сравнения, на базе данных обзора фильмов английской IMDB.

Ключевые слова: обработка естественного языка; сентимент анализ; создание сентимент словаря.

1. Introduction

The advancements of web 2.0 allow users to generate their content and express their views on politics, product quality, restaurant quality, etc. That encourages politicians, business owners, and other interested parties to do sentiment analysis on this data to advance their services (Tai et al., 2013). There are machine learning-based and lexicon-based sentiment analysis algorithms (Vohra et al., 2013). Lexicon-based sentiment analysis requires extensive sentiment lexicons. Thus, there is a need for generating sentiment lexicons.

Automatic sentiment lexicon generation is the process of building lists of words that are usually used to convey positive or negative sentiments (Feng et al., 2018). Sentiment lexicons play a key-role in lexicon-based sentiment analysis algorithms (Taboada et al., 2011) and in improving machine learning-based sentiment analysis systems (Al-Sallab et al., 2017).

Many sentiment lexicon extraction methods have been discussed in the literature. Some of them arrange words into two groups; positive and negative. In (Mohammad and Turney, 2013), the authors build human-labeled lexicons using crowdsourcing. (Warriner et al., 2013) ask the human annotators to give a value in the range 1-9 for each word on three dimensions. These dimensions are valence (the pleasantness), arousal (the intensity), and dominance of the emotion.

Other researchers use semi-supervised methods for sentiment lexicon induction. In (Turney and Littman, 2003), they start building their lexicon using seed words like good and bad. Then, for the remaining words in the corpus, they measure their closeness and difference from the seed words. Then, they label it accordingly. Another semi-supervised method is sentiment propagation (Hamilton et al., 2016) which use a word graph-based algorithm. Another direction is to use online reviews from movie or product review sites. (Potts, 2011) calculates the normalized likelihood of each word for each rating.

In this work, we present a novel semi-supervised algorithm for au-

tomatic sentiment lexicon generation. We summarize our contributions as follows.

- We introduce a novel method for automatic sentiment lexicon generation. We achieve that by training a sentiment classifier, then analyzing its behavior and extracting the words that affect its decision.
- We show a new method for extracting the positions of the sentences that carry high sentiment values in a movie review. Thus, we can classify the whole review using only the important sentences.

The rest of this paper is organized as follows. In section 2, we detail our method. Then, after demonstrating the effectiveness of our method in section 3, we conclude in section 4.

2. Method

2.1. Extracting Sentimental Words

Our method depends on training a classifier, then we extract sentimental words from a held-out dataset. After splitting our sentiment analysis dataset into a train and a validation sets, we train a classifier on the training split. Then, we extract words with a frequency above a threshold from the validation set. Let's call these words *candidate words*. Next, for each candidate word, we get the validation examples that contain it, let's call them candidate example set. Note that, each word has its own candidate example set. Subsequently, we test the

Algorithm 1: Extracting Sentimental Words

Input: *dataset* is the sentiment analysis dataset
frequency_threshold is the minimum word frequency
score_threshold is the minimum sentiment score
rounds_count is the total number of extraction rounds
Output: *senti_words* is a list of sentimental words with their score.

```
senti_words=[]
train_set, valid_set, test_set = split_dataset(dataset)
for round in rounds_count do
    cls=fastText(train_set) // Train classifier on the training set
    words_freq=extract_words(valid_set) // Extract words and thier frequencies
    for word, freq in words_freq do
        if freq ≥ frequency_threshold then
            word_valid_set=filter_exampels(valid_set, word) // Keep only examples from valid set that contains the
                word
            orig_f1=cls.test(word_valid_set)
            new_word_valid_set=remove_word(word_valid_set, word) // Remove the word from each example
            new_f1=cls.test(new_word_valid_set)
            score=‖new_f1 − orig_f1‖
            if score > score_threshold then
                senti_words.append((word, score))
            end
        end
    end
    // Remove sentimental words from the dataset
    train_set, valid_set, test_set=remove_word(train_set, valid_set, test_set, senti_words)
end
```

Fig. 1. Extracting sentimental words algorithm

classifier on the word's candidate example set and record the classifier's accuracy. Afterward, we remove the word from each example of its candidate example set, test the classifier again, and record the accuracy. If the difference between the first and the second recorded accuracies is above a threshold, we call this word a sentimental word. We repeat this procedure many rounds until we get a poor f1 score, which means that the classifier is not able to classify the sentiment. We summarize our algorithm in Fig. 1. It takes the sentiment analysis dataset, the minimum words frequency, the minimum sentiment score, and the count of extracting rounds as input, and generates the candidate sentimental words as output.

2.2. Sentimental Sentences Positions

A movie review usually consists of more than one sentence. The spread of the sentiment changes over the sentences of the review. While some reviewers express their sentiments in the first sentence, then justify their opinion in the following sentence, other reviewers start with the good points of the movie then proceed with bad points about the movie. Our aim here is to figure out whether there is a pattern or not in conveying the overall sentiment in certain positions of the review. In order to discover that, we analyze the importance of each sentence on the sentiment analysis task. First, we train a classifier on a movie review dataset. Then, we alter each example in a validation set as follows.

- We only keep the example's first sentence,
- We only keep the example's last sentence,
- We keep both the example's first and last sentences,
- We only omit the example's first sentence,
- We only omit the example's last sentence,
- We omit both the example's first and last sentences.

We monitor the predicted sentiment of each of them. We summarize our findings in section 3.

3. Experiments

We perform our analyses on Turkish and English using fastText classifier (Joulin et al., 2016) with 100-dimensional word vectors. In the rest of this section, we define the used datasets and present the results after applying our proposed model for extracting of emotional

words. Finally, we compared detection of the significant positions of sentimental sentences in a paragraph with the statistical method.

3.1. Datasets

For Turkish, we scrape movie reviews from beyazperde.com website, where we consider the positive review when the user's rating is 4 or 5 stars, and negative review when his rating is 1 or 2 stars. We divide it into 50k, 6.25k, and 6.25k examples for training, validation, and testing respectively. Fig. 2-A shows the histogram of the sentences' counts over Beyazperde dataset's examples. This histogram is useful for hyperparameter-tuning in the extraction of the significant sentences' positions.

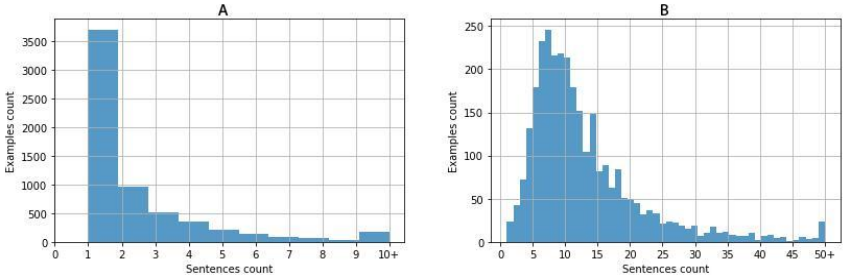

Fig. 2. (A) Histogram of sentences' count across Beyazperde dataset; (B) Histogram of sentences' count across IMDB dataset

We use IMDB sentiment analysis dataset (Maas, 2011) for English. We split it into 19k, 3k, and 3k examples for training, validation, and testing respectively. Fig. 2-B shows the histogram of the sentences' counts over IMDB dataset's examples. Note that the average of sentences' counts in the Turkish dataset is less than the average in the English one.

3.2. Sentimental Words Extraction

After preprocessing, we exclude rare words which occur in less than 100 examples. Then, we apply our method with a sentiment score threshold value of 0.02, and rounds count value of two.

In Table 1 and Table 2, we show a set of the sentimental words extracted from the first and second round of extraction from Beyazperde and IMDB datasets respectively, sorted by its sentiment score.

Table 1. A set of the sentimental words extracted from the first and second round of classification from Beyazperde dataset, sorted by its sentiment score

1st classification round		2nd classification round	
Word	Score	Word	Score
beğenmedim (I do not like)	0.072	gayet (very)	0.0507
gereksiz (unnecessary)	0.0535	olmasına (not to)	0.0476
sıkıcı (boring)	0.0524	hayal (dream)	0.04
altında (under)	0.038	beğendim (I liked it)	0.0384
kötü (bad)	0.0361	sanat (art)	0.0377
mükemmel (excellent)	0.0359	falan (like)	0.0344
harika (great)	0.0357	bişey (something)	0.0342
süper (super)	0.0325	izleyin (watch)	0.0305
başyapıt (masterpiece)	0.0322	mutlaka (surely)	0.0297
basit (simple)	0.0322	sürükleyici (absorbing)	0.0291
dışında (outside)	0.0305	muhteşem (spectacular)	0.0273
değildi (it was not)	0.0294	hele (especially)	0.0263
artık (no longer)	0.0284	kelimeyle (with word)	0.026
berbat (wretched)	0.0275	vakit (time)	0.0258

Table 2. A set of the sentimental words extracted from the first and second round of classification from IMDB dataset, sorted by its sentiment score

1st classification round		2nd classification round	
Word	Score	Word	Score
annoying	0.0406	worst	0.0791
poor	0.0362	sweet	0.0471
looks	0.0355	highly	0.0444
awful	0.0355	moving	0.0388
entertaining	0.0344	effort	0.037
weak	0.0327	waste	0.0368
relationship	0.0298	simple	0.0344
bad	0.0291	girls	0.0288
realistic	0.0283	sets	0.0288
excellent	0.028	comedy	0.0278

1st classification round		2nd classification round	
Word	Score	Word	Score
favourite	0.0275	richard	0.0277
boring	0.0267	terrible	0.0277
horrible	0.0263	hilarious	0.027
basically	0.0258	seriously	0.0265
ridiculous	0.0254	trying	0.0264

3.3. Comparison with a statistical method

We compare the proposed model with the statistical method introduced in (Potts, 2011), where we calculate the sentiment score as the maximum of the probability a word in positive and negative examples divided by the probability of the word in all instances.

{mükemmel, olmayan, gereksiz, muhteşem, mutlaka, süper, eğlenceli, vasat, kötü, sıradan, gayet, berbat, basit, harika, bişey, izleyin, müthiş, hayal, sıkıcı} in Turkish and {awful, favorite, worse, money, boring, highly, looked, worst, wonderful, excellent} in English are examples of the common words between the two methods.

{içinde, hep, beğendim, kelimeyle, müzikler, arada, artık, roman, güzeldi, güzel, ayrıca, hele, hiç, sürükleyici, olmasına, başyapıt, sonuç, vakit, klasik, beğenmedim, altında, kaybı, dışında, biraz, sizi, falan, sanat, değildi} in Turkish and {score, horrible, children, looks, brilliant, ridiculous, group, entertaining, lame, order, seems, annoying, true} in English are examples of the words that show the ability of the proposed method to find some words that have not been extracted by the statistical method. Note that these words are low-frequency words but affect the decision of the classifier.

There are some situations where the classifier does not rely on high-frequency words in only one class (either positive or negative). The following words are examples: {saçma, mi, önemli, yazık, birşey, gereken, para, özellikle, boş, etkileyici, hoş, başarılı, hayatımda, tamamen, başka, yok, oldukça, serinin} in Turkish and {half, stupid, save, amazing, perfect, waste, low, supposed, bad, attempt, minutes, beautiful, acting, nothing, loved, guess, bad, strong, supporting, poor, terrible, heart, person, bad} in English.

Therefore, we think it is better to integrate the two methods to obtain a better emotional words list. However, the use of different classifiers may lead to the extraction of more diverse words' lists.

3.4. Sentimental Sentences Positions Detection

We use the sentence tokenizer of Natural Language Toolkit (NLTK) (Loper and Bird, 2002). We apply the proposed model on the examples which have at least three sentences. Fig. 3 and Fig. 4 show the confusion matrix of the classifier's behavior in different cases. In Fig. 3-A and Fig. 4-A, we conclude that the first and the last sentences include more meaningful information for the classifier decision, while the sentimental expressions are replicated in the rest of paragraphs, which is demonstrated when the first and the last sentences are omitted. See Fig. 3-B and Fig. 4-B.

As a comparison of the concentration areas of feelings in the sentiment analysis text between Turkish and English, we can observe the following.

- In English, the emotional information is reflected in the middle parts of the paragraph compared to Turkish, since we get lower f1 score values in Turkish, when we remove the first and the last sentences. See Fig. 3-B and Fig. 4-B.
- In Turkish, the first and the last sentences, especially the last sentence, carries more emotional information than English, as noted in the differences of f1 scores. In Turkish, the decrease of f1 score between the original text and only keeping the first and last sentences is about %4 only, while the decrease of f1 score in English is about %13. See Fig. 3-A and Fig. 4-A.

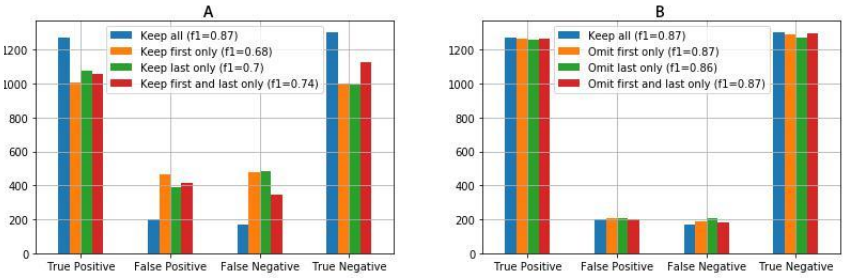

Fig. 3. The confusion matrix and f1 score of the classifier's behavior in different cases on English IMDB dataset; (A) compare the classification performance when we keep the first, the last, and the first and the last sentences only; (B) compare the classification when we remove the first, the last, and the first and the last sentences

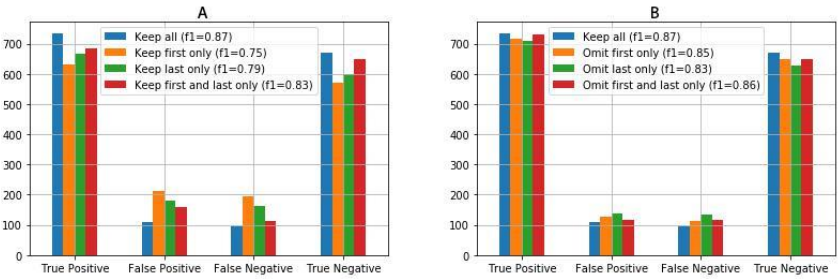

Fig. 4. The confusion matrix and f1 score of the classifier's behavior in different cases on Turkish Beyazperde dataset; (A) compare the classification performance when we keep the first, the last, and the first and the last sentences only; (B) compare the classification when we remove the first, the last, and the first and the last sentences

4. Conclusion

In this work, we introduce a novel semi-supervised method for sentiment lexicon generation. Our procedure unveils the words our classifier attends. Additionally, we analyze and visualize key sentiment sentences in movie reviews. We apply our method to a Turkish and an English movie review datasets. Our algorithm discovers sentiment words that classical methods do not. Our future work will explore building a morphology-aware sentiment lexicon.

REFERENCES

Tai, Y. J., & Kao, H. Y. (2013, December). Automatic domain-specific sentiment lexicon generation with label propagation. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services* (p. 53). ACM.

Vohra, S. M., & Teraiya, J. B. (2013). A comparative study of sentiment analysis techniques. *Journal JIKRCE, 2*(2), 313–317.

Feng, J., Gong, C., Li, X., & Lau, R. Y. (2018). Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews. *Wireless Communications and Mobile Computing*, 2018.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 267–307.

Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., & Badaro, G. (2017). Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16*(4), 25.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 142–150). Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436–465.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and (Warriner et al., 2013) dominance for 13,915 English lemmas. Behavior research methods, 45(4), 1191–1207.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4), 315–346.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016, November). Inducing domain -specific sentiment lexicons from unlabeled corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing (Vol. 2016, p. 595). NIH Public Access.

Potts, C. (2010, August). On the negativity of negation. In Semantics and Linguistic Theory (Vol. 20, pp. 636–659).

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.

УДК 81’33

END-TO-END TRANSLITERATION AND TRANSLATION BETWEEN OTTOMAN TURKISH AND MODERN TURKISH

Abdullah AL NAHAS, Muhammed Enes ALMAHDI, Yusuf Sinan AKGUL
Gebze Technical University, Gebze, Kocaeli, 41400, Turkey
aalnahas@gtu.edu.tr, almahdi@gtu.edu.tr, akgul@gtu.edu.tr

Turkic Languages have utilized various alphabets over the ages. Gokturk alphabet, Uighur alphabet, Arabic alphabet, and Latin alphabet are examples of them. More related to our work, the Turkish language has used an Arabic alphabet during the Ottoman Empire age, then a Latin alphabet in the early years of the Republic of Turkey. Together with the alphabet change, some Arabic and Persian words that were used in the Turkish language were replaced with their Turkish counterparts, and some French words started being used. Because of these changes, the Ottoman Turkish is not understandable for today’s generation. In this work, we build an automatic conversion system between Ottoman Turkish and modern Turkish. Unlike previous work, we develop an end-to-end translation system. Thus, our method does not only perform transliteration but also translates the old words too. We use the sequence-to-sequence framework to achieve that. Our initial results are promising. For the development of Ottoman Turkish to old Turkish parallel corpus, we re-implement a morphology-based transliteration tool in the other way around; from old Turkish to Ottoman Turkish.

Keywords: Natural Language Processing (NLP); Neural Machine Translation (NMT); Ottoman Turkish; Turkish Language.

КОМПЛЕКСНАЯ СИСТЕМА ТРАНСЛИТЕРАЦИИ И ПЕРЕВОДА МЕЖДУ ОСМАНСКИМ И СОВРЕМЕННЫМ ТУРЕЦКИМИ ЯЗЫКАМИ

Абдуллах Аль Нахас, Мухаммед Энес Альмахди, Юсуф Синан Акгуль
Технический университет Гебзе, Гебзе, Турция
aalnahas@gtu.edu.tr, almahdi@gtu.edu.tr, akgul@gtu.edu.tr

Тюркские языки на протяжении веков использовали различные алфавиты. Такими примерами являются: древнетюркский алфавит, уйгурский алфавит, арабский и латинский алфавиты. В нашей работе речь пойдет о турецком языке, который использовал арабский алфавит в эпоху Османской империи, а затем латинский алфавит в первые годы существования Турецкой Республики. Вместе с изменением алфавита,

некоторые арабские и персидские слова, которые использовались в турецком языке, были заменены их турецкими аналогами, также начали использоваться некоторые французские слова. Из-за этих изменений османский язык не понятен современному поколению. В этой работе описана созданная нами автоматическая система преобразования между османским и современным турецким языками. В отличие от предыдущей работы, мы разработали комплексную систему перевода с одного языка на другой. Таким образом, наш метод не только выполняет транслитерацию, но и переводит старые слова. Для достижения данной цели мы используем систему перевода предложение в предложение. Наши первые результаты являются многообещающими. Для развития параллельного корпуса османского языка и старого турецкого языка, мы переоснащаем основанный на морфологии инструмент транслитерации в обратном направлении, от старого турецкого языка к османскому языку.

Ключевые слова: обработка естественного языка; нейронный машинный перевод; османский язык, турецкий язык.

1. Introduction

Political events, accepting a new religion, technological advancement are some factors that affect a language (Lightfoot et al., 2006). The Turkish language is an example of a dramatically evolving language (Lewis et al., 1999). For example, during the Ottoman Empire days, the Turkish language (known as the Ottoman Turkish (Lewis et al., 1999)) used to use a modified version of the Arabic alphabet, imported Arabic and Persian words and phrases, plus some grammatical structures from Arabic and Persian (Ensari et al., 2015). On the other hand, today's Turkish uses a modified version of the Latin alphabet, less imported Arabic and Persian words and has got rid of most exotic grammatical structures (Lewis et al., 1999). These changes in the Turkish language, among others, made the Ottoman Turkish obscure for the period's generation (Lewis et al., 1999). Furthermore, what makes it challenging to learn the Ottoman script is the lack of a one-to-one correspondence between the Ottoman alphabet and the modern Turkish alphabet (Hagopianet al., 1907).

To make the Ottoman legacy accessible, qualified writers have made significant efforts to translate and transliterate some of the Ottoman Turkish documents into modern Turkish. Although that is helpful, manual translation/transliteration is not always an affordable solution. Hence, the need for a handy computerized tool. We dedicate this work to address the need for a fast and inexpensive automated tool. We summarize our contributions as follows.

We develop an Ottoman Turkish - modern Turkish parallel corpus, by implementing a morphology-based transliteration tool in the other way around; from Latin-written Turkish to Ottoman Turkish.

We build an end-to-end neural machine translation model from Ottoman Turkish to modern Turkish.

The rest of this paper is organized as follows. In section 2, we mention related work.

In section 3, we talk about the development of our parallel corpus and the Latin-written Turkish to Ottoman Turkish transliteration algorithm. Then, we describe our method in section 4. After that, we show our experiments and results in section 5. Finally, we conclude with section 6.

2. Related Work

We group related work that into two categories; (a) translating and transliterating between Ottoman Turkish and modern Turkish, (b) transliterating between two different scripts of the same language.

2.1. Translating and Transliterating from Ottoman Turkish to Modern Turkish

Scientific research in this area is somewhat limited (Kurt et al., 2011) (Erg et al., 2017) (Korkut et al., 2019), and mainly depends on morphological analysis of the input Ottoman text.

In (Kurt et al., 2011) researchers describe a six-step Ottoman Turkish to modern Turkish transliteration system. Their proposed framework starts by morphological analysis of each word. The first step results in producing possible stem/suffix pairs. Second, their framework looks up each stem/suffix from a dictionary. Third, it synthesizes each word in with the Latin script. Fourth, it proceeds with word disambiguation, where it chooses the most probable words using an n-gram language model. Fifth, it deals with errors caused by previous steps. Those errors could be typographical, wrong segmentation of Ottoman words and unknown words. Finally, it detects noun adjuncts.

In (Erg et al., 2017), researchers build a word-based analysis tool of Ottoman Turkish that works as follows. First, it takes an Ottoman word as input and checks if it exists in an exception list. If yes, it immediately outputs the Latin mapping of it. If not, it morphologically analyses the word and generates possible stem/suffix pairs. Finally, it maps each Ottoman stem/suffix pair to its Latin counterpart and outputs the most common one of them.

In (Korkut et al., 2019), researchers do very similar work to (Kurt et al., 2011) but without ambiguity handling. When there are multiple outputs of a word, they output them all and tell the user to choose between them.

2.2. Transliterating between two different scripts of the same language

Some languages, other than Turkish, use more than one writing system. For example, some Arabic speakers romanize their written communications in social media and messaging applications and mix them with some English words (Darwish, 2014) (Al-Badrashiny et al., 2015). The romanized version of the Arabic they use is nonstandard and is called Arabizi. The motivation to transliterate from Arabizi to Arabic is to be able to use the already existing Arabic NLP tools. In (Darwish, 2014), researchers build the Arabizi-Arabic transliteration system as following. First, they use Moses (Koehn et al., 2007) to build a statistical phrase-based machine translation model. They treated words as sentences and letters as words. After generating candidate translations of a word, they check if it exists in a large monolingual corpus. Having multiple possible word-to-word translations, they pick the most suitable translation using a trigram language model.

In (Al-Badrashiny et al., 2015), researchers first train a character-level finite state transducer that generates all possible transliterations of an input Arabizi word. Then, they filter the generated list using a morphological analyzer. Finally, they choose the most appropriate words using a language model.

In (Le et al., 2019), researchers design a neural machine transliteration system between the English alphabet and the Vietnamese alphabet. Their purpose is to transliterate English named entities into the Vietnamese alphabet. Their work has two stages. First, they prepare a dataset by segmenting words, then looking them up from a pronunciation dictionary and aligning them on character level, so they have parallel data. Then, they train an RNN-based machine transliteration model using the collected data.

3. Building Parallel Corpus

In this section, we describe the development of the dataset we use to train our translation model. We use a book titled Nutuk, for which we

find an Ottoman Turkish, old Turkish and modern Turkish versions. It was originally written with Ottoman Turkish. Then, after the adoption of the Latin-based Turkish alphabet, it was re-written with it. Moreover, because the change of the Turkish language over the years, old Turkish became cryptic. Consequently, qualified writers who are good at both old and modern Turkish re-wrote Nutuk again in nowadays Turkish. We scan and OCR the modern version of Nutuk. To build the parallel corpus we need, we should also OCR the Ottoman version of Nutuk. Unfortunately, however, we could not find an accurate OCR tool for Ottoman Turkish documents. Therefore, another work around is needed. To overcome this, we re-implement a Latin-written Turkish to Ottoman Turkish morphology-based transliteration tool. Then, we use the mentioned transliteration tool to transliterate the old Turkish version of Nutuk to Ottoman Turkish. As a result, we the required parallel corpus.

```
1: procedure OTTOMANIZE(inputSentence, wordDictionary, suffixDictionary)
2:     sentenceWords ← splitIntoWords(inputSentence)
3:     for word in sentenceWords do
4:         wordIsFound ←FALSE
5:         suffix ←""
6:         while NOT wordIsFound or word ≠ "" do
7:             suffix ← suffix + word.lastChar()
8:             word ← word.removeLastChar()
9:             if word IN dictionary then
10:                wordIsFound ← TRUE
11:                currentWordInOttoman ← wordDictionary.lookup(word)+
   suffixDictionary.lookup(suffix)
12:            end if
13:            if wordIsFound then
14:                ottomanSentence ← ottomanSentence +
   currentWordInOttoman
15:            else
16:                return ""          ▷ Cannot transliterate this sentence.
17:            end if
18:        end while
19:    end for
20:    return ottomanSentence
21: end procedure
```

Fig. 1. A Morphology-based Old/Modern Turkish to Ottoman Turkish Transliteration Algorithm

3.1. Nutuk

Nutuk, meaning the speech, is a book compilation of the speeches delivered by the founder of The Republic of Turkey, Mustafa Kemal Atatürk, between the 15th and 20th of October 1927. It talked about the political and the marital situation of the last days of the Ottoman Empire, the Turkish War of Independence and the foundation of the Republic of Turkey. It was delivered at the second congress of Cumhuriyet Halk Partisi; the political party founded by Mustafa Kemal Atatürk.

3.2. A Transliteration Algorithm from Turkish with Latin characters to Ottoman Turkish

The transliteration from the Ottoman Turkish alphabet to the modern Turkish alphabet is complex. Especially if we try to do it using only morphological analysis. However, the other way round (i.e., from the modern Turkish alphabet to Ottoman alphabet) is more straightforward. Consequently, we re-implement a modern Turkish-Ottoman Turkish transliteration tool (Şik, 2015). Then we use it to build a parallel corpus to train an NMT model from Ottoman to modern Turkish. We show the old/modern Turkish – Ottoman Turkish transliteration algorithm in Fig. 1. We give statistics of our parallel corpus in Table 1.

Table 1. Some statistics on our old Turkish – modern Turkish parallel corpus

Split	No. Tokens	No. Sentences
Train (Ottoman/Modern)	364,493/353,984	29,670
Validation (Ottoman/Modern)	3,767/3,674	305
Test (Ottoman/Modern)	3,753/3,638	305

4. Method

For the purpose of translation, we use neural machine translation (NMT). The specific model we use is the RNN Encoder-Decoder . We breifly describe it as follows.

Given an input sequence = (1, 2,..,), each input is mapped to a vector using a word vector embedding lookup table, then the RNN encoder maps Figure

4.1 An illustration of sequence-to-sequence models. =(1,..,) to another sequence ′ , and because ′ is a function of 1′ ,.., −1′ , , it could be thought of as a summary of the whole input sequence . Let us call ′ the context vector. After that, the RNN decoder gets initialized with the context vector and is given a go token to trigger it to start generating the output sequence, one at each time step. Please refer to the original paper for more details.

5. Experiments and Results

The baseline we use for our experiments has two LSTM layers for both the encoder and the decoder with size of 250 and a word embeddings size of 500. We train the baseline for 40 epochs and get validation set BLEU score of 28 points. As an NMT implementation, we use Open-NMT-py (Klein et al., 2017). We initialize our weights using a uniform random distribution in the range (-0.1, 0.1). As optimization algorithm, we use Adam (Kingma et al., 2014), an initial learning rate of 0.001, and Noam learning rate schedule (Vaswani, et al., 2017). To prevent overfitting, we use dropout (Srivastava, et al., 2014) in the LSTM stacks with a probability of 0.65. Our batch size is 32.

We do parameter search to determine a better hyper-parameter configurations.

According to our experiments, we decide to use the following configurations.

We train our models for 48 training epochs, 500 dimensional word vectors, 500 dimensional LSTM cells,

Dot product global attention,

A bidirectional LSTM encoder,

We tie (share) the decoder's input and output embeddings, and

We use pre-trained word vectors for both the encoder's and the decoder's look-up tables, and fine-tune them while training the NMT model.

Our BLEU score after applying the aforementioned hyper-parameters is 41.84 which is 13.84 points improvement. We show some of the model outputs in the following lines.

. زب ، وب پرنسیپی ، اصلا مرجی اوالمازدق , Original: Human: biz bu prensibi , hiçbir zaman uygun bulamazdik . Model Output: biz , bu cloruldar , hiç yetencği biriyim .

. وب تلغرافڭ بر تورلی معنا سینی الڭلمیوردم Original: Human: bu telgrafin anlamini bir türlü kavriyamiyordum . Model Output: bu telgrafin bir türlü anlamini gerektiriyordu .

. دینجی مکتوب استادم ؛ علی کمال بـگ ، دون وا ذاتله وگورشمش Original: Human: yedinci mektup ustadim ; ali kemal bey dün o ki iyle görü mü . Model Output: yedinci mektup ustadim ; ali kemal bey , dün o ki iyle görü mü .

، ردد قوتی ، صالحیت ، شکل ، وضعیت ، بوگونکی بوجوم وا اولونجه ، ه بولونجه تصرف فرصتی اجرایی فعالیتی Original: ، تشکیلاتی سیاسی ی ه و اساسی ی ه نڭ من اولدیغی معین ردم ، تمام ی خالفت و سلطنت
Human: o var olunca bugünkü durum , biçim , yetki geçicidir , halife ve padi ah çali ma firsati bulunca , yasal ve anayasal kurulu un ne olacağı bellidir bilinmektedir.
Model Output: o öğrenilince bugünkü durum , biçim , yetki geçicidir , halifelik ve padi ahlik , i çali maya firsat bulunca , siyasi siyasi ve anayasanin ne olduğu belirlidir , bilinir .

، ه ردده کدمکرتسوگ زڭ اتانیاب و ه طون نان ولوا ح اض وغ لیبت
دولی ایتالفی ه نارسینی ه ای الڭ زڭ Original: اوالرق انطولی
ردده کدمم ایتبول قبی ل ایتسی ردنی صلرخم

Human: verilen nota ve sizin sözleriniz açikça göstermektedir ki , itilaf devletleri londra konferansina anadolu delegelerini tek ba larina kabul etmemektedir . Model Output: bildirilen nota ve ve açikça açikça göstermektedir ki , itilaf devletleri londra konferansina yalniz olarak anadolu adamlari kabul edilebilir .

6. Conclusion

As the Turkish language has changed largely since a hundred years, the Ottoman Turkish documents became cryptic to nowadays' generation. In order to make the Ottoman written heritage relatable, qualified writers who are good at both the modern Turkish and the

Ottoman Turkish are manually rewriting old documents into modern Turkish. That is a vital but a resource-expensive workaround. To work out this problem, we develop an Ottoman Turkish to modern Turkish end-to-end neural machine translation model. To develop the parallel corpus, we build a morphology-based Latin-written Turkish to Ottoman Turkish transliteration tool. Our base-line Ottoman Turkish to modern Turkish translation model has a BLEU score of 28 points. Our improved model has a BLEU score of 41.84. Our Future work will explore the following directions: (a) build an OCR model for Ottoman Turkish documents, (b) ensemble NMT models with a morphology-based system for the translation and transliteration of Ottoman Turkish texts.

REFERENCES

Lightfoot, D. (2006). How new languages emerge. Cambridge University Press.

Lewis, G. (1999). The Turkish Language Reform: A Catastrophic Success: A Catastrophic Success. OUP Oxford.

Ensari, M. A. (2015). Osmanlica İmla Müfredati. Istanbul: Hayrat Vakfi Yayinlari.

Hagopian, V. H. (1907). Ottoman-Turkish conversation-grammar: a practical method of learning the Ottoman-Turkish language. J. Groos.

Kurt, A., Bilgin, E. F. (2011). The Outline of an Ottoman-to-Turkish Automatic Machine Transliteration System," Istanbul: in First Workshop on Language Resources and Technologies for Turkic Languages. ERGİŞİ, A., Şahin, E. (2017). Dervaze: A Spelling Dictionary for Digital Translation," International Journal of Languages' Education and Teaching, vol. 5, no. 3, pp. 78–84.

Korkut, J. Morphology and lexicon-based machine translation of Ottoman Turkish to Modern Turkish. Darwish, K. (2013). Arabizi detection and conversion to Arabic. arXiv preprint arXiv:1306.6755. Al-Badrashiny, M., Eskander, R., Habash, N., & Rambow, O. (2014, June). Automatic transliteration of romanized dialectal Arabic. In Proceedings of the eighteenth conference on computational natural language learning (pp. 30–38).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions (pp. 177–180).

Le, N. T., Sadat, F., Menard, L., & Dinh, D. (2019). Low-Resource Machine Transliteration Using Recurrent Neural Networks. ACM

Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(2), 13.

Şik, Z. (2015). Osmanlica Çeviri. [Online]. Available: https://play.google.com/store/apps/details?id=com.tahirhoca.osmanlicaceviri&hl=tr. [Accessed 16 06 2019].

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929–1958.

УДК 81’33

FREE/OPEN-SOURCE TECHNOLOGIES FOR TURKIC LANGUAGES DEVELOPED IN THE APERTIUM PROJECT

Jonathan Washington,[a] Ilnar Salimzianov,[b] Francis M. Tyers,[c,d] Memduh Gokirmak,[e] Sardana Ivanova,[f] Oguzhan Kuyrukcu[g]

[a]Linguistics Department, Swarthmore College, Swarthmore, USA;
[b]Independent Scholar, Kazan, Russia;
[c]Department of Linguistics, Indiana University, Bloomington, USA;
[d] Школа лингвистики, Высшая Школа Экономики, Москва;
[e]Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic;
[f]Department of Computer Science, University of Helsinki, Helsinki, Finland;
[g]Department of Linguistics, Boğaziçi University, İstanbul
jonathan.washington@swarthmore.edu

Free/open-source language technology for Turkic languages has been under development under the Apertium open-source project for over 8 years. Morphological transducers for a large number of Turkic languages are production-ready, and machine translation systems between a number of the languages have been released. Additionally, several prototype MT systems between Turkic and non-Turkic languages have been developed. This paper describes the current capabilities of these technologies, demonstrates the type of work that goes into them, illustrates applications for their use in empowering marginalised linguistic communities and assisting with language revitalisation efforts, and outlines future work.

Keywords: Turkic languages; language technology; finite state transducers; morphological analysis; spell checking; machine translation; linguistic rights.

БЕСПЛАТНЫЕ/ОТКРЫТЫЕ ТЕХНОЛОГИИ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ, РАЗРАБОТАННЫЕ В ПРОЕКТЕ APERTIUM

Джонатан Вашингтон[a], Илнар Салимзианов[b], Фрэнсиз М. Таэрс[c,d], Мемдух Гёкырмак[e], Сардана Иванова[f], Огузхан Куйрукчу[g]

[a]Факультет языкознания, колледж Свартмор, Свартмор, США
[b]Казань, Россия
[c]Факультет языкознания, Индианский университет, Блумингтон, США

[d]*Школа лингвистики, Высшая Школа Экономики, Москва, Россия*
[e]*Карлов университет, Прага, Чехия*
[f]*Хельсинский университет, Хельсинки, Финляндия*
[g]*Босфорский университет, Стамбул, Турция*
jonathan.washington@swarthmore.edu

Технология бесплатных/открытых исходных текстов для тюркских языков разрабатывается в рамках проекта открытый источник Apertium более 8 лет. Морфологические преобразователи для большого числа тюркских языков готовы к производству и выпущены системы машинного перевода между несколькими языками. Кроме того, разработано несколько прототипных систем МТ между тюркскими и нетюркскими языками. В данной статье описываются текущие возможности данных технологий, демонстрируется их работа, иллюстрируются приложения для использования в соответствующих ограниченных лингвистических сообществах, а также, описывается дальнейшее развитие проекта.

Ключевые слова: тюркские языки; языковая технология; автоматы конечных состояний; морфологический анализ; машинный перевод с проверкой орфографии; языковые права.

1. Introduction

This paper introduces free/open-source (FOS) language technology for Turkic languages developed as part of the Apertium open-source project[1], and overviews their current states. This includes morphological transduc-ers at all levels of performance, and machine translation (MT) systems between Turkic languages and between Turkic and non-Turkic languages–also at various levels of performance. One goal of this paper is to overview the existing tools and their states at the time of writing, as well as to demonstrate how they have been designed.

Another goal of this paper is to argue that this sort of work can empower marginalised linguistic com-munities and assist with language revitalisation efforts. Specifically, we argue that symbolic approaches to language technology and releasing the technology as free/open-source software (FOSS) are both crucial to these ends.

Section 2 motivates the development of FOS symbolic language technology for Turkic languages; sec-tion 3 overviews currently existing morphological transducers and their creation; section 4 overviews currently existing machine translation systems and their creation; section 5 discusses directions for future work; and section 6 concludes.

[1] http://apertium.org; all code at http://github.com/apertium; Turkic resources may be tested at http://turkic.apertium.org.

2. Background

This section motivates the paper. Subsection 2.1 motivates the need for improved linguistic rights across the Turkic-speaking world; subsection 2.2 explains how language technology can serve as a means to this end; subsection 2.3 expounds on symbolic approaches as an effective approach to this kind of language technology; and subsection 2.4 drives the need for releasing this technology to the public under Free and Open licenses.

2.1. Linguistic rights in Turkic-speaking communities

There are around 40 Turkic languages[1], spoken in communities across Eurasia, and in a number of diaspora communities through the world.

While Turkic languages were once at the heart of civilisation[2] and scientific advancement across the Eurasian continent (Starr 2009), Turkic languages have since lost prestige in the areas where they were once spoken, and in an increasing number of cases have entirely ceased to be spoken by communities that once spoke them.

Today only 6 Turkic languages enjoy offcial status at a national level (Azerbaycani, Kazakh, Kyr-gyz, Turkish, Turkmen, and Uzbek), and another approximately 13 have offcial recognition as minority lan-guages or are offcial at an immediately sub-national level (Altay, Bashqort, Chuvash, Crimean Tatar, Gagauz, Karachay-Balkar, Khakas, Kumyk, Noghay, Sakha, Tatar, Tuvan, and Uyghur). Hence, most Turkic languages do not enjoy a status that affords speakers access to resources in their language; in fact, speakers and learn-ers of many of the latter set of languages may have trouble finding support for use of the language. Instead, speakers of most Turkic languages access resources in the prestige languages of the areas they live in, including Russian, Chinese, Persian, and even English. While multilingualism is not detrimental to these communities, the lack of resources in or encouragement to use the Turkic language marginalises it. Often even primary education in a given Turkic language is not available within the community that speaks it, and those growing up with the languages

[1] Counts of Turkic languages vary, sometimes by quite a bit, mostly depending on which varieties are considered separate lan-guages or not. In particular, some sources offer a number higher than 40.

[2] We use the term «civilisation» with some reservation. We have in mind a meaning along the lines of cultural, economic, and military strength and influence, which we fully understand does not equate to «beneficial to all.»

come to view them as unnecessary relics–even as holding them back socially–and so these speakers either passively fail to pass the language on to the next generation, or actively avoid doing so (cf., the case of Chuvash, per Alòs i Font 2014).

The development of open textual resources in these languages, such as Wikipedia, can mitigate to some extent the lack of educational and textual resources in a language and, more importantly, can signal to speakers that the language is not a relic and that is has a place in the digital age. However, these resources are maintained by volunteer native speakers, often by dedicating huge amounts of uncompensated time to the project.

All of this represents a situation where the linguistic rights of Turkic language speakers are in jeopardy in many of the communities where the languages are used. Linguistic rights include not just the right of interlocutors to use the language of their choosing with one another, but also the right to not be discriminated against based on language use, the right to access the judicial system regardless of linguistic knowledge, and the right to education in ones own language–to name a few that are outlined in the Universal Declaration of Human Rights (United Nations 1948).

One of the main goals of the work described in this paper is to enhance the situation with regard to the linguistic rights of Turkic language speakers. In more practical terms, several of the uses for the tools we describe have the potential to offset these limitations, for example by assisting with learning of Turkic languages, by expanding access for Turkic-language speakers to resources in languages other than their native language(s) (through machine translation from better resourced languages, including other Turkic languages), and by making it possible to more easily and quickly expand the range of materials available in a given Turkic language.

2.2 Motivating Turkic language technology

Bird (2009, p. 473) notes that «We live during a brief period of overlap between the mass extinction of the world's languages and the advent of the digital age,» and proposes that computational linguists «[focus] some of our efforts on a new kind of computational linguistics, one that accelerates the documentation and description of the world's endangered linguistic heritage, and delivers tangible and intangible value to future generations.» The work described in this paper, then, may be understood as a response to this call-to-arms.

While many Turkic languages are not in immediate danger of not being passed on to future generations, a number are. Of the 39 non-extinct Turkic languages currently recognised as distinct by Ethnologue (Eberhard et al. 2019), 13–exactly one-third–are at level 6b «threatened» or worse. We have not developed resources for any of these languages yet, but most of the remaining Turkic languages–including most of those for which we have developed resources–face challenges nonetheless.

The diffculty of this situation is explored in more depth by Kornai (2013), who suggests that while two-thirds of the languages on the planet are not considered endangered, only 5% of all languages have a suffcient level of access to language technology to survive in the digital age. The proportions of world lan-guages discussed by Kornai appear to scale well to Turkic languages: as mentioned above, two-thirds of Turkic languages are not considered endangered, yet nearly all Turkic languages lack the most basic of language tech-nologies (e.g., spell checking). For Kornai (ibid.), access to language technology is critical for the continued survival of any currently used language. Barriers to using a language online lead to the first of three stages that reflect imminent language loss: loss of function (when a language is no longer useful in a range of domains), loss of prestige (when attitudes towards the language worsen), then loss of competence (when a generation of «semi-speakers» appears). So the loss of function associated with barriers to using a language in digital communication may initiate the interruption of intergenerational transmission of a language.

A recent survey of language use among 525 Kazakhstani business specialists (Aimoldina 2019) reports, despite positive attitudes towards Kazakh, that Kazakh is the least used of Kazakh, Russian, and English in domains related to digital technology. Respondents were asked to choose which language of the three they most prefer to conduct different activities in. For several categories, no respondent chose English, e.g., *in the shop*, *on transport*, *at the bank*. For some categories, respondents preferred Kazakh to Russian or English, including *listening to the radio* (50.8% Kazakh, 43.6% Russian, and 5.6% English) and *in state bodies* (51.25% Kazakh, 48.75% Russian, and 0% English). The lowest response rates for preference of using Kazakh were *working on a computer* (9.18% Kazakh, 56.62% Russian, and 34.2% English) and *searching the internet* (11.2% Kazakh, 50.65% Russian, and 38.15% English), followed distantly by *reading non-fiction* (18.1% Kazakh,

68.1% Russian, and 13.8% English) and *reading fiction* (25.6% Kazakh, 58.9% Russian, and 15.5% English). This is in spite of the fact that there is more financial support for resources in Kazakh–including state and corporate support for the development of language technology–than for most other Turkic languages. Our experience anecdotally suggests that the trend to prefer using computers and the internet in better-resourced languages over a Turkic language is a pervasive problem.

The take-away of all this is that even languages with millions of speakers which are not categorised as being in any immediate danger of losing intergenerational transmission may still be considered in danger when the opportunities for speakers to use the language in digital media are limited. Combined with other issues of access to materials and services in a language as discussed in section 2.1, the situation exhibits a need for intervention.

This paper outlines the development of two main types of language technology which have the potential to feed into other types of language technology. Specifically, we have developed morphological transducers and machine translation systems.

On their own, machine translation systems have the ability to make existing materials in one language available to speakers of another in several ways. Machine translation may be used by an individual to make sense of text in an unknown language. It may also be used by a translator to accelerate and reduce the diffculty of translating materials. MT may also serve as a component of other systems, including language learning tools. These are important tools for language revitalisation, as they can facilitate improvement of language skills by those interested in a language at all levels of knowledge.

Morphological transducers serve not only as components of our machine translation systems, but can be employed in language learning tools and as spell checkers (a technology highlighted by Kornai 2013), not to mention as components in other tools. The importance of spell checking to language survival originates in the symbolic nature of the red underline, which communicates to users that a language it does not support is always «wrong», and is not meant to be written (at least not on a computer).

2.3. Motivating symbolic approaches

The tools presented in this paper are all symbolic, meaning that they make use of explicit linguistic knowl-edge instead of distributional properties of corpora. While they may be augmented with statistical

or related approaches, they do not rely on the availability of large amounts of data.

Indeed, one of the main advantages of symbolic approaches to language technology is that tools may be bootstrapped with no need for large corpora. Another major advantage is that the source of errors returned by symbolic tools may be easily identified and fixed.

Developers of symbolic tools spend most of their time formalising linguistic knowledge in computational representations. Like statistical and related approaches, this does require specialist knowledge– however, the type of specialist knowledge needed is knowledge of a language, which speakers of languages who would like to see more technology in their language often possess. The additional knowledge of the computational formalisms for encoding that knowledge (and conventions and reasoning from the field of Linguistics) may be learned more easily than other natural language processing (NLP) methodologies, as no programming or math is needed. Furthermore, the work of implementing linguistic knowledge in computational formalisms may be seen as less tedious and time-consuming than creating or translating immense amounts of text to use as the training data needed for corpus-based approaches.

This paper provides examples of all of the formalisms used to encode linguistic knowledge that are needed to develop the tools described.

2.4 Motivating FOS

Releasing language technology as free/open-source software is critical to its ability to benefit a community. Releasing software under a free/open-source license means that similar and related software does not have to be reimplemented multiple times. Such licenses allow others to freely use, improve, and redistribute software. These freedoms in turn support other freedoms, such as freedom of speech and individual privacy (FSFE 2019).

Wheeler (n.d.) conducted an extensive, data-supported comparison of free/open-source software and the alternative, proprietary software. He finds that FOSS is a «reasonable or even superior approach» to using software as compared to proprietary software based on popularity, reliability, performance, scalability, se-curity, and total cost of ownership. Developing FOSS may also offer a reasonable profit model for business, although certainly not always.

In terms of language technology specifically, Streiter et al. (2006) present a series of arguments for why work on technology for what they term «noncentral» languages[1] should always be released as FOSS. The main points they raise are that open licenses encourage sharing of data, formats, and programs; ensure continuity between projects which do not overlap in time; prevent conflicting standards in data formats and encoding; promote a collaborative atmosphere between experts in various disciplines, including members of the language communities; allow for maximal use of limited resources (data, labour, computing power, and finances); and help to overcome the isolationism of working on noncentral languages. Each of these topics (and more) is addressed in depth by Streiter et al. (ibid.), and the reader is strongly encouraged to refer to that work for deeper discussion of these issues.

Pedersen (2008) presents several more points in favour of Open Source when working on language technology in general–not just for languages with fewer existing resources. By releasing code early and releasing often, a project published in the academic press ends up with a code base that contains better software and better documentation, and that can be used more easily to reproduce results by both other users and the original authors. Additionally, while potentially counter-intuitive, code released publicly under an Open license faces less chance of being «stolen» or «scooped», while at the same time making it easier for others to build on and cite the work. Pedersen (ibid.) reiterates the point by Streiter et al. (2006) that releasing software openly enhances the possibility of project survivability, and also points out that Open licenses offer a way for academics to give back to the greater good.

One of the major advantages of releasing code publicly under FOS licenses that we would like to em-phasise is how it can facilitate continued community involvement and use of software. For example, the Apertium-developed Kazakh transducer has been developed in collaboration with Kazakh speakers, and vari-ous research projects,

[1] This term refers to languages which lack some portion of resources which put «the blessings of digital culture [within] reach», such as the following: a writing system, Unicode support, font support, spell checkers, information retrieval systems, corpora, stemmers, taggers, parsers, and MT systems. Most Turkic languages fall into this category to some extent or other, though many are becoming more «central», in part through the work represented in the current paper, and for some, also through corporately developed resources. The depth and breadth of available resources are still far from ideal for all Turkic languages.

both in Kazakh-speaking communities and elsewhere, have used, built on, and cited the transducer in other work.

3. Morphological transducers

This section overviews in general terms what morphological transducers are and how they work (section 3.1), presents their main uses (section 3.2), describes their implementation using HFST (section 3.3), and describes the status of the transducers already developed for Turkic languages (section 3.4).

3.1. Overview

A morphological transducer maps between linguistic forms and analyses in both directions–i.e., it maps both form to analysis and analysis to form.

An example of this is the mapping between the form *алмалардан* 'from the apples'[1] and the analysis `алма<n><pl><abl>`, the latter of which can be read to mean that the lemma is алма, the category is noun (`n`), and it is a plural (`pl`) form and is in the ablative (`abl`) case. Morphological analysis is the conversion of an orthographic form into an analysis, and morphological generation is the conversion of an analysis into an orthographic form, both of which a transducer can be used for.

A simple morphological transducer corresponding to the above example, but containing three additional forms of the same stem–the ablative singular *алмадан*, the nominative singular *алма*, and the nominative plural *алмалар*–is shown in Figure 1.

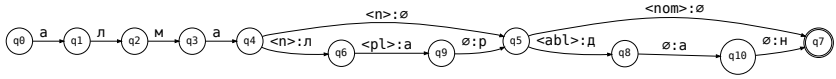

Figure 1. A simple transducer containing four form-to-analysis mappings: *алма* ↔ `алма<n><nom>`, *алмалар* ↔ `алма<n><pl><nom>`, *алмадан* ↔ `алма<n><abl>`, and *алмалардан* ↔ `алма<n><pl><abl>`. States are represented with circles, the final state with a double circle, and directional paths between the states are represented with arcs terminated by arrows. The : symbol separates the values of the two tapes; path labels without a : have the same value on both sides of the tape.

[1] This form and gloss are correct in a number of Turkic languages which use Cyrillic, including Kazakh, Qaraqalpaq, Kyrgyz, Tatar, Crimean Tatar, and Turkmen.

The transducer, also called a finite-state transducer (FST), is a finite-state automaton (FSA) consisting of two tapes, one for the orthographic forms and one for the analysis. When taking forms or analyses as input, the transducer processing program matches the input, character-by-character, to symbols on the appropriate tape (as defined by the user); when a match is found for a given character, both tapes advance along the arc for that character to the next state and the value of the second tape for that arc is read and retained in an output buffer. Only when a final state is reached and no more input remains can the contents of the output buffer for the set of arcs that led to the final state be output. The «reversible» nature of the transducer is due to the fact that either tape may be used for input or output.

Other properties of transducers that result from their implementation are that they process text very quickly, and that they can store many forms very effciently. Note that in the above example, four forms consisting of 28 total characters with analyses consisting of 22 total symbols are stored in an 11-state transducer with 12 arcs.

3.2. Main uses

Morphological transducers may be used in machine translation (MT) systems. Morphological analysis of the input language text, followed by lexical and structural transfer of the resulting analysis, and then morphological generation of the resulting analysis is one way to translate text from one language to another. Section 4 explains in more depth how such rule-based MT systems work and the role of morphological transducers in these systems.

Additionally, there is a role for morphological transducers in other types of systems, i.e., non-rule-based systems. For example, Toral et al. (2019) used the Apertium-developed Kazakh transducer as a segmenter, and integrated it into a neural MT system, reporting improvements over a baseline MT system. Silfverberg and Tyers (2019) use FSTs to generate data for neural morphological analysers for Uralic languages that come close to the performance of symbolic (FST) analysers.

The 2019 VarDial shared task on cross-lingual morphological analysis (Zampieri et al. 2019) contained a task on cross-lingual morphological analysis for the Turkic languages. Teams were given data generated from Wikipedia as well as Apertium transducers for six Turkic languages (Bashqort, Crimean Tatar, Kazakh, Kyrgyz, Tatar, and Turkish) and were asked to create a model to analyse a surprise

language, Karachay-Balkar. Three teams took part –from Russia, Mexico, and Norway/Germany–, and while none of the models came close in terms of *F* -score to a hand-developed analyser (the best system achieving a score of 39.46), they presented the first ever results for cross-lingual morphological analysis for Turkic languages.

Various sorts of linguistic research or other automated tasks may also leverage morphological trans-ducers. For example, searches in corpora for particular lemmas or parts of speech can be automated. Also, morphological analysers can be used to e.g. bootstrap syntactically annotated corpora, as was done for the Universal Dependencies Kazakh Treebank (Tyers and Washington 2015).

There are several uses for morphological transducers besides MT systems and linguistic research. One use that empowers linguistic communities is as a spell checker. Since a morphological transducer is imple-mented as a two-tape FSA, it is trivial to turn it into a one-tape FSA, or an acceptor. An acceptor version of a morphological FST simply accepts valid forms in the transducer and rejects forms that are not in the transducer. By defining weights to common misspellings (e.g., common character substitutions), spelling suggestions are also possible.

One use for morphological transducers that can help with language revitalisation is use in language-learning software. For example, Johnson et al. (2013) present work on creating comprehension dictionaries (which allow language learners to look up words in a dictionary even when presented with a form that is not in the dictionary) for morphologically-complex languages, such as the Sámi languages, using finite-state transducers. A related technology, Revita[1] (Katinskaia et al. 2018), uses Apertium transducers for several Turkic languages (Ivanova et al. 2019): currently Sakha and Kazakh are publicly available, and Tatar is in testing. The site presents the user with texts, categorised at different levels of diffculty. Word forms that are known by the transducer may be clicked on; the word's stems as given by the transducer are then looked up in a third-party bilingual dictionary. This allows a language learner to read a text without spending a lot of time looking up unknown words in the dictionary, and the use of the transducer means that lots of time is saved by the preparers of a text in a morphologically rich language, and that time is not wasted looking up incorrectly segmented versions of the word. There is also a mode where the language learner is presented with a number of stems from the text and asked to put them in the right form in the context of the text.

[1] https://revita.cs.helsinki.fi/

Another use for a morphological transducer in support of language learning is as a morphological gen-erator in a paradigm generation program. A prototype language-agnostic paradigm generation program that queries Apertium transducers through Apertium's web API (Cherivirala et al. 2018) was developed recently by a participant in Google Code-In[1].

```
Multichar_Symbols

%<n%>     ! Noun
%<pl%>    ! Plural
%<abl%>   ! Ablative
%<nom%>   ! Nominative
%{D%}     ! Archiphoneme
%{L%}     ! Archiphoneme
%{A%}     ! Archiphoneme
%>        ! Morpheme boundary

LEXICON Root

Nouns ;

LEXICON N-INFL

%<n%>%<pl%>:%>%{L%}%{A%}p N-CASES ;
%<n%>: N-CASES ;

LEXICON N-CASES

%<abl%>:%>%{D%}%{A%}н # ;
%<nom%>: # ;

LEXICON Nouns

алма:алма N-INFL ;  ! "apple"
```

Figure 2: A complete `lexc` file that implements four forms of the word *алма*: nominative singular, nominative plural, ablative singular, and ablative plural. The : symbol separates the two tapes, % is an escape character (allowing special characters to be used on the tapes), ! is a comment character (all text after it on a line is ignored, # is an end state, and ; ends content lines. Content lines define content on both tapes and then, after a space, direct to another lexicon or an end state.

3.3. Morphological transducers with `lexc` and `twol`

Developed transducers for Turkic languages are implemented in HFST (Helsinki Finite State Technology, Linden et al. 2011)–specifically

[1] The code is available at https://github.com/avyayv/apertium-paradigm-generator.

in the `lexc` and `twol` formalisms, using the general approach described by Koskenniemi (1983) and Beesley and Karttunen (2003).

The `lexc` formalism is used for the mapping of an analysis (lemma + tags) onto a morphological form. The morphological form is a phonologically abstract form which allows a given suffx to have a minimal number of representations when there may otherwise be a large number of phonologically conditioned surface forms.

For example, the plural suffx on nouns in Kyrgyz has twelve forms: *лар, лер, лор, лөр, дар, дер, дор, дөр, тар, тер, тор, төр*. The correct form is entirely predictable based on a couple phonological properties of the stem the suffx attaches to and some generalisations about Kyrgyz morphophonology: the first consonant is normally *л*, but surfaces as *д* when following voiced consonants of equal or lower sonority to *л* and as *т* after voiceless consonants; the vowel agrees in backness with the preceding vowel (*а, о* represent back vowels; *е, ө* represent front vowels), and also in roundness (*а, е* represent unrounded vowels; *о, ө* represent rounded vowels), with the exception that the unrounded *а* follows the rounded *у*. I.e., the surface value of the first consonant of the suffx is conditioned on the last segment of the stem, and the surface value of the vowel of the suffx is conditioned on the last vowel of the stem.

These two segments of the suffx are hence represented with special symbols that allow them to be treated differently from other segments: `{L}` and `{A`[1].The suffx, then, is encoded as `{L}{A}p` at the level of the morphological representation. The morphological representation is encoded entirely in `lexc` format (compiled with `hfst-lexc`), as shown in the example in Figure 2, which implements the morphological representation of the example in section 3.1 as would be done for several languages that have these particular forms.

The morphotactic transducer compiled from this `lexc` file, shown in Figure 3, maps between analyses such as `алма<n><pl><abl>` and morphological forms such as `алма>{L}{A}p>{D}{A}н`. While these form are not the final forms of the language, they allow stems of

[1] Despite the somewhat different implications of the use of the term in phonology, we call these symbols archiphonemes. Note that the choice of symbol is somewhat arbitrary, but we tend to use upper-case versions of something like the underlying phoneme for symbols that can have multiple surface representations, and lower-case versions of something like the underlying phoneme for symbols that can be present or absent and potentially also have multiple surface representations.

similar grammatical categories to share identical continuation classes, instead of having e.g., twelve different continuation classes for nouns.

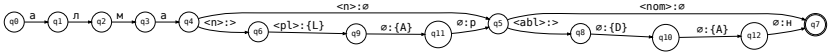

Figure 3: A simple transducer containing four mappings between analyses and morphological forms: `алма<n><nom>` ↔ `алма`, `алма<n><pl><nom>` ↔ `алма>{L}{A}р`, `алма<n><abl>` ↔ `алма>{D}{A}н`, and `алма<n><pl><abl>` ↔ `алма>{L}{A}р>{D}{A}н`. Conventions are as in Figure 1

The morphophonology is resolved by creating a two-level phonology transducer (cf. Beesley and Kart-tunen 2003) and compose-intersecting it (using `hfst-compose-intersect`) with the morphotactic transducer. The resulting transducer is a complete transducer–when done correctly for this example, the same as that in Figure 1. The two-level phonology is written in `twol` format (compiled into a transducer with `hfst-twolc`); an example including phonology for this example and other potential outcomes of the same set of processes for Kyrgyz is shown in Figure 4.

The transducers compiled from each rule are shown in Figure 5. The single FSA which shares start and end states for these transducers is compose-intersected with the transducer that maps between analyses and morphological forms, resulting in a complete morphological transducer. For this example, the resulting morphological transducer would be that in Figure 1.

In sum, a transducer is written in `lexc` format that maps between an analysis and a morphological rep-resentation, and a morphophonological transducer is written in `twol` that maps between morphological rep-resentation and orthographic form. The composition of these transducers results in a single morphological transducer that maps between analysis and orthographic form. This approach is more effcient and human-readable than implementing everything in one transducer or coding the transducers by hand.

It should be noted that the morphological analyses we have decided on, as well as the tag names that go with them, are to a large extent arbitrary, and other solutions may be at least as valid. We do make attempts to develop principled analyses, but decisions that are made are not always necessarily the best options. An example of an inconsistency might be that the nominative, despite having no overt morphological marking, is encoded with a `<nom>` tag, while the singular (in contrast

to the plural), which also has no overt morphological marking, is *not* marked with a <sg> tag, despite this being the approach taken for many languages outside of Turkic. Both approaches have their merits, but when these individual decisions are taken together they may appear to present an internal inconsistency.

```
Alphabet
 А Б В Г Д Е Ё Ж З И Й К Л М Н Ң О Ө П Р С Т У Ү Ф Х Ц Ч Ш Щ Ь Ы Ъ Э Ю Я
 а б в г д е ё ж з и й к л м н ң о ө п р с т у ү ф х ц ч ш щ ь ы ъ э ю я
 %{A%}:а %{A%}:е %{A%}:о %{A%}:ө
 %{L%}:л %{L%}:д %{L%}:т
 %{D%}:д %{D%}:т
 %>:0
;

Sets

Cns = Б В Г Д Ж З Й К Л М Н Ң П Р С Т Ф Х Ц Ч Ш Щ Ь Ъ
      б в г д ж з й к л м н ң п р с т ф х ц ч ш щ ь ъ
      %{D%} %{L%} ;

VoicelessCns = К П С Т Ф Х Ц Ч Ш Щ
               к п с т ф х ц ч ш щ ;

VoicedLowSonCns = Б В Г Д Ж З Л М Н Ң
                  б в г д ж з л м н ң ;

Rules

"{L} Desonorisation"
%{L%}:д <=> :VoicedLowSonCns %>:* _ ;

"{L} and {D} devoicing"
Cx:Cy <=> :VoicelessCns %>:* _ ;
      where Cx in ( %{L%} %{D%} )
            Cy in (   т     т   )
      matched ;

"{A} vowel harmony"
%{A%}:Vy <=> [ :LastVowel :Cns* :Cns ]/%>:* _ ;
         where LastVowel in ( и ү е э ө я а ё о ы ю у )
                       Vy in ( е ө е е ө а а о о а а а )
         matched ;
```

Figure 4. A two-level rule file that defines three different rules for the processing of Kyrgyz forms like *алмалардан* from алма>{L}{A}р>{D}{A}н as well as other nouns with plural and/or ablative morphology. The symbols, sets, and rules are defined in three separate sections of the file. Some basics of the formalism in-clude, as with lexc formalism, that the : symbol separates the two tapes (in this case, morphological form and orthographic form), and symbols without : are understood to have the same value on both tapes. The escape character is % and _ defines the locus of a rule; the <=> operator means that the pair preceding it is restricted to the environment following it. More on twol syntax can be found in Beesley and Karttunen (2003)

It should also be noted that the configuration of tags is somewhat arbitrary as well. The general con-vention is that the main part-of-speech tag follows the lemma, e.g. <n> (for nouns), <v> (for verbs), <np> (for proper nouns). Following that should come any subcategory tags, such as <iv> (for intransitive verbs) or <ant> (for anthroponyms). Then come grammatical tags, although the order is mostly language-specific. In Tur-kic languages, the order of tags follows the order of morphemes, so that e.g., in most Turkic languages for noun morphology, after <n> comes <pl> or absence thereof, followed by tags for possessive morphology when present, followed by tags for case morphology. In Chuvash, on the other hand, possessive morphology precedes plural morphology, and the tag order reflects that, resulting in a different tag order for nominal morphology between Chuvash and other Turkic languages.

Furthermore, in some Turkic languages, there are multiple possible orders of morphemes within one system. For example, in Kyrgyz (and a number of other Turkic languages), depending on the analysis, it can be understood that there are multiple possible orders of tense and negation morphemes. Some canonical negative verb forms might include *барбайм*, *барбаптырмын*, and *барбасмын*–all negative first person singular finite verb forms of *бар* ‘go’, receiving the following analyses, respectively: бар<v><iv><neg><aor><p1><sg>, бар<v><iv><neg><ifi><evid><p1><sg>, бар<v><iv><neg><fut><p1><sg>. In these forms, the negative suffx *ма* immediately follows the verb, and is followed immediately by the tense suffx and then a person/number-agreement suffx. Other negative first-person singular finite verb forms in Kyrgyz include *барган жокмун*, *барган эмесмин* and *барчу эмесмин*. The order of morphemes in these forms is verb stem, tense suf-fix, negative marker, then person agreement–that is, the tense and polarity suffxes are in the reverse or-der. Nonetheless, the analyses our analyser returns for these forms have the negative tag before the tense tag: бар<v><iv><neg><ifi><p1><sg>, бар<v><iv><neg><past><p1><sg>, and бар<v><iv><neg><pih><p1><sg>, respectively. The main reasons we implemented it this way were for consistency in tagging within a single language’s transducer and for ease of structural transfer when morphological analysis or generation is used as a component of a machine translation system. Alternative analyses are also a bit less obvious; see Washington and Tyers (2019, §6.1) for more discussion of this general problem.

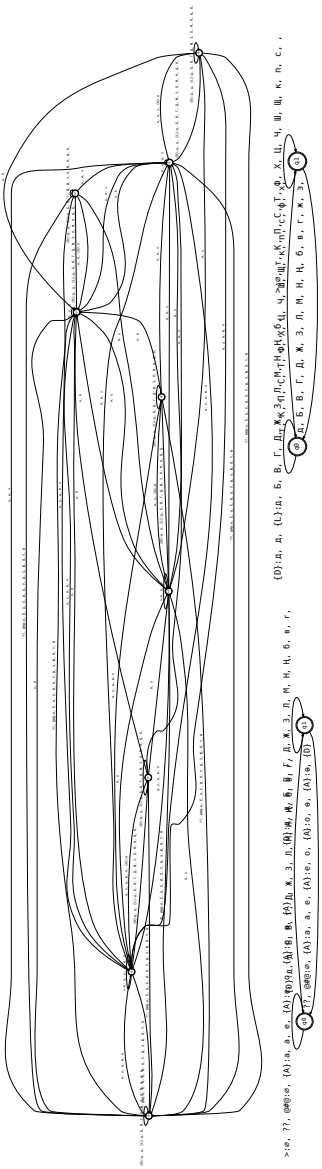

Figure 5: The finite-state transducers which are the result of compiling the twol rules in Figure 4; top: vowel harmony, bottom left: devoicing; bottom right: desonorisation

3.4. Developed morphological transducers

This section describes the Turkic morphological transducers that have been developed to date. Two main metrics are presented: number of stems in the lexicon, and naïve coverage. Naïve coverage refers to the number of forms in a corpus that receive an analysis, whether correct or not. The corpora used to arrive at naïve coverage are a full Bible translation (or when not available, New Testament translation) and the content of Wikipedia in the particular language (abbreviated 'wp'). The Bible translations are all around the same amount of content and hence size, whereas the Wikipedia corpora are of vastly different sizes.

Papers on the specific transducers (or systems that use them), referenced in this section, report further evaluation of the accuracy of each transducer. However, many of the transducers have received a lot of attention since those evaluations have been performed, so the numbers reported in those papers are mostly no longer accurate.

The languages we report on are divided into three groups based on quality of the transducer:

Production-level or **mature** transducers are those that have few known issues with morphotactics and mor-phophonology, and have a reasonably large lexicon.

Working transducers are those which either have a small lexicon or much work remaining with morphotactics and/or morphophonology.

Prototype transducers are those which have both a small lexicon of well-categorised stems and are lacking careful and complete work in the morphotactics and morphophonology.

Production-level transducers for Tatar, Kazakh, Turkish, Kyrgyz, Crimean Tatar, and Tuvan have been developed. The number of stems, coverage numbers, references, and special features of these transducers are reported in table 1[1]. Four of these transducers have been set up to generate spelling modules (Tatar, Kazakh, Kyrgyz, and Tuvan); one supports the language's two widely used orthographies (Crimean Tatar), and one supports segmentation (Kazakh). These transducers all perform at well above 90% naïve coverage. The somewhat low coverage number for the Crimean Tatar Bible is due to issues with Cyrillic analysis. Low numbers for Wikipedia corpora reflect both the

[1] Metrics reported in this paper reflect the state of the transducers in the weeks preceding the 2019 TurkLang conference, i.e., during September, 2019.

messiness of these corpora (content in other languages is often present, as are e.g. HTML symbols) and the wide range of domains covered.

Table 1: Production-level transducers, with number of stems, coverage on two corpora, previous publications, and special features

language		stems	coverage	previous publication	special features
eng name	code		Bible, wp		
Tatar	tat	59 755	98.31	Tyers et al. (2012), Salimzyanov et al.	spell checking
			92.23	(2013), and Washington et al. (2014)	
Kazakh	kaz	37 801	96.35	Salimzyanov et al. (2013), Washington	spell checking;
			90.60	et al. (2014), and Bayatlı et al. (2018b)	segmenter
Turkish	tur	22 652	93.02	Bayatlı et al. (2018b) and Gökırmak	
			87.42	et al. (2019)	
Kyrgyz	kir	15 886	94.31	Washington et al. (2012)	spell checking
			86.94		
Crimean Tatar	crh	13 631	90.86	Tyers et al. (2019) and Gökırmak	analyses Cyrillic and
			92.87	et al. (2019)	Latin orthographies
Tuvan	tyv	11 845	95.55	Washington et al. (2016)	spell checking
			88.15		

There are nine additional transducers that we consider «working»: Bashqort, Chuvash, Uzbek, Qaraqal-paq, Uyghur, Sakha, Karachay-Balkar, Gagauz, and Kumyk. These transducers perform well, but need more attention to become more robust. The Qaraqalpaq transducer analyses the Soviet-era Cyrillic orthography, as well as

several revisions of the pre-2017 Latin orthography and the current orthography–which is now its native implementation. However, as seen for the Wikipedia coverage numbers, the decade-old orthography of the Qaraqalpaq Wikipedia is not currently handled well by our transducer. Otherwise, these transducers all perform at around 90% naïve coverage, and sometimes above.

Table 2: Working transducers, with number of stems, coverage on two corpora, previous publications, and special features

language		stems	coverage	previous publication	special features
eng name	code		Bible, wp		
Bashqort	bak	56 463	93.08, 88.20	Tyers et al. (2012)	
Chuvash	chv	62 530	93.78, 90.74	–	
Uzbek	uzb	36 684	94.81, 89.89	–	
Qaraqalpaq	kaa	28 474	92.07, 67.18	–	analyses Cyrillic and multiple Latin orthographies
Uyghur	uig	25 385	90.40, 85.47	Littell et al. (2018)	
Sakha	sah	9 505	90.44, 89.73	Ivanova et al. (n.d.)	
Karachay-Balkar	krc	8 551	84.76, 88.79	–	
Gagauz	gag	6 470	91.45, 91.64	Bayatlı et al. (2018a)	
Kumyk	kum	4 949	93.01, 79.75	Washington et al. (2014)	

Another six Turkic transducers are considered prototypes: Azerbaycani, Turkmen, Noghay, Khakas, Altay, and Ottoman Turkish. Azerbaycani has a much larger number of stems than the others; however, many of these have not been categorised, and the morphotactics have not been fully implemented.

Figure 6 shows the number of stems by category of the stem. The lexicons of mature pairs contain mostly nouns and proper nouns, and the proportion of different types of stems is more consistent than the proportions of stems across the other languages.

Table 3: Prototype transducers, with number of stems and coverage on two corpora

language		stems	coverage
eng name	code		Bible, wp
Azerbaycani	aze	11 583	54.43, 45.85
Turkmen	tuk	2 986	74.40, 69.34
Noghay	nog	1 367	81.11, 75.50
Khakas	kjh	710	50.08, 50.66
Altay	alt	182	59.86, 47.04
Ottoman Turkish	ota	77	–,–

The category «Other» consists of pronouns, determiners, adverbs, postpositions, numbers, and numerals– and also stems that are not otherwise categorised. This latter category often includes stems that need to be sorted.

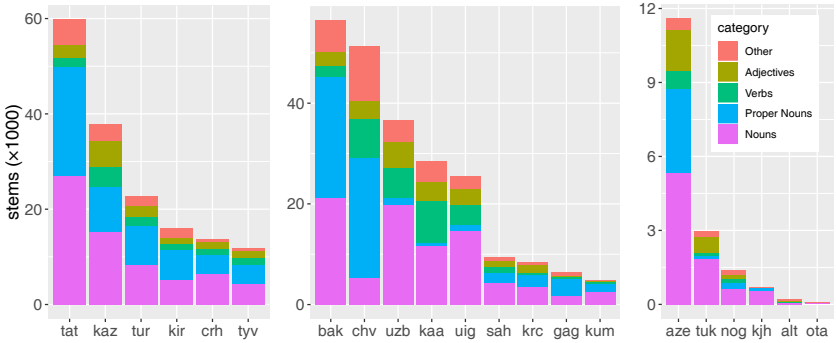

Figure 6: Number of stems by category in Apertium Turkic transducers, divided by quality of transducer

4. Machine translation systems

This section overviews what machine translation systems are along their main applications (4.1), describes the Apertium machine translation pipeline and its components (4.2), and presents the status of the machine translation systems already developed for Turkic languages in the Apertium project (section 4.3).

4.1. Overview and main applications

The purpose of a machine translation system is to render text in one language in another language. Machine translation technology has not reached the ability to replace human translators for production-ready translation, such as translation of legal documents, literature, or marketing/promotional materials. However, machine translation has the potential to save translators a lot of time in translating any material: post-editing the output of decent machine translation to create production-ready text can take far less time than translating from scratch (cf., Koponen 2016).

Another use of machine translation that does not require professional human translators is to assist a speaker of one language in understanding text written in another language. For this, the translation does not need to be production-ready or even grammatical, and may even be useful with inaccuracies.

Besides accelerating post-editing and making reading other unknown languages possible, machine trans-lation has the potential to assist in language learning. For example, using elements of a machine translation pipeline could be useful in resources like Revita (discussed in section 3.1) and other resources that use com-prehension dictionaries. Comprehension dictionaries allow language learners to look up words in a dictionary even when presented with a form that is not in the dictionary by lemmatising the form first. Disambiguation, dictionary lookup, and lexical selection stages may also be helpful for finding the best translation of a word. Geriaoueg[110] is a project that leverages elements of the Apertium pipeline for this purpose, with (currently) prototype browser plugins that enable users to dynamically look up words encountered in arbitrary internet material. Another common use of machine translation systems is as a substitute for a dictionary–i.e., if properly set up, they can be used for simple dictionary lookup.

All of these uses of machine translation systems have the ability to empower language learners, which we interpret to mean anyone learning a language, be it a foreign student, a heritage speaker, or a native speaker who wants to improve their competency.

4.2. The Apertium RBMT pipeline

The rule-based machine translation (RBMT) tools and pipeline presented by Forcada et al. (2011) and main-tained as FOSS by the

[1] `http://wiki.apertium.org/wiki/Geriaoueg`

Apertium project have been used to build the Turkic translation pairs presented in this paper. An overview of the pipeline's architecture is shown in Figure 7.

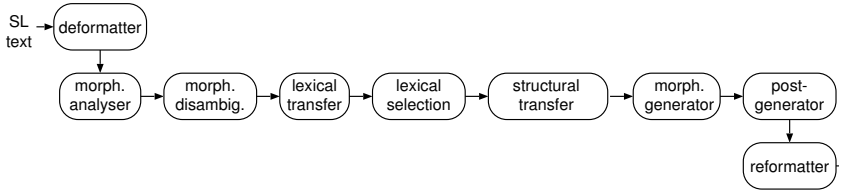

Figure 7: The standard Apertium RBMT pipeline

The pipeline consists of stages which process text output by the previous stage and feed it to the next stage. All stages of the pipeline are handled automatically by Apertium once defined for a translation pair.

One advantage of this pipeline is that it is not limited to rule-based tools: any stage is augmentable by statistical methods or replaceable by an equivalent module which could be an entirely statistical module. Furthermore, additional modules may be added arbitrarily. An example of replacing a module is that none of the Turkic languages described in this paper use the standard Apertium (`lttoolbox`) morphological transduc-ers, and instead use HFST transducers, as described in section 3. The reason for this choice is the effciency of development for suffxational agglutinative languages. Like `lttoolbox` transducers, HFST transducers are symbolic, and certain paths may also be weighted, either manually or based on the contents of a corpus (Keleg et al. n.d.). Other examples of augmenting a module statistically include learning lexical selection rules (Tyers 2013) and structural transfer rules (Sánchez-Martínez and Forcada 2009) from corpora, and learning weights for structural transfer rules from a corpus (Bayatlı n.d.). Examples of additional modules that have recently become available are presented in section 5.2.

The deformatter, provided by Apertium, removes formatting–such as font face, weight, style, etc.–and stores it for the reformatter to replace. A number of formatting engines are supported, including MS Word, LibreOffce, and `html`, allowing documents in these formats to be translated without (much) loss of document structure.

Each remaining stage of this pipeline is language- or translation pair-specific; i.e., it must be developed for a given language or

translation pair. The bilingual dictionary is used for translation in both directions, but needs a further lexical selection module because the mappings are many-to-many. A structural transfer module adjusts the tags and word order, and adds and removes any function words in order to render the translated wordforms as (hopefully) coherent sentences in the target language. An example of each of these remaining stages of the pipeline is provided in Figure 8 for Crimean Tatar to Turkish machine translation.

input	*yahşı vaqıt edi.*
morphological analysis	^yahşı/yahşı<adj>/yahşı<adj><subst><nom>/ yahşı<adj>+e<cop><aor><p3><sg> /yahşı<adj><subst><nom>+e<cop> <aor><p3><sg>$ ^vaqıt/vaqıt<n><nom>/vaqıt<n><attr>/ vaqıt<n><nom>+e<cop><aor><p3><sg>$ ^edi/e<cop><ifi><p3><sg>/ e<vaux><ifi><p3><sg>$^./.<sent>$
disambiguation	^yahşı/yahşı<adj>$ ^vaqıt/vaqıt<n> <nom>$ ^edi/e<cop><ifi><p3><sg>$^./.<sent>$
lexical transfer	^yahşı<adj>/iyi<adj>$ ^vaqıt<n><nom>/ vakit<n><nom>/süre<n><nom>$ ^e<cop><ifi><p3><sg>/ i<cop><ifi><p3><sg>$^.<sent>/.<sent>$
lexical selection	^yahşı<adj>/iyi<adj>$ ^vaqıt<n><nom>/ vakit<n><nom>$ ^e<cop><ifi><p3><sg>/ i<cop><ifi><p3><sg>$^.<sent>/.<sent>$
structural transfer	^iyi<adj>$ ^vakit<n><nom>+i<cop><ifi><p3> <sg>$^.<sent>/.<sent>$
morphological generation	*iyi vakitti.*

Figure 8: An example of a simple sentence at each stage of the Apertium pipeline. Translation from Crimean Tatar to Turkish is demonstrated with the simple Crimean Tatar sentence *yahşı vaqıt edi*. 'it was a good time.' as input. Single cohorts begin with ^ and end with $, while / separates units, where the first unit is an input form and subsequent units are possible outputs

4.2.1. Morphological analysis and generation

The morphological analyser provides analyses for each identified token in the source-language text. The mor-phological generator provides a linguistic form in the target language for each analysis output by the structural transfer element of the system. These are both implemented as a morphological transducer, as discussed in section 3.

4.2.2. Morphological disambiguation

The morphological disambiguator decides which of a series of analyses returned for a form is the correct one based on context. Rules are hand-crafted in VISL-CG3 (Bick and Didriksen 2015), and are often augmented by the HMM probabilistic tagger discussed by Forcada et al. (2011, p. 130).

In Figure 8, all cohorts returned by morphological analysis except the sentence-final period are am-biguous. Disambiguation rules in CG3 are written to remove incorrect analyses or select correct analyses. For example, the rule `REMOVE SUB:1 Cop IF (1C N) ;` removes analyses of *yahşı* (and other words) that have a subreading containing `<cop>` (an element of the list `Cop`) if the following token has an `<n>` tag (as defined by the list `N`). In other words, this rule essentially states that a form cannot be analysed as including a copula (morphologically realised as nothing for third person singular in Crimean Tatar) if the following word is a noun.

4.2.3. Lexical transfer and lexical selection

Lexical transfer is defined by a bilingual dictionary mapping between stems in the two languages of the trans-lation pair. An example is presented in Figure 9, showing the entries for *vaqıt*.

```
<e><p><l>vaqıt<s n=«n»/></l><r>vakit<s
n=«n»/></r></p></e> <e><p><l>vaqıt<s n=«n»/></
l><r>süre<s n=«n»/></r></p></e>
```

Figure 9: An excerpt from the Crimean Tatar ↔ Turkish dictionary for the Crimean Tatar noun *vaqıt* ‘time’, which translates to the Turkish nouns *vakit* and *süre*. `<e>` tags define entries, `<p>` tags enclose translation equivalents, `<l>` and `<r>` define content on the «left» and «right» sides of the dictionary, respectively, and `<s>` defines morphological tags to be matched and transferred. An optional `<par>` tag might appear after the `<p>` tag and define paradigms (the power of which is not leveraged here)

Since mappings can be many-to-many (cf. the one-to-many entry for *vaqıt* shown in Figure 9), there are often several translations to choose from. For this purpose, a lexical selection module chooses between correct translations based on the context. Figure 10 shows the lexical selection rules which choose between the correct Turkish translations of the Crimean Tatar noun *vaqıt*.

```
<rule weight=‘‘1.2’’>
    <match lemma=‘‘bir’’/>
    <match lemma=‘‘vaqıt’’ tags=‘‘n.*’’>
        <select lemma=‘‘süre’’ tags=‘‘n.*’’/>
    </match>
</rule>

<rule weight=‘‘1.0’’>
    <match lemma=‘‘vaqıt’’ tags=‘‘n.*’’>
        <select lemma=‘‘vakit’’ tags=‘‘n.*’’/>
    </match>
</rule>
```

Figure 10: Lexical selection rules to choose between two Turkish translations of the Crimean Tatar noun vaqıt

Generally, *vaqıt* translates from Crimean Tatar to Turkish as *vakit*, but the correct translation of the phrase *bir vaqıt* ‘(for) some time’ is *bir süre*. The second rule in Figure 10 has a lower weight (1.0), and accounts for the general condition, and when `vaqıt<n><*>` is matched, the translation `vakit<n><*>` is selected. The first rule, with a higher weight (1.2), overrides the general rule in the condition that the lemma `bir` immediately precedes `vaqıt<n><*>`, and outputs `süre<n><*>` instead.

There is an alternative approach to the translation problem in this particular example that would also work in the Apertium pipeline. Both versions of the approach would involve treating the phrase *bir vaqıt* / *bir süre* in each language as a single multi-word lemma (likely an adverb), and adding the translation between the two as a single entry in the bilingual dictionary; this would preclude the need for a mapping between *vaqıt* and *süre* in the dictionary as well as the need for `crh-tur` lexical selection rules dealing with choosing the correct translation. The first–and simplest–version of this approach would include these

multi-word lemmas as adverb entries in their respective morphological transducers. Some may see this as unnecessary «contamination» of the morphological transducers with multi-word expressions that are needed only for MT. The second version of this approach would «construct» the multi-word entry right after morphological analysis, and «deconstruct» it right before morphological generation. Under this version of the approach, the transducers would be clear of multi-word lemmas, and instead they would be encoded in a separate (also reversible) module. The `apertium-separable` module can handle multi-word lemmas in this way, and more, and is discussed in section 5.2.1.

```
<rule comment="REGLA: GAn soñ">
    <pattern>
        <pattern-item n="ger_past"/>
        <pattern-item n="soñ"/>
    </pattern>
    <action>
        <call-macro n="f_strip_tags"><with-param pos="1"/></call-macro>
        <let>
            <clip pos="1" side="tl" part="a_cas"/>
            <lit-tag v="abl"/>
        </let>
        <out>
            <chunk name="n" case="caseSecondWord">
                <tags>
                    <tag><lit-tag v="SP"/></tag>
                </tags>
                <lu>
                    <clip pos="1" side="tl" part="whole"/>
                </lu>
                <b/>
                <lu>
                    <clip pos="2" side="tl" part="whole"/>
                </lu>
            </chunk>
        </out>
    </action>
</rule>
```

Figure 11: An Apertium structural transfer rule which matches a past verbal noun form followed by the postposition *soñ* and outputs the sequence (inside a chunk named `n` of type `<SP>`) with the nominative case tag on the verbal noun replaced by an ablative case tag

4.2.4. Structural transfer

In an RBMT system like Apertium, structural transfer is what needs to be done to take an analysis-by-analysis translation from one language and turn it into analyses that may then be morphologically

generated to pro-duce understandable forms in another language. In other words, once the source language is analysed and disambiguated, and correct translations of each lemma are selected, there is still work to be done–usually even between very closely related languages–before this output can be generated to produce correct forms in another language. These differences can be ones of word order, grammatical tags (read: grammatical distinc-tions) that are different or absent/present in one of the languages, or even entire words that need to be added or removed in a given translation.

One simple example may be encountered in translating from Crimean Tatar to Turkish. The Crimean Tatar phrase *barğan soñ* 'after going' may be translated into Turkish as *gittikten sonra*. The non-cognate verb stems *bar* and *git* are easily dealt with during the lexical transfer and lexical selection stages, as are the partially cognate *soñ* and *sonra*. The analysed Crimean Tatar form (1a) is then rendered with different lemmas, but otherwise similarly to the input, upon output from the lexical selection stage (1b). The only remaining thing needed to be done in order to be able to render a correct Turkish form is to change the case of the verbal noun from nominative (<nom>) to ablative (<abl>), or to create the analysis in (1c).

(1) a. `^bar<v><iv><ger_past><nom>$ ^soñ<post>$` (Crimean Tatar)

b. `^git<v><iv><ger_past><nom>$ ^sonra<post>$` (intermediate)

c. `^git<v><iv><ger_past><abl>$ ^sonra<post>$` (Turkish)

The structural transfer rule in Figure 11 does just that. More on Apertium structural transfer rules can be read in Forcada et al. (2010, §3.5).

4.3. Developed machine translation pairs

Table 4 presents a list of the machine translation pairs developed between Turkic languages, along with naïve coverage in each direction on a Bible or partial Bible corpus, the number of entries in the bilingual dictionary, the number of lexical selection rules, the number of structural rules and macros, and any previous publications.

Table 5 presents a list of the machine translation pairs developed

between Turkic and non-Turkic lan-guages, along with naïve coverage in each direction on a Bible or partial Bible corpus, the number of entries in the bilingual dictionary, the number of lexical selection rules, the number of structural rules and macros, and any previous publications. Prototypes with less than 100 lexical entries are not shown.

More can be learned about the performance of published systems in their respective publications. Sys-tems may be used online at `http://turkic.apertium.org` and downloaded from GitHub `http://github.com/apertium`.

5. Future work

This section presents future work that we plan to undertake, divided into work on morphological transducers (§5.1) and machine translation systems (§5.2).

5.1. Morphological transducers

There are three main areas in which we would like to continue work on Turkic morphological transducers, be-sides expanding the coverage and accuracy of existing transducers: multiple-script support for more languages (§5.1.1), expanded spell-checker support (§5.1.2), and FSTs for more languages (§5.1.3).

5.1.1. Multiple-script support

We plan to make more of our transducers support multiple orthographies. Our current priorities are the following:

Uyghur. The Uyghur transducer only analyses and generates the Perso-Arabic script currently in use in China. However, other orthographies are used, particularly the Cyrillic alphabet as used by Uyghur-speaking communities in the former Soviet Union (Kazakhstan, Kyrgyzstan, etc.), and various Latin orthogra-phies.

Uzbek. The Uzbek transducer only analyses and generates the Latin script currently in use in Uzbekistan. However, Uzbek-speaking communities in Afghanistan use a Perso-Arabic script, and the Cyrillic script, while offcially superseded for nearly three decades in Uzbekistan, is still found in many ma-terials and is used frequently enough to warrant support.

Table 4: Turkic↔Turkic machine translation pairs developed within Apertium. A coverage value of '–' indicates that there was a problem using the translation direction to perform a coverage test when attempted

translation pair	direction	naïve coverage (Bible)	bilingual dictionary entries	lexical selection rules	structural transfer rules		previous publication
						macros	
apertium-tat-bak	tat → bak	97.34	57 710	3	3	0	Tyers et al.
Tatar↔Bashqort	bak → tat	94.05		3	3	0	(2012)
apertium-chv-tur	chv → tur	–	31 946	5	7	5	–
Chuvash↔Turkish	tur → chv	–		23	18	7	
apertium-chv-tat	chv → tat	86.62	29 512	2	9	4	–
Chuvash↔Tatar	tat → chv	86.16		11	29	5	
apertium-uzb-kaa	uzb → kaa	57.58	19 369	3	3	2	–
Uzbek↔Qaraqalpaq	kaa → uzb	58.67		3	3	2	
apertium-uig-tur	uig → tur	86.66	10 194	53	72	5	–
Uyghur↔Turkish	tur → uig	87.91		21	34	2	
apertium-kaz-tat	kaz → tat	95.03	9 956	44	36	8	Salimzyanov
Kazakh↔Tatar	tat → kaz	93.57		4	10	3	et al. (2013)
apertium-tur-aze	tur → aze	–	8 211	0	4	0	–
Turkish↔Azerbaycani	aze → tur	–		0	0	0	
apertium-kaz-kir	kaz → kir	92.64	8 177	59	33	6	–
Kazakh↔Kyrgyz	kir → kaz	89.17		3	7	3	
apertium-tur-kir	tur → kir	88.27	7 824	539	62	6	–
Turkish↔Kyrgyz	kir → tur	92.76		2	68	8	
apertium-kaz-tur	kaz → tur	89.05	7 636	96	88	4	Bayatlı et al.
Kazakh↔Turkish	tur → kaz	82.25		3	7	4	(2018b)
apertium-crh-tur	crh → tur	46.98	7 140	15	109	12	Gökırmak et
Crimean Tatar↔Turkish	tur → crh	77.31		3	31	5	al. (2019)
apertium-kaz-kaa	kaz → kaa	92.41	5 425	7	5	0	–
Kazakh↔Qaraqalpaq	kaa → kaz	88.24		2	5	0	
apertium-tur-uzb	tur → uzb	80.04	4 359	3	5	2	–
Turkish↔Uzbek	uzb → tur	84.92		18	43	2	
apertium-tur-tat	tur → tat	80.90	4 272	3	36	6	–
Turkish↔Tatar	tat → tur	92.21		5	23	5	
apertium-kaz-sah	kaz → sah	67.93	2 890	3	14	4	–
Kazakh↔Sakha	sah → kaz	63.47		3	4	2	
apertium-kaz-uig	kaz → uig	79.13	2 728	7	4	0	–
Kazakh↔Uyghur	uig → kaz	57.33		2	2	0	
apertium-kaz-kum	kaz → kum	56.36	563	3	10	0	–
Kazakh↔Kumyk	kum → kaz	–		0	0	0	
apertium-kir-uzb	kir → uzb	64.63	323	2	18	5	–
Kyrgyz↔Uzbek	uzb → kir	51.41		2	18	5	
apertium-kaz-tyv	kaz → tyv	57.56	159	3	10	3	–
Kazakh↔Tuvan	tyv → kaz	59.05		3	4	2	

Table 5: Turkic↔non-Turkic machine translation pairs developed within Apertium

translation pair	direction	naïve coverage (Bible)	bilingual dictionary entries	lexical selection rules	structural transfer		previous publication
					rules	macros	
apertium-eng-kaz	eng → kaz	96.15	33 002	78	351	35	Sundetova et al. (2014, 2015)
English↔Kazakh	kaz → eng	96.97		101	263	27	
apertium-kaz-rus	kaz → rus	91.90	29 777	35	198	28	–
Kazakh↔Russian	rus → kaz	98.76		23	216	24	
apertium-tat-eng	tat → eng	83.37	14 985	15	20	7	–
Tatar↔English	eng → tat	97.42		15	20	7	
apertium-tat-rus	tat → rus	90.88	6 036	64	38	10	–
Tatar↔Russian	rus → tat	–		0	0	0	
apertium-eng-kir	eng → kir	57.42	368	6	8	1	–
English↔Kyrgyz	kir → eng	59.85		45	85	12	
apertium-khk-kaz	khk → kaz	40.69	140	2	15	4	–
Kazakh↔Khalkha	kaz → khk	53.47		2	7	3	

Azerbaycani. The Azerbaycani transducer only analyses and generates the Latin script currently in use in Azerbaycan. However, Azerbaycani-speaking communities in Iran use a Perso-Arabic script, and two separate Wikipedias even exist using the two standards. The Cyrillic script used in Azerbaycan until the 1990s (and which continues to be used in other areas) could also be supported.

Kazakh. The Kazakh transducer currently only analyses and generates the Cyrillic orthography currently in use in Kazakhstan. However, Kazakh-speaking communities in China use a Perso-Arabic script, and Kazakhstan is transitioning to a Latin orthography. The transducer should support analysis and generation of both of these scripts.

5.1.2. Speller support

Another important step for Apertium transducers would be to make spell checkers available by default. Cur-rently Apertium transducers for only four Turkic languages are set up to generate spell checkers: Kazakh, Kyrgyz, Tatar, and Tuvan. These are usable «out of the box»

in LibreOffce and voikospell, and are packaged for Microsoft Word[1] as well.

In principle it is trivial to set up a language module to generate a spell checker from its transducer–we simply have yet to apply this to all the languages. We also have not yet developed a method to generate spell checkers for our multiple-orthography transducers, though in principle this should not be diffcult either.

We also strive to find ways to implement spellers into mobile device input systems.

5.1.3. Additional languages

There are also a number of Turkic languages for which Apertium transducers do not exist, such as Dolgan, Qashqai, Shor, and Salar. We hope to develop transducers for any language whose community might benefit from the existence of these tools.

5.2. MT systems

There are a number of new modules available for the Apertium RBMT pipeline that were not available when most of the existing MT systems were begun. While still unpublished, these modules offer powerful new ways to approach problems that were previously diffcult or impossible to deal with. While prototype-level use has demonstrated that these modules solve many problems, more work is needed to integrate them into production-quality language pairs.

A revised version of the Apertium pipeline with these three modules included is presented in Figure 12.

This extends the pipeline presented in Figure 7.

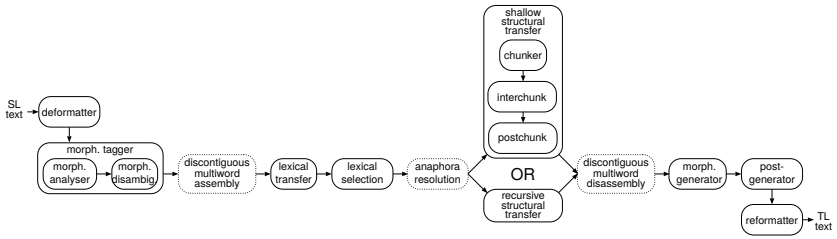

Figure 12: The Apertium pipeline with new optional modules (shown with broken-line borders) and the two options for structural transfer

[1]Available at https://apertium.projectjj.com/spellers/nightly/

The new modules, each described individually in the subsections that follow, include a discontiguous multiword and separable expressions processor (§5.2.1), a module for resolution of anaphora (§5.2.2), and a recursive structural transfer engine (§5.2.3). Separate papers on each of these modules are in preparation.

5.2.1. Multiwords and separable expressions

The `apertium-separable` module was developed by Irene Tang as a Google Summer-of-Code project in 2017 to deal with multi-word expressions (MWEs), and has yet to be published. Processors for this module may be included in two places in the Apertium RBMT pipeline: immediately following morphological tagging and preceding lexical transfer, and immediately following structural transfer and preceding morphological gener-ation. The former use allows «assembly» of source-language MWEs for transfer, and the latter «disassembles» transferred target-language MWEs for morphological generation.

Example (2) shows a Kyrgyz sentence with corresponding English translation which benefits from the `apertium-separable` module in both translation directions.

(2) a. *Мугалим сынык калемди ташгады.*
teacher broken pen– throw away– PAST.3 (Kyrgyz)

b. *The teacher threw the broken pen away.* (English)

When translating from Kyrgyz to English, the best translation we might be able to hope for without a comparable module in the pipeline is *Мугалим сынык калемди ары ыргытты* ‘The teacher threw the broken pen [in a direction away from themself]’. What `apertium-separable` does before lexical and structural transfer is assembles «threw» and «away» as single lexical unit, despite being separated by an intervening noun phrase.

The general principles for how this works follow. The module accepts flat definitions of phrase types, so for instance, a noun phrase might be defined such that one match for it is a series of analyses where the first contains a `<det>` tag, the second contains an `<adj>` tag, and the last contains a <n> tag. This pattern matches *the broken pen*. A pattern is then set up that matches the analysis `throw<vblex><*>`,

followed by a noun phrase, followed by away<adv>. This pattern is one tape of a two-tape FSA; the second tape contains throw# away<vblex> on the path that matches throw<vblex>, is empty on the path that matches away<adv>, and has the noun-phrase definition in the same path on both tapes. This results in a sequence of throw<vblex><*> [NP] away<adv> being replaced with throw# away<vblex><*> [NP], where <*> and [NP] both have the same value in the output as in the input. It is then trivial to add an appropriate entry for throw# away<vblex> in the bilingual dictionary.

In generating English text from e.g. Kyrgyz text, the same transducer may be used, with the input and output sides reversed, as is the bilingual dictionary transducer. The result is a single Kyrgyz dictionary entry like ташта<v><tv> being translated to throw# away<vblex>, and then being matched by the apertium-separable transducer. In the example sentence above, throw# away<vblex><*> [NP] is disassembled to throw<vblex><*> [NP] away<adv>, generating correct English.

The necessity of this module for generation may not seem as important as for analysis, since the English sentence *The teacher threw away the broken pen* is also grammatical and has the same meaning as (2b). How-ever, not all separable expressions in English are able to be rendered both ways. Consider for example the verb ***ысыт*** 'make hot', the causative of ***ысы*** 'be/get hot'. With a dictionary entry that matches ысы<v><iv><caus> with make# hot<vblex> and no module for processing separable expressions, a translation of a sentence like *Күн балдарды ысытты* 'The sun made the kids hot' would be generated as *The sun made hot the kids*, which is ungrammatical. A simple rule for expressions like «make hot» solves this problem.

Simple contiguous multi-word expressions can also be handled by this module, allowing for more robust bilingual dictionary entries with fairly vanilla morphological transducers, as discussed in section 4.2.3. For example, it may not make sense to have an entry for *little brother* in an English morphological transducer that already contains the component words, but it is useful to have in a bilingual dictionary with a language like Kyrgyz that has two words for brother with the difference in meaning associated with relative age to a sibling. The apertium-separable module, then, converts between the analysis of *little brother* as an adjective and a noun and the analysis of it as a multi-word noun.

5.2.2. Anaphora resolution

The `apertium-anaphora` module was developed by Tanmai Khanna as a Google Summer-of-Code project in 2019 to deal with problems of anaphora resolution, and has yet to be published. The processor for this module may be placed in the Apertium RBMT pipeline immediately preceding structural transfer.

An example of the utility of this module in resolving anaphora might be to select the right third-person singular pronoun in English (*he, she, it*) when translating from a Turkic language, where only one third person pronoun is used. The module can leverage existing categories (such as gender tags on personal and family names) or can use encoded lists of lemmas in various categories (e.g., *girl, aunt, hen* and *boy, uncle, rooster*) to identify anaphora in preceding text and decide on the correct translation of a given word.

5.2.3. Apertium-recursive

The `apertium-recursive` module (Swanson et al. n.d.) was developed by Daniel Swanson as a Google Summer-of-Code project in 2019 to deal with structural transfer issues that are diffcult or impossible to solve with Apertium's shallow transfer module. Such issues especially encompass any sort of deep embedding, including recursion.

Rules in this module match patterns and produce output based on the matched material, but potentially in different configurations (in terms of word order, tags, etc.). The output of a rule forms a node that can then be matched by another rule. Since application is recursive, syntactic parsing is exhaustive. An example of a parse tree built for a relatively simple sentence is shown in Figure 13.

An application for this module that would be diffcult or impossible to implement using Apertium's shallow transfer module might be to transfer sentences with multiple verb phrases in a subordinate or relative clause relationship with one another. For example, sentences like *I said that the teacher threw the broken pen away* and *I saw the teacher throw the broken pen away* have a main verb phrase (headed by *said* and *saw*, respectively) and a secondary verb phrase (headed by *throw*). In most Turkic languages, translations of these sentences would need the main verb to follow the secondary verb phrase, not to mention the supporting morphological adjustments (e.g., the secondary verb phrase in the *saw* example would need to be an accusative verbal noun

in many Turkic languages). The `apertium-recursive` module is able to handle situations like this, whereas the shallow transfer module is not able to.

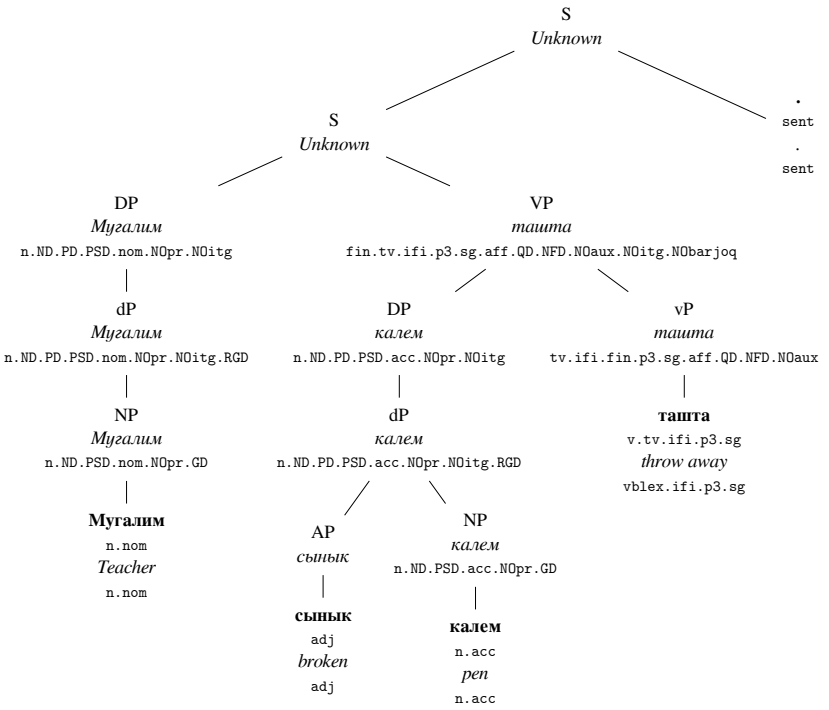

Figure 13: A parse tree of the Kyrgyz sentence in (2a) as output by the current rule set of the `apertium-recursive` model in the `apertium-eng-kir` pair

6. Conclusion

This paper has overviewed free/open-source language technology for a number of Turkic languages, consist-ing of morphological transducers and machine translation systems. Various uses for these technologies were outlined, including in ways that have the potential to mitigate language loss, and the decisions to develop the resources under free/open-source licenses and as symbolic models were motivated. It was argued that these ap-proaches to this work have the potential to better engage the language communities, contribute to the linguistic

rights of speakers of the languages, and enhance the potential for the languages' continued intergenerational transmission.

Acknowledgements

The authors would like to acknowledge the generous support Apertium has received under the Google Sum-mer of Code and Google Code-In programmes, which has made this work much more feasible. The work on *apertium-kaz* was also partially supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan, contract #346/018-2018/33-28, IRN AP05133700. We would additionally like to recognise the work of everyone who has contributed time to development of the tools presented by this paper; besides the authors and those already cited, this includes Hèctor Alòs i Font, Nikolay Aleksandrov, Anna Zueva, Kantörö Erkulov, Gianluca Grossi, Sharapat Kalabaev, Mansur Saykhunov, Beknazar Abdikamalov, Assel Baltabayeva, Zhenisbek Assylbekov, Akın Dalkı, Ağarahim Sultanmuradov, and Tolgonay Kubatova.

REFERENCES

Aimoldina, Aliya (2019). «Three Languages of Business Discourse in Kazakhstan: Achievements, Challenges and What is Next?» Central Eurasian Studies Society 20th Annual Conference (Oct. 12, 2019). Washington, D.C. URL: https://nomadit.co.uk/conference/cess2019/paper/50424.

Alòs i Font, Hèctor (2014). «Chuvash Language in Chuvashia's Instruction System: An Example of Educational Language Policies in Post-Soviet Russia». In: *Journal on Ethnopolitics and Minority Issues in Europe* (4), pp. 52–84. URL: https://www.infoecmi.eu/index.php/new- jemie- online- volume- 13- issue-42014/.

Bayatlı, Sevilay (n.d.). «An unsupervised maximum-entropy approach to transfer-rule selection in rule-based machine translation». In preparation.

Bayatlı, Sevilay, Güllü Karanfil, Memduh Gökırmak, and Francis M. Tyers (2018a). «Finite-state morpho-logical analysis for Gagauz». In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European Languages Resources Associa-tion (ELRA). URL: https://www.aclweb.org/anthology/L18-1411.

Bayatlı, Sevilay, Sefer Kurnaz, Ilnar Salimzyanov, Jonathan North Washington, and Francis M. Tyers (2018b). «Rule-based machine translation from Kazakh to Turkish». In: *Proceedings of the 21st Annual Con-ference of the European Association for Machine Translation (EAMT 2018)*. Ed. by Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Maja Popović, Celia Rico, André Mar-tins, Joachim Van den

Bogaert, and Mikel L. Forcada, pp. 49–58. : 978-84-09-01901-4. URL: http://hdl.handle.net/10045/76020.

Beesley, Kenneth R. and Lauri Karttunen (2003). «Two-Level Rule Compiler». URL: https://web.stanford.edu/~laurik/.book2software/twolc.pdf.

Bick, Eckhard and Tino Didriksen (2015). «CG-3 – Beyond Classical Constraint Grammar». In: *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. Vilnius, Lithuania: Linköping University Electronic Press, Sweden, pp. 31–39. URL: https://www.aclweb.org/anthology/ W15-1807.

Bird, Stephen (2009). «Natural Language Processing and Linguistic Fieldwork». In: *Computational Linguistics* 35 (3), pp. 469–474. DOI: 10.1162/coli.35.3.469.

Cherivirala, Sushain, Shardul Chiplunkar, Jonathan Washington, and Kevin Unhammer (2018). «Apertium's Web Toolchain for Low-Resource Language Technology». In: *Proceedings of the AMTA 2018 Work- shop on Technologies for MT of Low Resource Languages (LoResMT 2018)*. Boston, MA: Association for Machine Translation in the Americas, pp. 53–62. URL: https://www.aclweb.org/anthology/W18-2207.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, eds. (2019). *Ethnologue: Languages of the World*. Online version. Dallas, Texas: SIL International. URL: http://www.ethnologue.com.

Forcada, Mikel L., Boyan Ivanov Bonev, Sergio Ortiz Rojas, Juan Antonio Perez Ortíz, Gema Ramírez Sánchez, Felipe Sánchez Martínez, Carme Armentano-Oller, Marco A. Montava, and Francis M. Tyers (2010). «Documentation of the Open-Source Shallow-Transfer Machine Translation Platform *Aper-tium*». URL: http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers (2011). «Aper-tium: a free/open-source platform for rule-based machine translation». In: *Machine Translation* 25.2: *Free/Open-Source Machine Translation*, pp. 127–144.

FSFE, Free Software Foundation Europe (2019). *Reasons for public code*. Date accessed: 2019-10-17. Public Money, Public Code: Code paid by the people should be available to the people! URL: https://publiccode.eu/#arguments.

Gökırmak, Memduh, Francis M. Tyers, and Jonathan North Washington (2019). «A free/open-source rule-based machine translation system for Crimean Tatar to Turkish». In: pp. 24–31. URL: https://www.aclweb.org/anthology/W19-68#page=30.

Ivanova, Sardana, Anisia Katinskaia, and Roman Yangarber (2019). «Tools for supporting language learning for Sakha». In: *Proceedings of the*

22nd Nordic Conference on Computational Linguistics (NoDaL-iDa'19). Linköping University Electronic Press.

Ivanova, Sardana, Francis Tyers, and Jonathan Washington (n.d.). «A free/open-source morphological analyser and generator for Sakha». In preparation.

Johnson, Ryan, Lene Antonsen, and Trond Trosterud (2013). «Using Finite State Transducers for Making Ef-ficient Reading Comprehension Dictionaries». In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Vol. 85, pp. 59–71.

Katinskaia, Anisia, Javad Nouri, and Roman Yangarber (2018). «Revita: a language-learning platform at the intersection of ITS and CALL». In: *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*. Miyazaki, Japan.

Keleg, Amr, Nick Howell, Tommi A. Pirinen, and Francis M. Tyers (n.d.). «Unsupervised weighting of finite-state transducers for morphological analysis». In preparation.

Koponen, Maarit (2016). «Is machine translation post-editing worth the effort? A survey of research into post-editing and effort». In: *The Journal of Specialised Translation* (25), pp. 131–148.URL: https : //www.jostrans.org/issue25/art_koponen.pdf.

Kornai, András (2013). «Digital Language Death». In: *PLOS ONE* 8.10, pp. 1–11. DOI: 10 . 1371 / journal . pone.0077056.

Koskenniemi, Kimmo (1983). *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. 11. Helsinki: University of Helsinki Department of General Linguistics.ISBN: 951-45-3201-5.

Linden, Krister, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen (2011). «HFST– Framework for Compiling and Applying Morphologies». In: *Systems and Frameworks for Compu-tational Morphology*. Ed. by Cerstin Mahlow and Michael Pietrowski. Vol. 100. Communications in Computer and Information Science, pp. 67–85.

Littell, Patrick, Tian Tian, Ruochen Xu, Zaid Sheikh, David Mortensen, Lori Levin, Francis Tyers, Hiroaki Hayashi, Graham Horwood, Steve Sloto, Emily Tagtow, Alan Black, Yiming Yang, Teruko Mitamura, and Eduard Hovy (2018). «The ARIEL-CMU situation frame detection pipeline for LoReHLT16: a model translation approach». In: *Machine Translation* 32.1, pp. 105–126. ISSN: 1573-0573. DOI: 10. 1007/s10590-017-9205-3.

Pedersen, Ted (2008). «Empiricism Is Not a Matter of Faith». In: *Computational Linguistics* 34.3, pp. 465–470. DOI : 10.1162/coli.2008.34.3.465.

Salimzyanov, Ilnar, Jonathan North Washington, and Francis Morton Tyers (2013). «A free/open-source Kazakh-Tatar machine translation system». In: *Proceedings of the XIV Machine Translation Summit*. Ed. by K.

Sima'an, M.L. Forcada, D. Grasmick, H. Depraetere, and A. Way. European Association for Machine Translation, pp. 175–182. URL: http://www.mt-archive.info/10/MTS-2013-Salimzyanov.pdf.

Sánchez-Martínez, Felipe and Mikel L. Forcada (2009). «Inferring Shallow-Transfer Machine Translation Rules from Small Parallel Corpora». In: *Journal of Artificial Intelligence Research* 34, pp. 605–635.

Silfverberg, Miikka and Francis Tyers (2019). «Data-Driven Morphological Analysis for Uralic Languages». In: *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*. Tartu, Estonia: Association for Computational Linguistics, pp. 1–14. URL: http : / / aclweb. org/anthology/W19-0301.

Starr, S. Frederick (2009). «Rediscovering Central Asia». In: *The Wilson Quarterly* (Summer 2009). URL: http://archive.wilsonquarterly.com/essays/rediscovering-central-asia.

Streiter, Oliver, Kevin P. Scannell, and Mathias Stuflesser (2006). «Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers». In: *Machine Translation* 20.4, pp. 267–289. DOI: 10.1007/s10590-007-9026-x.

Sundetova, Aida, Mikel Forcada, and Francis Tyers (2015). «A free/open-source machine translation system for English to Kazakh». In: *Proceedings of the International Conference «Turkic Language Processing» (TurkLang 2015)*. Kazan, Tatarstan, pp. 78–90.

Sundetova, Aida, Aidana Karibayeva, and Ualsher Tukeyev (2014). «Structural transfer rules for Kazakh-to-English machine translation in the free/open-source platform Apertium». In: vol. 7. 2. Türkiye Bilişim Vakfı, pp. 48–53. URL: https://dergipark.org.tr/tr/pub/tbbmd/issue/33580.

Swanson, Daniel G., Jonathan N. Washington, Francis M. Tyers, and Mikel L. Forcada (n.d.). «Apertium-recursive: A Free/Open-Source Structural Transfer Module for the Apertium Machine Translation Plat-form». In preparation.

Toral, Antonio, Lukas Edman, Galiya Yeshmagambetova, and Jennifer Spenader (2019). «Neural Machine Translation for English–Kazakh with Morphological Segmentation and Synthetic Data». In: *Proceed-ings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Flo-rence, Italy: Association for Computational Linguistics, pp. 386–392. URL: https://www.aclweb.org/ anthology/W19-5343.

Tyers, Francis M. and Jonathan N. Washington (2015). «Towards a Free/Open-source Universal-dependency Treebank for Kazakh». In: *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pp. 276–289.

Tyers, Francis M., Jonathan Washington, Darya Kavitskaya, Memduh Gökırmak, Nick Howell, and Remziye Berberova (2019). «A biscriptual morphological transducer for Crimean Tatar». In: vol. 1, pp. 74–80. URL: https://scholar.colorado.edu/scil-cmel/vol1/iss1/10.

Tyers, Francis Morton (2013). «Feasible lexical selection for rule-based machine translation». PhD thesis. Universitat d'Alacant.

Tyers, Francis, Jonathan North Washington, Ilnar Salimzyan, and Rustam Batalov (2012). «A prototype ma-chine translation system for Tatar and Bashkir based on free/open-source components». In: *Proceedings of the First Workshop on Language Resources and Technologies for Turkic Languages at the Eighth In-ternational Conference on Language Resources and Evaluation (LREC'12)*. İstanbul, Turkey, pp. 11–14. URL: http : / / www . lrec - conf . org / proceedings / lrec2012 / workshops / 02 . Turkic % 20Languages %20Proceedings.pdf#page=16.

United Nations (1948). *Universal Declaration of Human Rights*. 217 (III) A. Paris: Offce of the United Nations High Commissioner for Human Rights. URL: http://www.un.org/en/universal-declaration-human-rights/.

Washington, J. N., A. Bayyr-ool, A. Salchak, and F. M. Tyers (2016). «Development of a finite-state model for morphological processing of Tuvan». In: *Родной Язык* 1.4, pp. 156–187. URL: http://rodyaz.ru/pdf/no.4_2016/Washington%20J., %20Bayyr- ool%20A., %20Salchak%20A., %20Tyers%20F.%20Development% 20of%20a%20finite-state%20model%20for%20morphological%20processing%20of%20Tuvan.pdf.

Washington, Jonathan North, Ilnar Salimzyanov, and Francis M. Tyers (2014). «Finite-state morphological transducers for three Kypchak languages». In: *Proceedings of the Ninth International Conference on*

Language Resources and Evaluation (LREC'14), pp. 3378–3385. URL: http://www.lrec- conf.org/ proceedings/lrec2014/summaries/1207.html.

Washington, Jonathan North and Francis M. Tyers (2019). «Delineating Turkic non-finite verb forms by syn-tactic function». In: *Proceedings of the Fourth Workshop on Turkic and Languages in Contact with Turkic (Tu+4)*. Ed. by Paloma Jeretič and Yağmur Sağ, pp. 132–146. DOI: 10.3765/ptu.v4i1.4587.

Washington, Jonathan, Mirlan Ipasov, and Francis Tyers (2012). «A finite-state morphological transducer for Kyrgyz». In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari et al. İstanbul, Turkey: European Language Resources Associa-tion (ELRA), pp. 934–940. ISBN: 978-2-

9517408-7-7. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/1077.html.

Wheeler, David A. (n.d.). *Why Open Source Software / Free Software (OSS/FS, FLOSS, or FOSS)? Look at the Numbers!* URL: https://dwheeler.com/oss_fs_why.html.

Zampieri, Marcos, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen (2019). «A Report on the Third VarDial Evaluation Campaign». In: *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*. TOBEFILLED-Ann Arbor, Michi-gan: Association for Computational Linguistics, pp. 1–16. DOI: 10.18653/v1/W19-1401. : https://www.aclweb.org/anthology/W19-1401.

УДК 81’33

BASIC MEANINGS OF THE ABLATIVE CASE IN TATAR (ON CORPUS DATA)

Alfiya Galieva
Tatarstan Academy of Sciences, Kazan, Russia
amgalieva@gmail.com

The task of developing an empirically oriented corpus grammar of the Tatar language involves a detailed description of individual grammatical categories, including cases. The Ablative is not only one of main ways to express spatial localization, but also to convey a significant number of other semantic roles, one way or another connected with expressing the prototypical meaning of motion away from something or source: reference point, cause, object and others.

This paper is aimed at inventorying the main ways of using of the Tatar Ablative case according to the corpus data, joining heterogeneous meanings into semantic groupings according to the basic semantic features and types of controlling predicates. Meanings of the Ablative are distinguished with a focus on semantic roles expressed by word forms.

The meanings of individual word forms with the ablative affix within syntactic items are not strictly determined by the semantic class of a lexeme; the material of the Tatar language confirms the thesis of typologists that a specific semantic role of the lexeme is part of the semantics of the controlling word (predicate) and this role reflects general properties of the participants in certain types of situations.

The need for this study is due to the fact that the existing Tatar grammars provide an extremely superficial and insufficiently systematic description of the case system of the Tatar language, including ablative.

Keywords: spatial cases; Tatar language; ablative; corpus data.

ОСНОВНЫЕ ЗНАЧЕНИЯ ИСХОДНОГО ПАДЕЖА В ТАТАРСКОМ ЯЗЫКЕ (НА КОРПУСНЫХ ДАННЫХ)

А. М. Галиева
Академия наук Республики Татарстан, Казань, Россия
amgalieva@gmail.com

Задача разработки эмпирически ориентированной корпусной грамматики татарского языка предполагает детальное описание отдельных грамматических категорий, в том числе падежей. Аблатив – исходный падеж – является не только выразителем пространственной локализации, но и значительного числа других семантических ролей, так или иначе сопряженных с выражением прототипического значения исходной точки и источника: точки отсчета, причины, стимула, косвенного объекта и ряда других.

В статье ставится задача инвентаризации основных случаев употребления татарского исходного падежа по корпусным данным, объединение гетерогенных значений в семантические группировки по основным значениям и типам управляющих предикатов. Значения аблатива выделяются с фокусом на семантические роли, выражаемые словоформами.

Значения конкретных словоформ с аффиксом аблатива в составе синтаксического целого не детерминированы жестко семантическим классом исходной лексемы; материал татарского языка подтверждает тезис типологов о том, что конкретная семантическая роль лексемы является частью семантики управляющего слова (предиката) и отражает общие свойства участников определенных типов ситуаций.

Необходимость данного исследования обусловлено тем, что существующие грамматики татарского языка дают крайне поверхностное и недостаточно системное описание падежной системы татарского языка, в том числе и аблатива.

Ключевые слова: пространственные падежи; татарский язык; аблатив; корпусные данные.

1.1. Введение

Существующие на сегодняшний день описательные грамматики татарского языка созданы без учета количественных данных и статистических закономерностей (Закиев, 1993; Хисамова, 2015; Зәкиев, 2016). К настоящему времени разработаны корпусы татарского языка (Татарский национальный корпус «Туган тел» http://tugantel.tatar/ (Невзорова, Мухамедшин, Билалов, 2015), Письменный корпус татарского языка http://www.corpus.tatar/) (Ибрагимов, Сайхунов, 2014), представляющее собой новые общедоступные инструменты для получения объективных данных о языке, лингвистического анализа, изучения активных процессов в языке. Проведение описательных и лингвостатистических исследований на корпусных данных – первый и необходимый шаг для создания современной корпусной грамматики татарского языка. Необходимо отметить еще один важный аспект: большая часть существующих исследований по татарской грамматике, выполненных татарстанскими учеными, не учитывают достижений современной лингвистической типологии, добившейся существенных результатов в исследовании языков, отличающихся от индо-европейских. Поэтому вопрос о пересмотре и уточнении грамматических описаний, в том числе и традиционно выделяемых категорий татарского языка, сохраняется актуальным, с уче-

том имеющихся инструментов для получения репрезентативных данных и достижений в теории языка.

В данной статье ставится задача выделения, анализа и систематизации значений татарского исходного падежа (аблатива) на данных Татарского национального корпуса «Туган тел» (Невзорова, Мухамедшин, Билалов, 2015). Перевод примеров на русский язык выполнен автором статьи.

1.2. Исходный падеж в татарских грамматиках

В академических грамматиках падежная система татарского языка описана недостаточно полно, с выделением лишь основных значений падежей (Закиев, 1993; Зәкиев 2016). В частности, в новой татарской грамматике 2016 года значения исходного падежа определены лишь в самом общем виде: отмечены значение места, времени, причины, объекта, меры и степени действия (Зәкиев, 2016, с. 53). В грамматике Ф. М. Хисамовой (2015) представлено значительно более полное описание значений татарских падежей, для исходного падежа отмечены следующие значения, которые иллюстрируются примерами:

1) исходная точка действия или состояния;
2) место действия или события;
3) время;
4) косвенный объект действия;
5) сравнение;
6) материал;
7) причина;
8) цена (при употреблении со словами, обозначающие денежные единицы);
9) случаи употребления с послелогами *соң, башка, бирле, тыш, элек* и др. (Хисамова, 2015, с. 88–89).

Тем не менее приведенные значения требуют большей систематизации и уточнения, особенно в аспекте связи частных значений исходного падежа с семантическими классами управляющих лексем.

1.3. Основные значения аблатива по корпусным данным

Во многих современных исследованиях функции падежей связываются с выражаемыми ими семантическими ролями; В. А. Плунгян указывает, что в любой не редуцированной до край-

ности падежной системе падеж с точки зрения функции является в первую очередь показателем роли имени и лишь вследствие этого – маркером синтаксической зависимости имени (Плунгян, 2003, с. 167).

Показатели направления в языках конкретизируют различные векторы движения, а также отсутствия движения пространственно характеризуемого объекта относительно некоей заданной области локализации (Кибрик, 2005: 203).

В тюркских языках для выражения локализации и направления используются набор (подсистема) из 3 падежей, иллюстрируемых Таблицей 1.

Таблица 1. Локализация в татарском языке (базовые падежи)

Пространственные падежи	**Основные значения**	**Пример**
Директив (направительный)	Двигаться к *х*	*мәктәпкә* ‘в школу’
Локатив (местно-временной)	Находиться в/на *х*	*мәктәптә* ‘в школе’
Аблатив (исходный)	Двигаться от *х*	*мәктәптән* ‘из школы

Пространственные падежи татарского языка характеризуются сложной и разветвлённой сетью непространственных употреблений, и далеко не во всех случаях эти частные значения являются симметричными для разных падежей.

Ниже рассмотрим основные значения татарского аблатива.

1. Источник, исходная точка

1.1 Исходная точка движения

При глаголах движения – исходная точка движения или перемещения объекта:

Казаннан *килгән язучы сез буласызмы?* (Ягсуф Шафиков). ‘Это вы писатель, который приехал **из Казани**?’

Югыйсә хәзер сине ***урманнан*** *куам!* (Разиль Валиев). ‘Иначе сейчас тебя прогоню **из леса**’.

Лексема со значением движения может быть подразумеваться, например, при указании происхождения:

Яңа автомобильләрдә өйләренә кайтучыларның дүртесе – ***Казаннан*** **(**«Татар-информ»). ‘Четверо из тех, кто приехал домой на новой машине, **из Казани’**.

Значение исходной точки при переносных употреблениях глагольных лексем:

Әтидән *уйларым авылга, авылның халкына күчә* (Нур Ахмадиев). ‘Мысли **от отца** перешли к деревне, к ее жителям’.

Атасыз үскән ***малайдан*** *нинди ир чыксын?* (Кояш Тимбикова). ‘Какой мужчина может выйти **из мальчика**, росшего без отца?’

Значение исходной точки для аблатива можно считать базовым, прототипическим, и именно оно дало название падежу – «исходный».

1.2. Траектория перемещения

При глаголах движения – обозначение траектории – совокупности точек, в которых находился или находится движущийся объект:

Шушы ***юлдан*** *якындагы поселоклардан балалар укырга, Дим бистәсендә яшәүчеләр эшкә йөри иде* (Ильдар Фазлетдинов). ‘**По** этой **дороге** ходили дети из ближайших поселков учиться, рабочие из Демской слободы – на работу’.

Җәен әйләнгеч ***юлдан*** *шактый йөрергә туры килгән килүен* («Татар-информ»). ‘Летом довольно долго пришлось ходить **по** окружной **дороге**’.

Нәкъ Гәрәйнең ***баскычыннан*** *менеп баралар* (Галимджан Ибрагимов). ‘Поднимаются именно **по лестнице** Гарая’.

Карале, нигә ***тәрәзәдән*** *йөрисең әле син?* (Захид Махмуди). ‘Послушай, почему же ты ходишь **через окно**?’

Чиләкләре дә зур шул, тигез ***җирдән*** *атлаганда да җиргә тиям-тиям дип баралар* (Лира Ибрагим-Валиди). ‘И ведра большие, даже при ходьбе **по** ровной **земле** едва не касаются земли’.

В данном случае могут использоваться как глаголы ненаправленного, так и направленного движения.

1.3. Место (участок пространства, в котором локализована ситуация)

Часто – объект приобретается/добывается в месте, названным словом с аффиксом исходного падежа, для субъекта:

Акчага булгач, ***базардан*** *да чыгып алам мин аны* (Аманулла). ‘Если за деньги, я и **на базаре** возьму’.

Солтан ***урманнан*** *озын киштәләр кисеп, атларына бәйләде* (Тагир Набиуллин). ‘Султан **в лесу** нарезал длинные вязанки и привязал к лошади’.

Урманнан *коры чыбык-чабык җыйган өчен генә дә урманчы Василий Маратның балтасын алып калган* (Анас Хасанов). ‘Лес-

ник Василий отобрал у Марата топор только за то, что он собирал **в лесу** валежник'.

Урманнан *ниндидер каенны эзләү диңгез төбеннән тишек чуерташ эзләүгә тиң булмасмы?* (Нур Ахмадиев). 'Искать **в лесу** какую-то березу – разве не то же самое, что искать булыжник на дне моря?'

Данное значение отчасти синонимично основному значению локатива, но если даже глагольный предикат допускает локатив (в татарском языке предикаты в примерах выше чаще используются с аблативом, нежели с локативом), описываемая ситуация предполагает, что приобретённый объект выводится из места, обозначенного словом в исходном падеже.

1.4. Источник

Без класска телевизор алдык, ремонт ясыйбыз, һәр ***баладан*** *2–3 мең сум акча җыябыз* (Малика Басыйр). 'Купили для класса телевизор, делаем ремонт, **с** каждого **ребенка** собираем по 2-3 тысячи рублей'.

Бер ***гектардан*** *уртача 28,8 центнер уңыш чыкты* («Татар-информ»). 'С одного гектара получено 28,8 центнеров урожая'.

В данном случае лексемы с аффиксом локатива могут не иметь значения локализации и относиться к широкому классу различных семантических классов.

1.5. Источник сообщения

С некоторыми глаголами речи и интеллектуальной деятельности:

Тукта, ***әтидән*** *сорыйм әле* (Тази Гиззат). 'Постой, спрошу **у отца**'.

Кичә кич ***радиодан*** *хәбәр иткәннән бирле бөтен авыл халкы малаең турында гына сөйләшә* (Аманулла). 'После того, как вчера сообщили **по радио**, все только говорят о твоем сыне'.

Телевизордан *чыгыш ясады* (Захид Махмуди). 'Выступил **по телевизору'**.

Ниндидер бер татарча ***китаптан*** *укып, «Диңгез сугышы» дигән уен да өйрәндек* (Адлер Тимергалин). 'Прочитав в какой-то книге на татарском языке, разучили игру «Морской бой»'.

Чыгышлардан *аңлашылганча, кайберәүләр бурычлылар белән эшләү өчен махсус хезмәткәр, юрист алган* («Татар-информ»). 'Как стало ясно **из выступлений**, некоторые наняли специальных сотрудников, юристов для работы с долгами'.

Слово с аффиксом аблатива обозначает источник или носи-

тель информации, соответственно, чаще всего он относится к классу лиц или средств передачи информации.

1.6. Материал или блоки для строительства чего-либо

При глаголах созидательной деятельности:

Сандык шикелле кечкенә генә, ***агачтан*** *салынган ак өй үзеннеке* (Айрат Зиятов). ‘Маленький, как сундук, белый дом, построенный **из дерева**, мой’.

Сиксәненче еллар башында шәп ***бүрәнәләрдән*** *Баратынскийның йорт-музеен салып куйдык* (Захид Махмуди). ‘В начале 80-ых **из** добротных **бревен** построили дом-музей Баратынского’.

Ефәктән *капчык тегмиләр,* ***Киндердән*** *капчык була* (Туфан Миннуллин). ‘**Из шелка** мешок не шьют, мешок бывает из холстины’.

В данной группе слово с аффиксом аблатива является исходным материалом для преобразования, для творческой деятельности, соответственно, управляющие предикаты относятся к классу трудовой созидательной деятельности.

1.7. Источник эмоции, стимул:

При глаголах, выражающих страх (*курку* ‘бояться’, *шүрләү* ‘побаиваться, опасаться’, *шөлләү* ‘побаиваться, опасаться’, *өркү* ‘пугаться’ и т. п.):

Бала чакта без ***әтидән*** *курка идек* (Розалина Нуруллина). ‘В детстве мы боялись **отца**’.

Әтидән *байлар дер калтырап тордылар* (Абдулла Алиш). ‘Богатеи тряслись **перед отцом**’.

При глаголах, обозначающих подозрение, опасение *(шикләнү* ‘подозревать’, *шөһбәләнү* ‘подозревать, сомневаться’) слово с аффиксом аблатива обозначает источник подозрения:

Нидән *шикләнәләр?* (Адлер Тимергалин). ‘Чего опасаются?’

Аннан соң, волисполком акчасын урлауда Бакый ***иптәштән*** *шикләнәләр* (Тази Гиззатт). ‘И потом, в краже денег волисполкома подозревают **товарища** Баки’.

Значение источника эмоции (стимула) может быть выражено при отдельных глаголах некоторых других семантических классов, например, при словах, обозначающих зависть, ревность, брезгливость и др.:

Укымышлы кешеләрдән *көнләшә дә, шул ук вакытта аларны сөйми дә иде ул* (Сабир Шарипов). ‘Он завидовал **ученым людям** и одновременно недолюбливал их’.

Ә мин синең янда ***озак утыручылардан*** *да көнләшә идем* (Фанис Яруллин). ‘А я ревновал даже тех, кто долго сидел рядом с тобой’.

Кешедән *җирәнә торган гадәтегез булса, үзегезнекен алып килегез* (Ильдина Ядкарова). ‘Если брезгуете **людьми**, приносите свое’.

Өйрәнгәннәр ***кешедән*** *көлеп яшәргә* (Фанзаман Баттал). ‘Привыкли смеяться **над людьми’**.

Требовать формы исходного падежа могут отдельные глаголы, обозначающие самые разны эмоции и чувства.

1.8. Каузируемый субъект, источник каузируемого действия

При каузативных глаголах: словоформа с аффиксом аблатива может обозначать каузируемый субъект, на который субъект-каузатор оказывает воздействие:

Бер дә яла түгел, ни өчен ***әтидән*** *сиксән потлап икмәк саттырдылар?* (Тази Гиззат). ‘Это не клевета, почему отца заставили продать около восьмидесяти пудов хлеба?’

Аллага шөкер, телчәк Гымрый кызы булсак та, ***кешедән*** *алдатмадык әле, мыскыл иттермәдек* (Радиф Сагди). ‘Слава аллаху, будучи дочерью речистого Гимри, не дали **людям** обмануть себя, не дали насмехаться’.

Менә ул хәзер кайтыр, чишенеп ташлап, мәрмәр бассейннарда йөзәр, ***хезмәтчеләрдән*** *җаны теләгән нәрсәләрне ашарга әзерләтер* (Фаузия Байрамова). ‘Вот сейчас вернется, разденется, поплавает в мраморном бассейне, велит приготовить слугам что душа пожелает’

В этом случае словоформа с аффиксом аблатива обозначает источник (исполнителя) каузируемого другим субъектом действия.

1.9. Прекращение действия

Сочетания имен действий с аффиксом аблатива с глаголом *туктау* ‘остановиться, прекратить’ передают значение ‘прекратить что-либо делать, перестать что-либо делать’:

Халык кайчан ***эчүдән*** *туктар?* («Татар-информ»). ‘Когда люди прекратят пить?’

Машина белән ***йөрүдән туктау*** *авыр булачак* («Безнең гәҗит»). ‘Трудно будеть перестать ездить на машине’.

В данном случае происходит движение от совершаемого действия к его прекращению, которое метафорически представляется

как перемещение из точки совершения действия к точке его отсутствия.

2. Точка отсчёта, ориентир

2.1. Собственно точка отсчёта

В этом случае можно выделить три основных случая.

1. При обозначении расстояния:

Авылдан *ерак түгел иске керәшен зираты да бар (*Гаяз Исхаки). ‘Недалеко **от деревни** есть старое кладбище татар-кряшен’.

Безнең ***коймадан*** *йөз-йөз илле метр чамасы ераклыкта <...> авыл клубы хәтле йорт корып куйганнар иде* (Талгат Галиуллин). ‘На расстоянии ста – ста пятидесяти метров **от** нашего **забора** поставили дом размерами с сельский клуб.’

Здесь используются слова, обозначающие пространственную близость.

2. При обозначении времени:

Берәр атнадан *китәм* (Аманулла). ‘Уеду примерно **через неделю**.’

Оркестр ***дүшәмбедән*** *концертта уйналачак көйләрне кабатлый башлады* (Наил Алан). ‘Оркестр **с понедельника** начал играть произведения, запланированные для концерта’.

Ярминкә ***беренче көннән*** *башлап күп халыкны җәлеп итте* (Ленар Мөхәммәдиев). ‘Ярмарка **с первого дня** привлекла много народу’.

При обозначении времени предикатами могут быть слова самых разных семантических классов, обозначающие действие.

3. Начало интервала, последовательности, ряда, множества (также для слов самых разных семантических классов):

Бүгенге көндә 1,5 ***яшьтән*** *алып 2,5 яшькәчә чаклы булган 800 бала чиратта тора.* («Татар-информ»). ‘На сегодняшний день в очереди стоят 800 детей от 1,5 **года** до 2,5 лет’.

Бүген тәпи атлап киткән ***баладан*** *алып карт әби-бабайларга тикле кесә телефоны белән яхшы таныш* («Татар-информ»). ‘Сегодня с сотовыми телефонами хорошо знакомы все начиная **от детей**, едва начинающих ходить, до старых бабушек и дедушек’.

Аннары фатирларга түләү бәяләренең традицион үсеше 1 ***июльдән*** *дә, 1* ***сентябрьдән*** *дә көтәргә кирәк була.* («Татар-информ»). ‘Традиционный рост коммунальных платежей можно ждать и **с 1 июля,** и **с 1 сентября**’.

Конец интервала (ряда) может быть явным образом обозначен или подразумеваться из контекста, например: по настоящее время.

2.1. Основание для сравнения

Стандартно, относится к качественным прилагательным или наречиям с аффиксом компаратива или без:

Синнән *матуррак, акыллырак, бераклар бар, ә ул гомерлек яры итеп сине сайлаган!* (Хадиджа Шарафетдин). ‘Есть люди красивее, умнее, богаче тебя, а она все-таки выбрала тебя’.

Хактыр, ***мәхәббәттән*** *көчле хис юк* (Гайса Хусаинов). ‘Воистину нет чувства сильнее любви’.

Может использоваться при глаголах с семантическим компонентом ‘количество’, вводящих сравнение:

Хәтеребез ялгышмаса, Татарстан ***Башкорстаннан*** *территориясе буенча да, халык саны буенча да шактый* ***калыша*** *шикелле* (Ильдар Фазлетдинов). ‘Если не изменяет память, Татарстан уступает Башкортостану и по территории, и по населению’.

Чуашстанда мөселманнар сәҗдә кыла ала торган мәчетләрнең саны ***кырыктан артып*** *бара* (Асия Юнусова). ‘В Чувашии число мечетей, где совершают молитвы мусульмане, превышает **сорок’**.

При помощи таких словоформ строятся суждения о сходстве и различии объектов, выявляются количественные и качественные характеристики предметов.

3. Причина

Значение причины может выражаться при словах самых разных семантических классов:

Мин ***кайгыдан*** *эчәм* (Гаяз Исхаки). ‘Я пью **с горя**’.

Шул ***шатлыктан*** *биеп җибәрдем* (Тази Гыйззәт). ‘**От** такой **радости** я заплясал’.

Мәхәббәттән *янган ике йөрәк Зур даланы айкап уздылар* (Алсу Наджми). ‘Два сердца, горящие **от любви**, объехали всю степь’.

1. При словах, выражающих изменение состояния:

Ромео белән Мжульетта сыман ***мәхәббәттән*** *шашынган яшьләр булуы, әлбәттә, зур сөенеч* (Линар Закиров). Большая радость, что есть люди, пьяные **от любви**, словно Ромео и Джульетта.

Кайгысыннан *кибеп саргайды инде бичара Әсхәпҗамал апай!* (Галимджан Ибрагимов). Тетушка Асхапджамал **от горя** пожелтела и высохла!’

2. При словах, указывающих на последствия и ущерб от стихийных и других бедствий:

*Яуган **яңгырдан** су басу республика өчен хас нәрсә түгел* (Резеда Галикаева). ‘Наводнения **вследствие** выпавших **дождей** не новость для республики’.

*Халык **ачлыктан** интегә, хәлләре көннәнкөн начарлана бара* (Радиф Сагди). ‘Народ страдает **от голода**, ситуация ухудшается с каждым днем’.

*Уфада **су басудан** каза күрүчеләргә гуманитар ярдәм әыю пунктлары атна буе актив эшләде* («Татар-информ»). ‘В Уфе в течение недели активно работали пункты гуманитарной помощи для тех, кто пострадал о**т наводнения**.’

4. Косвенный объект

Объектное значение реализуется при предикатах различных семантических классов, отдельно выделим несколько самых частотных случая.

1. При предикатах с семантическим компонентом ‘пользоваться’:

*Вазыйфаи **вәкаләтләрдә**н файдаланып кылынган әлеге җинаять өчен реаль колония янаганын аңламый түгел сержант* (Наиль Вахитов). ‘Сержант понимает, что за это преступление, совершенное пользуясь служебным **положением**, грозит колония’.

***Социаль челтәрдән** кулланучылар әлеге видеоязманың Баймак шәһәренең 4 нче урта мәктәбендә төшерелүен белдерә* (Ильдар Фазлетдинов). ‘Пользователи **социальных сетей** сообщают, что видеозапись сделана в школе № 4 города Баймаково’.

*Авыл кешесе тир түгеп үстергән **ашлыктан**, тапкан **байлыктан** хуҗалар, алыпсатарлар, түрәләр рәхәт күрә* («Татар-информ»). ‘**От урожая**, выращенного потом сельчан, **от богатства** будут получать удовольствие хозяева, торгаши, сановники’.

*Сирәк кенә булса да Галиябануга **телефоннан** да шылтыратканалады* (Захид Махмуди). ‘Хотя бы изредка звонил Галиябану **по телефону**’.

2. При предикатах со значениями ‘отделить’, ‘оставить без’, ‘избавить’, ‘лишить, ‘ ‘удалить’ и т.п. (каузативные и некаузативные употребления). Часто используются следующие глаголы: *чистатрту* ‘очищать’, *арындыру* ‘очищать’, *колак кагу* ‘лишиться’, *котылу* ‘избавиться, спастись’, *азат итү* ‘освободить’, *баш тарту* ‘отказаться’ и др.

*Кызны тынычландыру, авыр **уйларыннан** арындыру бик читен булды* (Анас Хасанов). ‘Успокоить девушку, избавить **от** тяжелых **мыслей** было очень трудно’.

*Хәзер күбесе артык **җирдән** баш тарта* (Ягсуф Шафиков). ‘Сейчас многие отказываются **от** лишней **земли**’.

*Өйләнәчәк кеше буларак ул җитди ир булып кылана, әмма **балалыгыннан** котыла алмый* (Ильдар Юзеев). ‘Нацеленный на брак, он притворяется серьезным человеком, но **от ребячества** избавиться не может’.

*Аның хезмәт урыны, эшләү шартлары аны аерым кешеләргә **антлар итүдән** азат итә иде* (Адель Кутуй). ‘Его должность, условия работы освобождали его **от присяги** отдельным людям’.

Объектное значение может выражаться также при некоторых глаголах других семантических классов.

5. Элемент целого, множества

Здесь мы выделяем два разных случая: 1) совокупность и ее элемент обозначаются разными словами, 2) совокупность и ее элемент обозначаются одной словоформой.

1. Совокупность названа существительным, а его элемент – словом с аффиксом аблатива. Данное значение наиболее часто реализуется при глаголе бытия *тору* ‘быть’:

*Кечкенә, дүрт ятим **баладан** торган төркем Актүбә тимер юл стансасына таба атлады* (Зәки Зәйнуллин). ‘Маленькая группа, состоящая из четырех детей, зашагала к железнодорожной станции Актюбинска’.

*Русиядә 4 **китаптан** торган, рус телендә ислам энциклопедиясе нәшер ителәчәк* (Назифа Каримова). ‘В России на русском языке будет издана исламская энциклопедия, состоящая **из** четырех **книг**’.

Существенно, что элементы множества не теряют своей самостоятельности и не испытывают трансформаций, в отличие от значения «Материал».

2. Совокупность называется существительным или признаковым словом с аффиксом множественного числа, к которому присоединён аффикс аблатива (с выделительным значением ‘из числа ...’). В данном случае аффикс множественного числа формирует совокупность, а аффикс аблатива выделяет ее элемент:

*Азнакай, Түбән Кама, Сарман районнары әнә **шундыйлардан*** (**«**Татар-информ»). ‘Азнакаевский, Нижнекамский, Сармановский районы – **из числа подобных**’.

*Ник дигәндә, ул табигый бәла-казалар күзлегеннән караганда җир шарының иң тыныч географик **төбәкләреннән** санала*

(Захид Махмуди). ‘Это с точки зрения стихийных бедствий считается одним **из** самых безопасных **регионов**’.

Әле кайчан гына тугыз катлы бу йорт шәһәрдә иң биек һәм иң ***матурлардан*** *берсе иде* (Рөстәм Галиуллин). ‘Еще недавно этот девятиэтажный дом был одним **из** самых высоких и **красивых**’.

6. Определительное значение

В этой группе самым частотным и характерным является обозначение цены за единицу измерения (‘по цене *x*’), словоформы с аффиксом аблатива используются с глаголами, обозначающими реализацию продуктов (товаров) и действия за вознаграждение: *сату* ‘продавать’, *сатып алу* ‘закупать’, *алу* ‘принимать’, *җибәрү* ‘отправлять’, *яллау* ‘нанимать’ и т.п.

Шикәр комының килосы уртача ***29 сум 78 тиеннән*** *сатылса, йомырканың дистәсе –* ***38 сум 87 тиеннән*** *тәкъдим ителә* («Татар-информ»). ‘Если килограмм сахарного песка прдается по **29 рублей 78 копеек**, то десяток яиц предлагается по **38 рублей 87 копеек**’.

Ә Самарада барлык марка бензин иң ***түбән бәядән*** *сатыла* («Татар-информ»). ‘А в Самаре все марки бензина продаются **по** самой **низкой цене**’.

Арышның тоннасы ***6,2 мең сумнан*** *сатып алыначак* («Татар-информ»). ‘Рожь будет закупаться по **6,2 тысяч рублей** за тонну’.

Нефтьнең тоннасын 8 сумга сата идек, ә «Дуслык» үткәргече аша башка илләргә ***15 сумнан*** *җибәрдек* (Ягсуф Шафиков). ‘Мы продавали нефть по 8 рублей за тонну, а по нефтепроводу «Дружба» отправляем в другие страны по **15 рублей**’.

Предполагается, что определительное значением монет быть использовано не только при обозначении цены, но и в ситуациях с другими предикатами.

Подведем некоторые предварительные итоги.

Инвариантное значение татарского аблатива – исходная точка (исходный пункт) – получает дальнейшее семантическое наполнение за счет значения (с учетом тематических и акциональных семантических компонентов) управляющих слов. Конкретная реализация значения падежа по существу определяется предикатом и в меньшей степени зависит от семантического класса лексемы; верней, для слов разных семантических классов представлен широкий набор семантических ролей, в зависимости от управляющего слова:

коймадан сикерде ‘прыгнул с забора’,

коймадан йөрмә ‘не ходи через забор’,
коймадан ясады ‘сделал из забора’,
коймадан башлап сана ‘считай начиная с забора’,
коймадан ике метр читтәрәк ‘на два метра дальше от забора’,
коймадан биегрәк ‘выше забора’,
коймадан шикләнәм, аумасмы ‘опасаюсь забора, как бы не упал’,
ихата ике коймадан тора ‘ограждение состоит из двух заборов’ и т.п.

То есть конкретная семантическая роль лексемы определяется семантикой управляющего слова (предиката) и отражает общие свойства участников определенных типов ситуаций. Безусловно, для слов со значением места в большей степени естественно выражать значение исходной точки, для слов с абстрактных существительных – значение причины, но в действительности все слова реализует широкий спектр значений.

Во многих случаях трудно дифференцировать отдельные значения, например, исходную точку и источник, источник и причину, объект и причину и т.п.

1.4. Заключение

В данной статье представлены предварительные результаты проведенного исследования. В качестве основных мы выделяем следующие значения:

- исходная точка (исходный пункт), источник (базовое, протитипическое значение);
- точка отсчета, ориентир;
- причина;
- объект;
- элементы целого, множества;
- определительное значение.

Данные значения не являются случайным набором и имеют определенные точки пересечения, выражаемые, например, семантическими свойствами предикатов.

Тема требует дальнейшего исследования во многих аспектах, в частности: семантический класс лексемы и возможные значения форм аблатива, соотношение акциональных и тематических классов предикатов, управляющих словоформами с аффиксом аблатива. Предполагается, что с учетом новых данных и их анализа клас-

сификация значений аблатива будет выглядеть более стройной и последовательной.

ЛИТЕРАТУРА

Закиев, М. З. (ред.) (1993). *Татарская грамматика*. Т.2. Морфология. Казань: Татар. кн.изд-во.

Зәкиев, М. З. (ред.) (2016). *Татар грамматикасы*. Т. 2. Казан: ТӘhСИ.

Ибрагимов Т. И, Сайхунов М. Р. (2014). Письменный корпус татарского языка: структурные и функциональные характеристики. *Актуальные проблемы диалектологии языков народов России: Материалы XIV Всероссийской научной конференции* (Уфа, 20–22 ноября 2014 г.), Уфа, С. 261–263.

Кибрик, А. Е. (2005). *Константы и переменные языка*. – СПб.: Алетейя,

Невзорова О. А., Мухамедшин Д. Р., Билалов Р. Р. (2015). Корпус-менеджер для тюркских языков: основная функциональность. *Труды международной конференции «Корпусная лингвистика – 2015»*. СПб.: С.-Петербургский гос. университет, филологический факультет, С. 344–350.

Плунгян, В. А. (2003). *Общая морфология: Введение в проблематику*. М.: Едиториал УРСС.

Хисамова, Ф. М. (2015). *Татар теле морфологиясе*. Казан: Татарстан китап нәшрияты.

УДК 655.28.022.1

BIBLIOGRAPHIC FILE OF STUDIES ON THE CRIMEA, PUBLISHED IN TURKEY

Ranetta Gafarova
Ismail Gasprinsky Eurasian Institute of Development

For Turkey, the priority has always been and remains – What has been done for the Crimea and the Crimean Tatars? Various historical and cultural facts, linked by certain actions, have been preserved in each region of Turkey. Therefore, the development of some actions will directly affect the life of Turkey. It should be noted that any historical events that take place in Turkey are reflected in the regions with which it is historically and politically connected.

For some time between Turkey and Crimea, economic, cultural, political and other ties have weakened a little. Scientific research works that will assess the ongoing events between the states are essential to improve the objective attitude of Turkey to the problems of the Crimea. Therefore, the number of scientific research works has recently increased.

The paper presents a bibliographic file of studies on the Crimea, published in Turkey, in the format of books, monographs and materials. In our opinion, one of the most important problems faced by researchers is determining whether the carrying out studies in a particular area had not been previously conducted. The bibliographic list includes materials from 1207 to the present. The range of sources is diverse – these are historical, sociological, literary, linguistic, theological, economic, anthropological and art history sources.

The work presents an ordered scheme of sources with indication of the author, year and place of publication, volume and type of research. The study also presents Turkic scientific works about the peoples living in the Crimea: Krymchaks, Urum.

Keywords: Crimea, Turkey, history, Crimean Tatars, sources, bibliography.

БИБЛИОГРАФИЧЕСКАЯ КАРТОТЕКА ИССЛЕДОВАНИЙ О КРЫМЕ, ИЗДАННАЯ В ТУРЦИИ

Р. И. Гафарова
Евразийский институт развития им. И. Гаспринского, РФ, Крым, Симферополь

Для Турции всегда приоритетным направлением было и остается: Что сделано для Крыма и крымских татар.

В каждом регионе Турции сохранились различные исторические и культурные факты, связанные между собой определенными действия-

ми. И поэтому развитие каких-то действий будет напрямую влиять на жизнь Турции.

Следует отметить, что любые исторические события, которые происходят в Турции, отражаются в тех регионах, с которыми она исторически и политически связана.

Какое-то время между Турцией и Крымом, экономические, культурные, политические и другие связи немного ослабли. Для оздоровления объективного отношения Турции к проблемам Крыма, доказательными являются научно-исследовательские работы о Крыме и крымских татарах, которые дадут оценку происходящих событий между государствами. Так в последнее время увеличилось число научно-исследовательских работ.

В работе представлена библиографическая картотека исследований о Крыме, изданная в Турции, в формате книг, монографий, материалов. На наш взгляд, одной из наиболее важных проблем, с которыми сталкивались исследователи, является определение того, исследования, проведенные в той или иной области, материал, над которым они работали, ранее не был изучен.

Библиографический перечень включает материалы с 1207 года по настоящее время. Диапазон источников разнообразен – это исторические, социологические, литературные, лингвистические, теологические, экономические, антропологические и искусствоведческие источники.

В работе представлена упорядоченная схема источников с указанием данных автора, года и места издания, объем и типа исследования. Также в исследовании представлены тюркологические научные работы о народах, живущих в Крыму: крымчаки, урумы.

Ключевые слова: Крым, Турция, история, крымские татары, источники, библиография.

Hızla değişen dünyada zamanla yarışan araştırmacılar için bilgi kaynağına kısa sürede ulaşabilmek çok önemlidir. Bunun için bir konuyu derinlemesine araştırmak isteyen kişinin doğru yol izlemesi gerekir. Araştırılması gereken konuyla ilgili yapılacak ilk okumalara öncelikli olarak konunun bilinen temel kaynaklarından başlayıp, bu kaynakların sonunda yer alan kaynakçalardan hareketle okumayı derinleştirmek bilinen bir yoldur. Fakat birkaç kaynağın arkasındaki kaynakça, araştırmacıya ancak sınırlı bir ilerleme sağlayabilir. Bu yüzden belli alanlarda yazılmış kitap ve dergiler taranarak derlenen kaynakçalar, araştırmacıya büyük kolaylıklar sunmaktadır. Bazı sorunları bulunsa da, uzun süreli periyodiklerin belli aralıklarla geçmiş sayılarına yönelik hazırladıkları bibliyografyalar ve belli alanlarda yapılan araştırmaların bibliyografyaları bugün azımsanamayacak kadar çoktur. (Y.Karayalçın, 1954: 28)

Kırım'ın ve Kırım Türklüğü 'nün önemine ve Türkiye'nin neler yapıp, neler yapmadığına ve neler yapması gerektiğine geçmeden önce bu mevcut hali açıklamak istiyorum.

Bu yeni düzenin ve mücadelenin olduğu her bölgede Türkiye'nin tarihsel ya da kültürel ama mutlaka bir bağı var. Dolayısıyla yaşanılan her gelişme Türkiye'yi doğrudan etkileyecektir.

Dikkat edilmesi gereken asıl mesele cereyan eden hadiselerin Türkiye'nin tarihsel, kültürel ve siyasi bağının olduğu bölgelerde yaşanıyor olmasıdır.

Türkiye'nin fiili olmayan sınırları ya da bir başka deyişle milli güvenlik alanı Batı Trakya, Kırım, Kafkaslar, Azerbaycan, Musul, Kerkük, Halep, Kıbrıs ve Adaları içerisine alan bölgedir. Bu bölgedeki her denge Türkiye'nin kaderini belirler. Yine bu alanlarda yaşanılan her kazanç ya da kayıp Türkiye'nin geleceği açısından hayati derecede önemlidir. Bu coğrafyada vukuu bulan her hadisede Türkiye'nin etkin rol alma ve dengeler şekillenirken akılcı davranıp, şekillendirilmek istenilen her dengede mutlaka ağırlığının bulunması mecburidir.

Gündemden Kırım söylemleri düşmüyor. Türkiye, Kırım ile ilgili çeşitli karar alırken Kırım'ın Türkiye için önemini, Kırım'ın nerede olduğunu tekrar hatırlamakta fayda var. Kırım'ın konumu tartışılırken, tarihte Kırım'ın Türkiye için önemi, Kırım'ın diğer ülkeler açısından önemi ve Kırım'ın coğrafi olarak nerede olduğu ile ilgili pek çok ifade yer alıyor. İşte tarihsel gelişmelerle Kırım ve Kırım Türkleri.

Çalkantılı coğrafyalardan bir tanesi de Kırım! Bir taraftan Rus hükümetinin demeçleri, bir taraftan Türkiye Cumhuriyeti'nin demeçleri, vatandaşlar tarafından Kırım'ın Türkler açısından önemini ve Kırım'ın genel önemini sormasına neden oluyor. Geçmişten bugüne Kırım'ın önemi, her iki ülke açısından da önemli bir konu haline geldi ve yüzyıllardır da gündemden düşmüyor. Peki, Kırım'ın önemi nedir? Kırım Türkler için neden ve ne şekilde önemli? Kırım, coğrafi olarak nerede yer alıyor? İşte tarihsel gelişmeleri ve tüm detaylarıyla Kırım.

Fatih Sultan Mehmet'in Kırım'ı Osmanlı idaresine alması, Karadeniz ve kuzeyindeki bölgelere yönelik stratejik bir yaklaşımın sonucuydu. Karadeniz'in Türk gölü haline gelmesinin temelinde Fatih'in bu stratejisi ve Kırım'ın alınması vardı. Kırım Hanlığı'nın kurulması ile İstanbul'un fethi birbirine çok yakındır.

Kırım Hanlığı 1441'de kurulmuş, 1453'te İstanbul'u fethi, ona Osmanlı'nın gücünü göstermişti. 1466'da Kırım Hanlığı'nın kurucusu Hacı Giray'ın vefatı üzerine oğulları arasında taht kavgası başlamış, bazı Tatar büyükleri bu konuda Osmanlı'dan yardım istemişti. Fatih Sultan Mehmet, bu fırsatı değerlendirerek Gedik Ahmet Paşa komutasında Osmanlı donanmasını bölgeye göndermiş, Kefe, Sudak, Balıklağo, Azak, Taman ve Mankub kaleleri fethedilmiş, Kırım'da Mengli Giray iktidara getirilmişti.

Dolayısıyla 1783'e kadar sürecek güçlü bir ittifak oluşmuştu. Mengli Giray, Fatih Sultan Mehmet'e «Senin düşmanın benim düşmanım, senin dostun benim dostumdur» güvencesi vermişti.

Kırım Hanlığı Osmanlı'nın yönetimindeydi ama geniş özerkliğe sahipti. Kırım süvarileri savaşlarda Osmanlı ordusunun önemli güçlerinden birini oluşturuyordu.

Kırım Hanları İstanbul'u ziyaret ettiklerinde görkemli törenlerle karşılanır, Padişah «Hoş geldin Han kardeş» derdi. Bu yüksek protokol sadece Kırım Hanı'na uygulanırdı. Kırım ordusu Osmanlı ordusuna katıldığında top ve tüfek atışlarıyla karşılanır, Rumeli Beylerbeyi karşılamaya çıkardı. Kırımlıların tüm cephelerde Osmanlı ordusuna katkılar yaptıkları genel kabul görür. Kırım 1475'te Fatih Sultan Mehmet zamanında Osmanlı topraklarına katılmıştır. 1475'de tahtını Osmanlı ordusunun Kırım yarımadası sahilini tümden fethedip, 1478'de tahtını Osmanlı himayesi ile ele geçirebilen Menli Giray Hanın Osmanlı vasalı olması sağlanabilmiştir.

Osmanlılar yönetime el koyunca Mengli Giray, «han» ilan edildi. Kırım kuvvetleri, bir Osmanlı savaşına ilk defa, Sultan II. Bayezid'in, 1484'teki Akkirman Seferi'nde katıldılar. 300 yıl Osmanlı yönetiminde kalan Kırım; Osmanlı diplomasi geleneğine göre Kırım'daki Giray Hanedanı üyeleri, Osmanlı Hanedanı mensuplarının ardından Osmanlı İmparatorluğu hiyerarşisinde ikinci sırada yer alırdı Kırım Hanlığı dış ilişkilerinde bağımsız bir devlet gibi davranabiliyordu. 1532 yılından sonra Kanuni tarafından Kırım'ın hükümdar ailesi Giray Han'lardan bir ya da birkaç kişi İstanbul'da ve hemen yakınındaki mülkleri olan, avcılığıyla ünlü Çatalca'da yaşadı. (Halil İnancık, 2017)

Rusya ile savaşlarda rolleri önemliydi. 1571'de Devlet Giray'ın ordusu Moskova'yı yakıp yıkmıştı. Ruslar da 1736'da Bahçesaray'ı yağmalamıştı. Rus Çarı Petro, iki kez Kırım'ı almayı denemişti. 1696'da Azak Kalesi'nin kaybedilmesi Rusya'nın yükselen gücünün habercisiydi. Sonuçta 1783'te Kırım kaybedildi, Kırım Hanlığı tarihe karıştı. Osmanlı ordusu Kırım Tatarlarının katkısını kaybetti. Rusya, Karadeniz'in kuzeyine hakim oldu. Tatarların çilesi başladı. Kırım tarihi, acının, sürgünün, hüznün, baskının, zulmün tarihidir, hem de aynı zamanda mücadelenin tarihidir. Kırım Tatarların tarihini 6 evrelerden oluşuyor: 1. Kırım Hanlık Dönemi (1441–1783), 2. Rus Hâkimiyeti: Kırım ilhakı (1783–1917), 3. Rus Devrimi ve Erken Bolşevik Yönetimi Sırasında Kırım: Kırım Tatar Cumhuriyeti (1917–1922), 4. 1944 yıl- Sürgünlük (1944–1987), 5. Vatana Dönüşü (1987 – bugüne kadar), 6. Bugünkü Kırım Tatar halkının mücadelesi.

Yirmi yıl önce 1990 tarihinde Kırım'ına uzun süre sürgünlükte yaşamış yerli halkı Kırım Türkleri dönüşünden sonra Türkiye'de Kırım dâhil olmak üzere bütün Sovyet Birliğinde Türk Cumhuriyetlerine yönelik hem kişisel hem bilimsel ilgi artmış ve artarak devam etmektedir. Sovyet Birliğinin dağılması ve kapalı demir perdeleri açılması bunun temelinde son yıllarda Türkiye'nin iktisadi ve siyasi anlamda gösterdiği büyük ilerleme yatmaktadır. Ancak Türkiye'nin, Türk Cumhuriyetleriyle olan ilişkilerinin daha sağlıklı devam etmesi dil, tarih ve kültürel konularda bilimsel araştırmalara ağırlık verilmesine, Türk Cumhuriyetleri konusunda uzmanların yetiştirilmesine bağılıdır. Bu da ümit ederiz ki, Türkiye'deki akademik çevrelerin, Kırım dâhil olmak üzere Eski Sovyet Birliğinin Türk Cumhuriyetlerinin ve topluluklarının kültürüne, tarihine, toplumsal, ekonomik ve siyasi gelişmelerine olan ilgisini daha da artıracaktır. Ulu önder Atatürk'ün, Türkiye Cumhuriyetinin kurulduğu ilk yıllarda söylediği «Bu gün Sovyetler Birliği dostumuzdur, komşumuzdur, müttefikimizdir. Bu dostluğa ihtiyacımız vardır. Fakat yarın ne olacağını kimse bugünden kestiremez. Tıpkı Osmanlı gibi, tıpkı Avusturya Macaristan İmparatorluğu gibi parçalanabilir, ufalanabilir. Bu gün Rusya'nın elinde sımsıkı tuttuğu milletler avuçlarından kaçabilirler. Dünya yeni dengeye ulaşabilir, işte o zaman Türkiye ne yapacağını bilmelidir. Bizim, bu dostumuzun idaresinde dili bir, inancı bir, özü bir kardeşlerimiz vardır. Onlara sahip çıkmaya hazır olmalıyız. Hazır olmak yalnız o günü susup beklemek değildir. Hazırlanmak lazımdır. Milletler buna nasıl hazırlanır? Manevi köprüleri sağlam tutarak... Dil bir köprüdür. İnanç bir köprüdür. Tarih bir köprüdür. Köklerimize inmeli ve olayların böldüğü tarihîmiz içinde bütünleşmeliyiz. Onların bize yaklaşmasını bekleyemeyiz, bizim onlara yaklaşmamız gereklidir » (İ. Bozdağ, 1975) .

Sovyet Birliğinin dağılmasının öncesinde Kırım'la ilgili konularda Türkiye'de neler yapıldığının, binlerce yıllık geçmişi olan Kırım kültürü üzerine akademik alanda Türkiye'de yapılan araştırmaların sayısı ve bibliyografya tespitine yönelik bir çalışma yapmaya karar verdik. Sovyet Birliği'nin dağılması ve Kırıma Kırım Türklerini dönüşü, Milli davası, son dört yılında tekrar Rusya Hükümetine Kırım yarımadası geçmesi bu süreçte «Türk ve akraba topluluklarla ilişkilerin yoğunlaşması, her açıdan konuya ilgiyi arttırmış, bu ilgi artışı yapılan yayın sayısına da yansımıştır» (F. Solak, 2007:7)

Yukardaki gösteren Kırım ve Kırım Türklerin tarihinin 6 tane evrenlerini Türkiye'de yaklaşık 700'den fazla Türkçe kitap, Türkiye'de yapılan yüksek lisans, doktora ve doçentlik tezleri, bilimsel makale,

analitik inceleme ve edebi eserler çalışma kılavuzlarının bibliyografyası verilmiştir. Bu çalışmalar Türkiye Cumhuriyetine aittir.

Araştırmamızın amacı, yeterlik belgelerini yazarken Türkiye'de yayınlayan Kırım hakkında bilgi ve bibliyografik yardım sağlamaktır; içerik analizine dayanan farklı türde konuşma ilişkileri söyleminin çalışması için verimli Kırım alanındaki adımları atan araştırmacıların bilimsel amacını yönlendirmek.

Çalışmamızda Türkçe, Rusça, İngilizce, Almanca, Lehçe ile yazılan eserleri içeriyor. Koleksiyon, Türkiye'de Kırımın gelişmesindeki ana eğilimleri ve kısmen Kırım yarımadası Türkiye'deki Kırım çalışmalarının ana eğilimlerini anlatan kaynaklarını sunuyor. Araştırma malzemeleri Türkiye'de Milli Kütüphanesi, İstanbul Üniversitesi Kütüphanesi, İSAM kütüphanesi kataloglarına ve resmi elektron makale kataloğu dayanmaktadır.

Bibliyografik indeks tematik olarak 8 bölüm halinde yapılandırılmıştır, her biri bölümlere ve alt başlıklar bölümlerine ayrılmıştır. İndeksin her bölümündeki malzemeler kronik bir yapısına sahiptir. Bazı eserlere kısa özet açıklamaları ve tablolar eşlik eder.

Aşağıdaki tablo ve grafikte görüldü gibi, Türkiye'de Kırım konulu yayınlar ilk dönemlerden günümüze giderek artan bir seyir izlemiş, en az kaynaklar 1207 y.-1921 yıllara kadar 27 toplam araştırma sayısı, 22-42'cı yıllarda 45, 40-60'cı yıllarda 45, 60-80 yıllarda 89, 1985 yıllından bugüne kadar 544 rakamına ulaşmıştır.

Tablo ve grafik yayınların dönemlere göre dağılımı

	1207-1335	1856-1900	1901-1921	1922-1942	1943-1963	1964-1984	1985-2005	2006-2017	Toplam
Tezler	–	–	–	–	–	–	59	119	**178**
Kitaplar	16	8	3	23	16	21	91	90	**268**
Makaleler	–	–	–	4	29	68	80	106	**287**
Toplam	**16**	**8**	**3**	**27**	**45**	**89**	**230**	**315**	***733***

Kırım Bibliyografyası konulu içerik ve odak noktalarından hareketle değerlendirdiğimizde 12 alt başlık altında tasnif etmemiz mümkündür:

1. Kırım antropolojisi
2. Kırım Savaşı (1853–1856 yıllar)
3. Kırım Hanlığı

4. Kırım'dan göç konusu
5. Kırım Türklerin sürgünü (1944 yıl)
6. Kırım Türklerin Kırıma dönüşü (1987–2017 yıllar)
7. Kırım Türklerin Milli Mücadelesi
8. Kırımın edebiyatı ve bibliyografik.
9. Kırım Türklerin ve Kırım'da yaşayan halkların dili
10. Kırım Kültür-İdeoloji (sanat ve müzik)
11. Kırım edebi eserlerde
12. Diğer

1. İncelediğimiz kaynaklar «Kırım antropolojisi» konulu 10 kitap tane tespit edilmiştir. Bu kaynaklarda, genel olarak Kırımın ve Kırım'da yaşayan halklar tarihi. Onların arasında, yer alan Martin Bronevskiy; çeviren: Kemal Ortaylı «Kırım» (1970), Prof. Dr. Mehmet Maksudoğlu «Kırım Türkleri» (2009), Alan Fischer; çeviren: Eşref B. Özbilen. «Kırım Tatarları» (2009), Kemal Çapraz «Kuzeydeki Yavru Vatan Kırım» (2016) kitapları ve bş., Kırım'ı geçmişten bugüne değerlendirmeleriyle farklı boyutlarıyla ele alan eserler. İki Yüksek Lisans tezi Oğuz Timur «Eski Polatlı Köyü'ne yerleşen Kırım Türklerinde gelenek ve adetler» (2009), Gözde Mirza «Toplumsal cinsiyet ve milliyetçilik bağlamında İstanbul'da yaşayan kırım Türk-Tatar kadınları» (2007) Türkiye'de yaşayan Kırım kökenli göçlerin antropolojisini açıklıyorlar. 28 Makalelerin arasında en çok sayısın olanlardan Alan Fişer 12 bildiri, Prof. Dr. Nadir Devlet – 8, Edige Kırımal – 5, Cafer Seydamet – 3 derleme metinlerinde anlatıcıların icrasını etkileyen unsurlar üzerinde de durulmuştur.

2. Bildirimizde «Kırım Savaşı (1853–1856 yıllar) » inceleyen çok sayıda yaklaşık 70 toplam kitap, tezler ve makaleler ayrı bir başlık altında değerlendirilmiştir. Bunun sebebi, araştırmamıza temel teşkil eden bildirilerde öne sürülen yeni araştırma tekniklerinin Kırım tarihi ve orada yaşayan halkların hakkında çalışmaları açısından önemidir. «Kırım savaşı, Osmanlı Devletinin toprak bütünlüğünün korunması isteğinden çok, Avrupa'ya özgü düşüncelerle yürütüldü ve önemli olan Avrupa'nın siyasal statüsüydü. İngiltere için önem taşıyan Avrupa'daki güç dengesinin korunmasıydı ve bunun için savaştı». (İA, 4 cilt, 145). Kırım Savaşı çoğunlukla Doğu Meselesindeki çarpıcı olaylardan biri olarak görülür; hem mağlup Rusya'nın hem de savaşçı olmayan Avusturya'nın çapını küçülterek Avrupa'daki güç dengesini değiştiren kesin bir mücadele olarak öne çıkmıştır. Birkaç tanesinin açıklaması yer almaktadır: Orlando Figes ; çeviren: Nurattin Elhüseyni. «Kırım : son haçlı seferi (Kırım Savaşı, 1853-1856) (2012), Alan Palmer «Kırım

savaşı ve modern Avrupa'nın doğuşu» (2014), Candan Badem «Kırım Savaşı ve Osmanlılar» (2017) kitapları, doktora tezi Budak Mustafa «1853–1856 Kırım Savaşı'nda Kafkas Cephesi» (1993), yüksek lisans tezleri: Mehmet Rezan Ekinci «Osmanlı-Rus ilişkileri çerçevesinde Kırım Savaşı (1853–1856)» (2005), Fatih Akyüz «Kırım savaşının lojistiğinde İstanbul'un yeri» (2006) ve bş., makaleler Nilüfer Bayatlı « Rusya'nın Kırım Savaşı'ndan Aldığı Dersler» (1992), Nejat Birinci «1853–1856 Kırım Savaşı'nı Anlatan bir Eser: Manzume-i Sevastopol» (1981–1982) ve bş. Ayrıca bir bilimsel çalışması Serap Torun «Kırım Savaşı'nda hasta bakımı ve hemşirelik» (2008) Deontoloji ve Tıp Tarihi alanında yapılmış.

3. Bildirimizde «Kırım Hanlığı» konusu geniş bir coğrafyaya yayılmış olmasının yanı sıra, pek çok anlatı türünü de etkilemiş, müstakil anlatı türlerinin ortaya çıkmasına olanak sağlamıştır. Kitap – 32, tezler – 36, makaleler – 25 adet bulunmaktadır. Joseph Von Hammer : çev: Seyfi Say « Kırım Hanlığı Tarihi» (2013) kitabı zengin bir kaynakçaya sahip olan eserde yazar, Osmanlıca birçok el yazma eser, resmi yazı, kanunname ve tezkirelerin yanı sıra Osmanlı Devleti tarihiyle ilgili birçok eserden de yararlanmıştır. Bu yönüyle Osmanlı tarihine meraklı olanları da ilgilendiriyor. Özellikle Osmanlı Devleti ile Kırım Hanlığı ilişkilerini ele almış olması kitaba bir özellik katıyor. Ayrıca sırada belli tarihçimizin Prof. Dr. Halil İnancık çalışmaları yer almaktadır. Son çıkan kitabı «Kırım Hanlığı Tarihi Üzerinde Araştırmalar (1441–1700), (2017) Halil İnalcık akademik hayatı boyunca Kırım üzerine yaptığı araştırmaları gözden geçirerek derlediği bu kitapta, hanlığın ileride yazılacak bir genel tarihinin omurgasını kuruyor. Orta Avrupa'daki Habsburg-Osmanlı, Doğu Avrupa'daki Leh-Osmanlı-Rus ve Kuzey'deki Rus-Osmanlı rekabetinin görünenden çok daha karmaşık, değişken ve çok aktörlü olduğunu gözler önüne sererken, İstanbul Bahçesaray ittifakının bu bölgelerdeki kilit niteliğini tarihî olaylar üzerinden anlatıyor. Kırım tarihine dair kısıtlı kaynakları ve araştırmaları eleştirel yaklaşımla okuyarak tarih yazmanın yollarını gösteriyor.

4. Bölüm «Kırım'dan göç konusu» 13 adet kaynaktan oluşuyor. İki kitap Ethem Feyzi Gözaydın «Kırım: Kırım Türklerinin Yerleşme ve Göçmeleri» (1948) ve Hakan Kırımlı «Türkiye'deki Kırım Tatar ve Nogay köy yerleşimleri» (2012), doktora tezi son kitap şeklinde yayınlanan Süleyman Erkan «Kırım «Kafkasya ve Doğu Anadolu göçleri (1878–1908)» (1993), yüksek lisans tezleri Ramazan Göktepe «Kırım Savaşı sonrası Osmanlı basınında Kafkasya ve Kırım göçleri» (2007), Berat YILDIZ «Rusya imparatorluğundan Osmanlı İmparatorluğu'na

göç: Yeni arşiv malzemeleri ışığında bir analiz» (2006) ve diğer. 1783 öncesinde de Kırım'dan Osmanlı topraklarına pek bilinmese de, azımsanmayacak boyutlarda göçler olmuştur. Kırım Tatar halkı bu tarihten itibaren Kırım'dan Osmanlı topraklarına doğru dalgalar halinde başlayan Kırım Tatar göçü 1920'lere kadar tek bir yıl bile durmadan devam etmiş, hatta bazı kesintilerle günümüze kadar sürmüştür. 1783–1922 yılları arasında Osmanlı ülkelerine göç eden Kırım Tatarlarının sayısı en az 1.800.000 idi. Göçlerin aslî sebebi hiç şüphesiz siyasîdir: Göç eden unsur yani Kırım Tatarları yerine göre canlarını, mallarını veya kimliklerini Rusya idaresinin doğrudan tehdidi altında hissettikleri için vatanlarını terk etmek zorunda olduklarını düşünmüşlerdir. Diğer taraftan, Rusya da onların Kırım'dan uzaklaşmalarını siyaseten yararlı görmüştür. Göçlerin ortaya çıkmasında büyük önem taşıyan dinî baskılar da nihayetinde bu siyasetin parçalarıydı. 1860-1861 göç dalgasından sonra da 1874, 1890 ve 1902'de yeni göç dalgaları olduysa da bunların sayıları yüz binlerle değil, on binlerle ifade olunuyordu (Kırımlı, 1996, s. 11–17).

5. İncelediğimiz kaynaklar «Kırım Türklerin sürgünü (1944 yıl)» konusuna toplam 21 adet kitap, tez ve makaleler ait. 1944 senesinde Kırım Türkleri yerlerinden, yurtlarından kazıyarak, hayvan vagonlarında Sibiryalara, Rusya içlerine sürmüştür. Stalin' in acımasız emriyle, bir günde binlerce yıllık vatanlarını – geride her şeylerini bırakarak – kundaktaki bebeğinden yaşlısına kadar, bilinmeyen bir yöne götürülen Kırım Türk halkının çileli yolculuğunu, kayıplarını, sürgünde karşılaştıkları zorlukları. Bu bölümde araştırmalar, Sovyetler Birliği döneminde Kırım Türklerinin maruz kaldığı kitlesel sürgün hareketini işlemektedir. Necip Hablemitoğlu'nun iki kitabı «Türksüz Kırım Yüz Binlerin Sürgünü» (1974), «Yüzbinlerin sürgünü Kırım'da Türk soykırımı» (2004), Kemal Özcan «Sovyet belgelerinde Kırım Dramı» (2007), «Vatana dönüş: Kırım Türkeri'nin sürgünü ve milli mücadele hareketi (1944–1991)» (2002), Kemal Çapraz «Sürgünde yeşeren vatan Kırım» (2012), Neşe Sarısoy Karatay; editör: Osman Gürkan «Gamalı haç ile kızıl yıldız arasında Türkler (Kırım Türkleri, Sürgünler, 1944–1991)» (2011), Özgür Fındık «Kara vagon : (Dersim-Kırım ve sürgün)» (2012) eserleri yer almaktadır.

6. «Kırım Türklerin Kırıma dönüşü (1987–2017 yıllar)» bölümde sosyoloji çalışmalarından oluşuyor. 1987 yıllından Kırım Türkler ana vatanına uzun sürgülün yıllardan sonra dönmeye başlamıştı. Nisan 1989'da, geri dönenlerin sayısı 40,000'i geçmişti. Altı ay sonra, Sovyetler Birliği yıkılmaya yüz tutmuşken, Sovyet Başkanlık Divanı, geçmişte haksızlığa uğrayan Kırım Tatarları ve Sovyetler Birliği'nin

diğer tüm halklarının haklarını güvence altına alan bir bildiri yayınladı. Müteakip yıllarda anayurda göç daha da büyük artış kaydetti. Fakat Kırım'da Kırım Türkleri yeni haksız zorluklarla karşılaşmıştı: evsizlik, işsizlik ve maddi desteksiz. Tekrar Kırım Türkler Kırım'da insan haklarını savunmaya devam etmektedir. Bu konu ilgili çok sayıda yüksek lisans tez araştırmaları. Ayla Göl «Sovyetler Birliği'nde Kırım Tatarlarının yer değiştirmesi sorunu» (1988), El`vis Beytullayev «Kırım politikası: 1990–2001» (2001), Gül Sarıkaya «Ukrayna jeopolitiği bağlamında Kırım ve Kırım Türkleri sorunu» (2014), Lyudmila Beybulayeva «Rusya-Ukrayna ilişkilerinde Kırım sorunu 1991–2014» *(2015),* Fethi Kurtiy Şahin «Ukrayna ulus inşasında Kırım Tatar etkisi ve Euro meydan: 2014 sonrası yenilikler ve değişimler» (2016) yüksek lisan tezleri Kırım Türklerin Kırıma dönüş sorularını incelenmektedir.

7. «Kırım Türklerin Milli Mücadelesi» ayrı bir bölüm olarak yer alıyor. Kırım Türklerin tarihi Dünya eylemin nasıl yapıldığını hak arayışı, bağımsızlık mücadelesi nasıl yapılır Kırım Türklerinden öğrenmeli. Bundan 230 yıl önce Osmanlı ile Rusya arasından imzalanan Küçük Kaynarca anlaşmasıyla Kırım'ı kaybettik. Ve o zamandan günümüze kadar o coğrafyada Ruslar Türker'i silmek isteyecek, büyük bir temizleme politikası uygulayacak ve insanı insanlığından utandıran olaylar gerçekleştirecekti. Bu tarihten Kırım Türklerin canı yanmıştır ama can yakmamışlardır. Kırım Tatarlarının gösterileri etkisini gösterdi ve Rusya gösterilerde yer almayan, milli mücadeleye katılmayanları şu tarihten başlanmıştı ve bugüne kadar Kırım'a yerleştirdi. Bu konuyu açıklayan aşağıdaki kaynaklarda bulunmaktadır. 1930'cı yıllarda Rusya Müslümanlarından Türkiye'ye göç etmiş kişilerin arasında Kırım Türk kökenli Cafer Seydametın faaliyeti Emel dergisinin editörü olarak kitap ve makaleler yayınlanmıştı: Yirminci Asırda Tatar Milleti Mazlumesi, (1911) (Şahap Nezihi takma adıyla), La Crimeé (Fransızca), (1921), Krym (Lehçe), (1930), Rus Inkılabı, (1930), Gaspırali Ismail Bey, (1934), Rus Tarihinin Inkılaba, Bolşevizme ve Cihan Inkılabına sürüklenmesi, (1948), Mefküre ve Türkçülük, (1965), Unutulmaz Göz Yaşları, (1975), Nurlu Kabirler (1992),Bazı Hatıralar, (1993). Romanya'daki Kırım Türklerden aydınların birisi Müstecip Ülküsal elinden çıkan eserleri: Dobruca ve Türkler (1940), Kırım Türk Tatarları (Dünü-Bugünü-Yarını, 1980), Kırım Yolunda Bir Ömür-Hatıralar (1999), İkinci Dünya Savaşında 1941–142 Berlin Hatıraları ve Kırım'ın Kurtuluş Davası. Bu bibliyografik listeyi Hakan Kırımlı, Saynur Giray Derman, Muhammed Furkan Engin ve bş. devam etmektedir. Toplam bu konuda 67 adet kaynak.

8. «Kırımın edebiyatı ve bibliyografik» bölüm geniş ve çok sayıda kaynaklardan oluşmaktadır. Kitaplar 37 adet, tezler – 33 adet, makaleler – 45 adet. «Kırım Türkleri yazılı edebiyatları meydana gelene kadar ideallerini, millî karakter özelliklerini, örf ve adetlerini, medeniyetlerini; sosyal, siyasî ve iktisadî durumlarını, arzu ve ümitlerini, dünya görüşlerini çok eski zamanlardan beri yırlar, takmaklar, çınlar, maniler, atasözleri, tapmacalar, lâtifeler, efsane ve destanlarla nesilden nesille geçen sözlü edebiyatlarıyla günümüze kadar getirmişlerdir». (Yüksel, 2003). Bölümde Kırım Türk yazarların edebi metin incelemesi, bibliyografik ve halk bilimi bilgilerimden göstermektedir. Mesele, İsa Kocakaplan «Kırım'ın edebi sesi Cengiz Dağcı: hayatı, kişiliği, sanatı, fikirleri, eserlerinin özet ve değerlendirmeleri, son röportaj» (2010), Zühal Yüksel «Kırım Tatar şairi Hamdi Giraybay» (2012), Zekeriya Karadavut «Kırım Tatar folkloru» (2013), Hüseyin Su «Çağdaş Kırım Tatar öyküsü (Kırım hikâyeleri, Antoloji) (2014) kitapları. Ayrıca Yavuz Akpınar İsmail Gaspıralı hakkında dört cilt kitapları «Gaspıralı Kırım'a sığmıyor. Kendisi Kırımlı ve Kırımlılar onunla ne kadar övünseler yeri. Kırım'ın yanı sıra Türk Dünyası'nı «dönüştüren» bir insan» kitabında ifade etmişti. Doktora tezleri 3 adet: İbrahim Şahin «Cengiz Dağcı'nın Hayatı ve eserleri» (1992), Kenan Acar «Kırımlı dilci Bekir Sıtkı Çobanzade» (1996), Işılay Işıktaş Sava «Kırım Tatar şairi Şakir Selim'in şiirleri: Metin – aktarma – inceleme» (2015). Yüksek Lisans tezlerinin teması çeşitli: Şerif Tiryaki «Bekir Sıdkı Çobanzade'nin Kumuk Dili ve Edebiyatı Tedkikleri adlı eseri; metin-inceleme-dizin» (2012), Susanna Mustafaıeva «Türk Romanında Kırım ve Kırım Tatarları» (2013), Gülşah Bulut Aktaş «Üriye Edemova'nın Ömürlik Yanımdasın adlı eserinden hareketle Kırım Tatar Türkçesinde kelime « (2014), İnci Yelda Şafakcı «Kırım Tatar aydınlarından Hasan Sabri Ayvazov (1878–1938): Hayatı, fikirleri ve eserleri» (2015), Ferhat Uzunkaya «İsmail Gaspıralı Anlatılarında Yapı ve İzlek» (2017).

9. «Kırım Türklerin ve Kırım'da yaşayan halkların dili» konulu bölümünde toplam 37 adet kaynak bulunmaktadır. Çoğulu tez ve makaleler şeklinde sunmaktadır. Doktora tezlerde Kırım Türkler ve Kırım'da yaşayan Karaylar, Krımçaklar dilleri incelemektedir: Zühal Yüksel «Kırım, Kazan ve Başkurt Türkçelerinde fiil» (1992), Arzu Sema Ertane Baydar «Kırım Tatar Türkçesinde anlamca kaynaşmış-deyimleşmiş birleşik fiiller ve bu fiillerin öğretimdeki yeri» (2002), Lenıyara Selımova «Kırım Tatar Türk ağızları (Akmescit, Bahçesaray, Güney kıyı bölgesi) (ses bilgisi)» (2006), Nesrin Güllüdağ «Kırımçak Türkçesi grameri» (2005); yüksek lisans tezlerin Türkiye'de yaşayan Kırım kö-

kenli Kırım Türklerinin dilinin özellikleri belirmektedir: Aydan Eryiğit Umunç «Türkiye Türkçesi ve Kırım Tatarcasında zaman» (1996), Arzu Sema Ertane Baydar «Eskişehir ve yöresi Kırım Tatar ağzı» (1999). Makaleler 19 adet.

10. «Kırım Kültür-İdeoloji (sanat ve müzik)» kaynakları sayısı kitap 7 adet, tezler – 5, makaleler -8. Bölümde Kırım Türklerin sanatı ile bağlı şu çalışmalar: I. Kúnos'un derlemesinden yayımlayan: Zsuzsa Kakuk «Kırım Tatar şarkıları» (1993), Oktay Aslanapa «Sanatı, tarihi, edebiyatı ve musikisiyle Kırım» (2003), H. Feriha Akpınarlı «Kırım el sanatlarının dünü ve bugünü» (2004), Nicole Kançal-Ferrari. «Kırım'dan kalan miras Hansaray» (2005), Yalkın Bektöre «Kırım türküleri : yaşamdan müziğe yansımalar» (2014) ve sayrı. Müzik sanat biriminde detaylı iki bilimsel çalışma Selma Agat'a ait «Kırım Türk Müziği» (1987) ve «Kırım Türkleri halk oyunları ve geleneksel giysileri» (2002).

11. Bölüm «Kırım edebi eserlerde» Türkiye'de belli yazarların eserlerinde Kırım kavramının yansıması. Bu listede Cengiz Dağcı'nın 18 eseri: Korkunç Yıllar (1956), Yurdunu Kaybeden Adam (1957), Onlar da İnsandı (1958), Ölüm ve Korku Günleri (1962), O Topraklar Bizimdi (1966), Dönüş (1968), Genç Temuçin (1969), Badem Dalına Asılı Bebekler (1970), Üşüyen Sokak (1972), Anneme Mektuplar (1988), Benim Gibi Biri (1988), Yoldaşlar (1991), Biz Beraber, Geçtik Bu Yolu (1996), Rüyalarda Ana ve Küçük Alimcan (2001), İhtiyar Savaşçı, Yansılar 1 (1988), Yansılar 2 (1990), Yansılar 3 (1991), Yansılar 4 (1993), Ben ve İçimdeki Ben (1994), Haluk'un Defterinden ve Londra Mektupları (1996), Hatıralarda Cengiz Dağcı (1998). Sevinç Coşkun ««Hilal Görününce» Türk edebiyatında hak ettiği kadar yer bulamamış bir coğrafyayı, Kırım'ı anlatıyor. Tatarlarla Osmanlı ilişkilerine de yer veren roman, 1853–1856 Kırım Savaşı'nın öncesi ve sonrasını kapsayacak şekilde ilerler. Yakında meydana çıkan Serra Menekay'ın iki kitabı « «Aluşta'dan Esen Yeller» (2015) İkinci Dünya Savaşı yıllarında Kırım'ı, sonrasında Kırım sürgününü, Kırım Türklerinin vermek zorunda kaldıkları yaşam savaşını ve bilahare vatanlarına geri dönüşlerini, Kırım Türklerinin siyasi ancak hep demokratik platformda kalan mücadelesini, sürgün olanların yanı sıra Almanya üzerinden Türkiye'ye kaçanların sıra dışı yaşam öykülerini konu alan bir romandır ve «Şefika» (2017) Romanlarında tarihi gerçeklikleri kurgu karakterler aracılığıyla anlatmayı tercih eden Serra Menekay dördüncü romanı Şefika'da ünlü Türkçü, eğitimci ve reformcu aydın İsmail Gaspıra'nın kızı Şefika Gaspıralı'nın inanılmaz hayat öyküsünü konu almış. Bi-

rimde daha birkaç eserler yer almıştır: Gönül Şamilkızı «Kırım Ateşi» (2017), Berkant Çolak Kırım Tatar Türkleri'nde kayıp zamanlar (Türk şiiri) (2013), Fevzi Altug «Dikenli ilişkiler : Kırımlı bir öğretmenin anıları ve şiirleri (Kırım şiiri)» (2005) ve diğer.

12. Son bölüm «Diğer» adlandırmaktadır. Bu birim çeşitli konuları tek kaynaklardan oluşuyor. Bölümde yüksek lisans tezi arkeoloji: Gülşen Abbasoğlu «Bizans Dönemi'nde Kherson» (2015), turizm: Esfer Katayev «Kırım'da turizm endüstrisi ve bölgesel kalkınma açısından rekreasyon ve turizm potansiyelinin değerlendirilmesi» (2015); ekonomi: Mehmet Çetin «On dokuzuncu yüzyıl Osmanlı İmparatorluğu'nun eko – lojistik analizi: Kırım Muharebesi örneği» (2015), Coğrafya: Fatih Koca «Kırım'ın ekonomik özellikleri ve Karadeniz Ekonomik İşbirliğindeki önemi» (2010); arşiv: Zeynep Özdem «Kırım Karasubazar'da sosyo-ekonomik hayat (17. yüzyıl sonlarından 18. yüzyıl ortalarına kadar)» (2006) ve sayrı. Toplam 20 adet kaynak.

Sonuç

Sonuç olarak, her alanda olduğu gibi bibliyografyacılık alanında da bazı sorunların olduğu gözden kaçmamaktadır. Çalışmalar arasında üslup ve yöntem birliğinin sağlanamamış olması başta gelen sorunlardandır. Şimdiye kadar yapılan çalışmalarda, künye sıralamasında bazı araştırmacıların yayın yılını esas aldığı görülürken, bazılarının da konu başlıklarını esas aldığı, çoğunluğun ise yazar soyadına göre alfabetik ya da kronolojik sıralamayı tercih ettiği görülmektedir. Hangi yöntem kullanılırsa kullanılsın, çalışmanın sonuna kapsamlı bir dizin konulması eserden istifadeyi azamî derecede artırmaktadır. Birçok yayında dizinin bulunmayışını önemli bir eksiklik olarak zikrediliriz. Güncellenmiş genel Kırım bibliyografyaları ve kronoloji çalışmalarının yeterli sayıda olmaması, bu alanda hâlâ büyük bir boşluğun bulunduğunu göstermektedir. Çalışmamızda toplam 733 adet bilimsel çalışma ve yayın listelenmiş olup, bunların 268'ı Kitap, 178'ı tez ve 287'sı makaledir. Bu sayılar da göstermektedir ki, tarama konumuz ile ilgili Türkiye'de ciddi bir literatür, ilmi ve fikri birikim oluşmuş bulunmaktadır. Hiç şüphesiz bu birikimin oluşum süreci, aşamaları, muhtevası; bir başka ifade ile genel olarak Türkiye'de Kırım ile ilgili çalışmaların kişiler, kurumlar, önemli eserler ve süreli yayınlar gibi çok farklı açılardan ele alınması ve oluşan birikimin tarihi, siyasi, sosyolojik açısından çözümlenmesi müstakil çalışmaları gerektirmektedir. Çalışmamızda yer alan araştırma ve yayınların dönemler ve konular itibariyle dağılımına genel olarak bakıldığında bazı hususların dikkate değer görülmektedir.

Türkiye için Kırım önemli bir tarihsel merkez gibi görünmektedir, bu mirası yaşatma konusundaki kararlılığı göstermektedir. Geniş ilgisini bütün Türk Dünyası için değerini de Türkiye ve Türk Dünyasından bu bilimsel çalışmalar bilim insanlarının katkısı açıkça ifade etmektedir. Gören çabaların nitelikli bir şekilde sürdürülmesi son derece önemlidir.

KAYNAKLAR

BOZDAĞ, İsmet (1975) Atatürk'ün Sofrası, Kervan Yayıncılık, İstanbul.

İNANCIK, Halil (2017), Kırım Hanlığı Tarihi Üzerine Araştırmalar (1441 – 1700), İstanbul, İş Bankası yayınları, 700 s.

KARAYALÇIN, Yaşar (1954) Bibliyografya Meselelerimize Umumi Bir Bakış, Ankara, Ankara Üniversitesi Hukuk Fakültesi, 28 s.

KIRIMLI, Hakan (1996), Tatarlarında Milli Kimlik ve Milli Hareketler, Türk Tarih Kurumu Yayınları, Ankara**,** 253 s.

SOLAK, Fahri (2007) «Türkistan ve Kafkasya Bibliyografyası», İstanbul, TDBB, 758 s.

YÜKSEL, Zühal, (11.02.2018)1917-1944 Dönemi Kırım Türk Edebiyatı, http://vatankirim.net/5-1917-1944-donemi-kirim-turk-edebiyati-577/.

http://www.islamansiklopedisi.info/dia/ayrmetin.php?idno=d040145 (11.02.2018).

УДК 81’322.4

MACHINE TRANSLATION IN THE TURKIC LANGUAGES (MAINLY IN THE AZERBAIJANI LANGUAGE)

Gulnara Jafarova
Azerbaijan National Academy of Sciences, Institute of Linguistics named after Nasimi, Azerbaijan, Baku
gulnarajafarova@yahoo.com

The issue of the machine translation in computer technologies appeared with the creation of the first type personal computer half a century ago. Machine translation has become one of the most important fields for researches (linguists, philosophers, programmers) of the computer in the last 60 years. The intellectual natural language processing is implemented automatically with the help of different technologies.

The article is devoted to the machine translation systems as one of the main issues of the artificial intellect in the Turkic World in the beginning of the XXI centuries. Main researches were dedicated to the development of the semantic models, language semantics, semantic analysis of the text, the applying of the linguistic resources, and the issues about thesaurus, machine translation, semiotic modeling and the creation of the National Language Base. The implementation of these issues in Turkey, Kazakhstan, Uzbekistan, Tatarstan and Azerbaijan is discussed.

Keywords: machine translation, semantic analysis, Turkic languages, electron dictionaries, semantic relation.

МАШИННЫЙ ПЕРЕВОД В ТЮРКСКИХ ЯЗЫКАХ (НА МАТЕРИАЛЕ АЗЕРБАЙДЖАНСКОГО ЯЗЫКА)

Гульнара Джафарова
Академия наук Азербайджана, Институт языкознания имени Насими, Азербайджан, Баку
gulnarajafarova@yahoo.com

Направление машинного перевода в информационных технологиях возникло с созданием первого персонального компьютера полвека назад. За это время машинный перевод стал одной из самых важных областей научных исследований для лингвистов, философов и программистов. Интеллектуальная обработка естественного языка осуществлялась автоматически, с использованием разных технологий обработки естественного языка. В статье рассматривается система машинного перевода как одна из основных проблем искусственного интеллекта в тюркском мире в начале XXI века. Основное направление было посвящено разработке семантических моделей, языковой семантике, семантическому анализу текста, применению лингвистических ресурсов, а также вопросам тезауруса, машинного перевода, семио-

тическому моделированию и созданию национальной языковой базы. В статье затрагивается реализация данных задач в Турции, Казахстане, Узбекистане, Татарстане и Азербайджане.

Ключевые слова: машинный перевод, семантический анализ, тюркские языки, электронные словари.

1. Main text

The researches belonging to the machine translation system started in the 40 years of XX century were elucidated in the science world, mostly. We will pay attention to the researches in the Turkic World in the article.

Machine Translation problem in Turkology was engaged in research comparatively for the Turkish and English languages by the guidance of Leon Dostert at the University of Georgetown in USA since 1961, English – Turkish Translation system was created. The algorithm and the dictionary were included to the system. The text consisting of 3500 words was chosen for the machine translation system prepared for those languages. The dictionary included to the system consists of 700 words. The words in the Turkish language dictionary are the equivalents of the English language words consist of the list of the base form and suffixes. The information coinciding to each of the word is noted here: that is the substitutions (the sound substitutions in the speech, the letter substitutions in the written form), the rhythm of the vowels, and the order of suffixes for each other were considered (Мельчук И. и др., 1967: с.188–192). This problem was researched again at the end of the 70th years. In the last years, the morphological analysis of the sentences in the Turkish language was researched by the guidance of Dr. Kemal Oflazer at the University of Bilkend in Turkey (Oflazer, Kemal, 2001).

The researchers of Tatar language such as F. A. Dreyzin (Дрейзин Ф. А и др.), P. K. Bayramova (Байрамова П. К., 1966) for syntax, the specialists of Uzbek language such as N. A. Xalitova (Халитова Н. А., 1960), X. F. Iskhakova (Исхакова Х. Ф., 1972) for morphological analysis, Yakubova (Якубова Н., 1979) (Uzbek language), Otkupshikova (Откупщикова М. И., 1963, с. 61–65) (Mongol language), G. O. Yefremov (Ефремов Г. О., 1963, стр. 79–88) (Chuvash language) for synthesis realized many research works in the machine translation system during 60–70 years.

Machine Translation problem takes an important place in the linguistics of many Turkic nations form the beginning of XX century. So

that, let's give some information about the research works realized in Kazakhstan, Uzbekistan, Turkey, Tatarstan and Azerbaijan.

The leader of the fund «Wikibilim» – Payan Kenjexanuli noted about the system of Kazak Tilі Google Translate that this system will be improved. This system can be worked on 90 languages currently and approximately 200 thousand users apply this system every day (automatic translation system was realized on the 12th of December and it was included to the server on the 21st of April in 2012 (www.navka.kz/page).

In 2009, the automatic computer translation system of English-Russian-Uzbek languages was created on the basis of the program prepared by Muftahan, Hakimov in Uzbekistan. 3855 biological terms and words from English, 6027 words from Russian, 6146 words from Uzbek, 984 words from Uzbek on ecology, 989 words from Russian were included to the base of the system. Except this, the Russian-Uzbek field terminology dictionary on economy and clerical work was prepared for enriching the base (http://www.centrasia.ru/newsA.php). We must note that, the given base «Таржимон – L – MX» was prepared for the automatic translation system.

The Academy of Sciences of the Republic of Tatarstan was prepared the project of the common republic automatic machine translation system for Russian – Tatar, Tatar – Russian, Russian – Turkish languages with ALS [ABBYYLS] (Russian linguistic company) in 2012 (http://als.ltd/mt_tatarstan).

The applying of the machine translation problem in the Azerbaijani linguistics was started at the end of the 60tieth years (Велиева К.А., 1971).

The issues such as the compiling of the automatic dictionary (Пинес В.Я., 1970), the synthesis of the word which one of the most important levels of system (Велиева К.А., 1971) were researched, the rules belonging to the agglutinative languages, the rhythm, the order of the morphemes in the word form, the formal description of the morphonological changings in the border of the root and morpheme were given and the compiled program on the basis of the created algorithm was checked on the computer, the results were got. So that, the true word forms were created following the rhythm during the combining of the root and the suffixes in the process of the synthesis, following the morphological changings in the order of the morphemes.

The level of the morphonological analysis being a special place in the Machine Translation system was touched at the end of the 70 years.

Differ from the synthesis process; the word forms in the morphological analysis are divided to the differential parts, here. So that, the list of the morphemes, the grammatical features of the bases was included to the software of the computer such as linguistic base for separating of these to the belonged classes automatically (Махмудов М. А., 1982, Махмудов М. А., 1991, Махмудов М. А., 1994). Except this, the thesauruses playing an important role in the machine translation system were involved to the research such as not only with the translation and analysis of different word forms, but also as the semantic component checking the correctness of the translation and analysis of the fixed word combinations in the text (Мамедова М. Г., 1984, Мамедова М. Г., 1990). The formal analysis of the text for its meaning, the providing of understanding of the text intends the implementation of two main actions, here automatically. The thesauruses of the information research realizes the functions of discovering of the semantic relations between words and the describing of these in this or the other form, so that, thesaurus was intended for discovering the relation among the semantic units of the language of information research describing the subject, firstly (Мамедова М.Г., 1990, 2. с. 38–39).

The syntactic analysis and synthesis being the complicated level in the machine translation system starting from the 90 years of XX century was attracted to the research (Vəliyeva K. A., 1997). The formal models of the language in the syntactic, morphological level were set up in the realizing research works, and the formal models of the sentence types were prepared in the same time.

The phonetic, morphological changings which will be happening in the process of the word forms were researched in the syntactic synthesis.

The issues of the checking of correctness of the sentence structure, the following of the word forms to the orthographic rules (as rhythm, the position of the morphemes for the root of the word, the checking of the reduction of the grapheme), the syntactic, semantic and grammatical relations among the word forms, the rules of the following of word-forms line-by-line and the rules of the using of punctuation marks were attracted to the research work for the aim of the editing of texts by computer automatically in the last years (Vəliyeva K. A., 1998, Vəliyeva K. A., 1999, Vəliyeva K. A., 1995, Vəliyeva K. A., 2001, Vəliyeva K. A., 2003).

We must noted that, the most of the research works realizing on the basis of the Machine Translation system was carrying theoretical char-

acter. The practical features were prefered in the machine translation system in this regard, AzSpellcheck system checking the correctness of the Azerbaijani language texts was created and realized including to the Windows system in the last years. Except this, many of the research works were realized. In this regard, the research work related to the Azerbaijani language in the electron place called «Elektron imkanlar: Azərbaycan dili və türk dillərinin inteqrasiya problemlərinə konseptual yanaşma» written by Firudin Ahmadov attracts our attention (Əhmədov F., 2007., səh.3). the problems of the main two directions relating to each other were touched in the article:

1. Using from the languages belonging to the Turkish group effectively for the development of the proximity and unity of the Turkish people and states;

2. The organization of the activity in the form of the relation of relevant institutions of states for the guarantee of protection, development and using of some Turkish languages (Əhmədov F., 2007., səh.3).

One of the research works assuming practical importance is the preparation of the digital modelling method by A.Fatullayev for the automation of the translation process from the Azerbaijani language to the other languages (Fətullayev Ə.B., 2006). It is needed to note specially one of the other interesting works – «Azərbaycan dilinin formal modellərinin yaradılması və onların əsasında linqvistik prosessorların qurulması» written by Z. Amirov (Əmirov Z. M., 2006, s. 3–4.). The coinciding mathematical algorithm and methods were used for the applying of the language properties and the analysis of these properties in the lexical and morphological level in the research work. The prepared algorithm was applied to the system of the Automatic Texts Processing and the compiled morphological dictionaries were included to the system of computer dictionary. Except this, the proposed methods were used in the system «Azərbaycan dilində mətnlərin düzgünlüyünü yoxlayan» (Azərbaycan dilindəki mətnlərin düzgünlüyünü yoxlayan Korrektor sistemi).

As we noted in the machine translation system, the preparation of the formal grammars is the most important problem, in this regard, «The formal grammar of the Azerbaijani language» written by Aladdin Khalili can be a sample which prepared in the last years (Xəlili Ə. M., 2009).

It is known that the most necessary base such as linguistic guarantee of the Machine Translation system is the compiling of the automatic dictionaries. In this meaning, the realized research works prepared by Zarifa Guliyeva in the last years have an undeniable importance

(Кулиева З.Ю., 2011). The principals of the formation of optimal structure of automatic dictionary were identified, the comparable analysis of the morphological, syntactic and semantic systems of applied languages (Azerbaijani and English) was realized, the formal models of the li8nguistic units were established, the knowledge base of the expert system supporting to the machine translation on the basis of the lexical – grammmatical information was prepared in the machine translation system (Кулиева З.Ю., 2011, s. 28–29).

As we noted above, the most difficult level in the machine translation system is the semantic analysis of the sentence which is impossible to formalize.

At first time, G.N.Jafarova (Aliyeva) touched to the semantic analysis of the sentence in Azerbaijan. She notes that semantic dictionaries were compiled for the semantic analysis of the sentence. Firstly, the belonging of the translated text to which field of science must be clarified in the process of analysis. The dependence on meaning among logic – base words and thesaurus dictionaries are used for identifying the field of science automatically, besides, it is necessary to take into account the absolute context (Cəfərova G.N., 2019, s. 193–196).

The complete explanation of all levels of Machine Translation was given in the monography named «Azərbaycan-ingiliscə tərcümə sistemində» written by K.A.Valiyeva.

In the Azerbaijani linguistics, specially, from the research works realized in the last years, we must note the appointment of the meaning of the words used in the context of the machine translation system prepared by A.Aliyev (Əliyev Ə. A., 2012:20). One thousand multiplicative words were chosen for its frequency in the English language, their meanings were researched, and the formal signs for every meaning were prepared and the compiled algorithms were included to the base of the Dilmanc Translation system in the research work.

Besides, many of the research works having an important place in the machine translation system nowadays were being realized. We can include many names of the young scientists to this rank: Konul Habibova, Gulnara Jafarova, Rena Mammadova, Maya Heydarova, Lala Muradova, Jala Mammadova, Aynur Alizadeh, Amina Rahimli, Azizli Natavan and the others (Габибова К.А., 2019 г./ Mammadova R.H. 2019, c. 103–107/ Məmmədova R.H., 2019, s. 168–179/ Heydərova M.İ., 2018, s. 187–200/ Cəfərova G.N., 2019, s. 193–196/ Məmmədova J.E., 2018, 143 s./ Əlizadə A., 2019/ Əzizli N., 2019).

REFERENCES

1. «Azərbaycan dilindəki mətnlərin düzgünlüyünü yoxlayan Korrektor sistemi».

2. Cəfərova G. N. «Maşın tərcüməsində sözün semantik təhlilində lüğətlər (Tezaurus lüğətləri)», AMEA Filologiya və Sənətşünaslıq, №1, ISSN 2663-4368, UOT 81, Bakı-2019, səh. 193–196.

3. Cəfərova G.N. «Maşın tərcüməsinə nəzəri baxış», Filologiya məsələləri №4, «Elm və təhsil» nəşriyyat, Bakı-2018, ISSN 2224-9257, səh. 147–157.

4. Əliyev Ə. A. 2012. İngiliscə-Azərbaycanca maşın tərcüməsi sistemində felin çoxmənalığının alqoritmik həlli. Fil.ü.f.d. NDA, Bakı, 20 s.

5. Əhmədov F. «Elektron imkanlar: Azərbaycan dili və türk dillərinin inteqrasiya problemlərinə konseptual yanaşma». B., 2007., s. 3.

6. Əmirov Z. M. «Azərbaycan dilinin formal modellərinin yaradılması və onların əsasında linqvistik prosessorların qurulması». NDA, B., 2006, s. 3–4.

7. Əlizadə A. «Принципы составления автоматического словаря в компьютерных программах обучения», Баку-2019.

8. Əzizli N. «Dildə sözlərin avtomatik sıralanması», Bakı-2019 (əlyazma).

9. Fətullayev Ə. B. «Azərbaycan-ingilis maşın tərcüməsisistemi üçün rəqəmsal modelləşdirmə m metodunun işlənib hazırlanması və tətbiqi». NDA, B., 2006.

10. Heydərova M.İ. «Azərbaycan dili elektron məkanda», Dilçilik İnstitutunun əsərləri, №2, Bakı: 2018, s. 187–200.

11. Xəlili Ə. M. «Deduktiv maşının biliklər bazsının tərkib hissəsi kimi» Məhdud Azərbaycan dilinin formal qrammatikasıiişlənib hazırlanması. NDA, B., 2009.

12. Mammadova R. H. «Dictionary block of the national corpuses of the turkic languages». Балтийский гуманитарный журнал. Том 8, №1 (26), 2019, НП ОДПО «институт направленного профессионального образования», Тольятти. s.103–107.

13. Məmmədova J. E. «Sözlərin elektron lüğətlərdə təqdimi və izahı prinsipləri». Fil.ü.fəl.dok. ... dis. Bakı, 2018, 143 s.

14. Məmmədova R. H. «Dil korpuslarında elektron lüğətlərin verilməsi üsaulları», Bakı-2019, №4, «Elm və təhsil», İSSN 2224-9257, s. 168–179.

15. Muradova L. «Kompüter dilçiliyinə dair bəzi məsələlərin proqramlaşdırma üsulları». Tədqiqlər №3, Bakı: 2018.

16. Oflazer, Kemal. «Developing a morphological for Turkish»// Proo. Of the NATO ASI on language Engineering for lesser-studied languages – NATO, ASI, July 2001, Ankara.

17. Orfoqrafiyanın avtomatik yolla yoxlanılması. AzSpellcheck.

18. Rəhimli Ə. «Yüksək səviyyəli proqramlaşdırma dilləri və verilənlər bazasının texnologiyası» (əlyazma).

19. Vəliyeva K. A. «Mətnin avtomatik sintaktik təhlili və sintezi». B., 1997, ADD.

20. Vəliyeva K. A. «Azərbaycan mətnlərində sözlərin düzgünlüyünün yoxlanılması yolları», Tədqiqlər, 2, B., 1998.

21. Vəliyeva K. A. «Mətndə söz birləşmələrinin düzgünlüyünün yoxlanılması yolları». Tədqiqlər 3, B., 1999.

22. Vəliyeva K. A. «Söz formalarının sətirdən-sətrə avtomatik yolla keçirilməsi» // Filologiya məsələləri: Nəzəriyyə və metodika. B., 1995.

23. Vəliyeva K. A. «Dildəki sapmaların aşkar edilməsi». Tədqiqlər, 2, B., 2001.

24. Vəliyeva K. A., Məmmədova M.H. «Mətnlərin avtomatik redaktəsi». B., 2003.

25. Велиева К. А. «Формальное описание синтеза азербайджанского слова». Дисс. на соискание ученой степени канд. филол. наук. М., 1971

26. Ефремов Г. О. «К вопросу о машинном переводе с русского языка на чувашский». «Уч.зан. Чувашского Педагогического Института», 1963, вып.15, стр. 79–88.

27. Габибова К. А., I Международная научно-практическая конференция «Современные тренды управления и цифровая экономика: от регионального развития к глобальному экономическому росту», MTDE, 14–15 апреля 2019 г., Екатеринбург, Россия.

28. Дрейзин Ф. А., Рашитов Р. С. «Принцип синтаксиса татарской обертки» // Мажей перевод. М., 1961.

29. Байрамова П. К. Переводы на татарский язык. ний. АКД. Л., 1966.

30. Исхакова Х. Ф. «Исследования по формальной морфологии тюркских языков (по содержанию татарского литературного языка в переписке с узбекским и узбекским языками)». АКД, 1972.

31. Кулиева З. Ю. «Определение оптимальной структуры автоматического словаря в системе машинного перевода». АКД, Б., 2011.

32. Мельчук И., Равич. А. «Автоматический перевод», 1949–1963, М., 1967, с. 188–192.

33. Махмудов М. А. «Формирование формально-морфологического анализа тюркских слов (на азербайджанском языке)» АКД, Баку, Элм, 1982.

34. Махмудов М. А. «Система автоматической обработки турецкого текста в лексико-морфологическом классе», Б, 1991.

35. Махмудов М. А. «Система автоматической обработки азербайджанских текстов», DDA, B., 1994.

36. Мамедова М. Г. «Автоматизированная лексическая лексика в информационных учебниках на основе терминологической терминологии». Авт. дисс. канд. тех. наука М., 1984.

37. Мамедова М. Г. «По терминологии терминология азербайджанского языка – советская поэзия», 1990.

38. Мамедова М. Г. «О терминологии терминологической терминологии азербайджанского языка – советская поэзия, 1990, 2. с. 38–39.

39. Откупщикова М. И. «Один из возможных способов формулирования передовой морфологии.» // Материал Математическая лингвистика и машинное обучение. Сб. Л., 1963, с. 6-1-65.

40. Пинес В. Я. «Моделирование структуры азербайджанских глагольных форм в связи с проблемой автоматического словаря». АКД, 1970.

41. Халитова Н. А. «Машина перевод с татарского языка на русский. Проблемы туркологии и истории российского востоковедения» – Казань, 1960.

42. Халитова Н. А. Закирова Р. А., Гимадутдинова Р. «Морфологический родной язык в машинном языке в татарском языке» // Учен. Записки Казанского университета, 1962, стр. 4.

43. Якубова Н. «Формальный синтез узбекского словаря», Ташкент, 1979.

44. [www.navka.kz/page]

45. [http://www.centrasia.ru/newsA.php]

46. [http://als.ltd/mt_tatarstan]

УДК 81’33

REPRESENTATION OF DIALECT TEXTS IN THE ORAL CORPUS OF THE KHAKAS DIALECTS

Anna Dybo[a,b], Vera Maltseva[a,b]
[a]Institute of linguistics, Russian Academy of Science, Moscow, Russia
[b]Institute for Oriental and Classical Studies,
Higher School of Economics, Moscow, 105066, Russia
[b]Tomsk State University, Laboratory of Linguistic anthropology,
Tomsk, Russia
adybo@iling-ru.ru

The paper deals with the principles of presenting dialectal texts for their further processing by an automatic parser designed for the literary Khakas language. The parser contains a digital grammatical dictionary of the Khakas language. The dictionary uses standard spelling notation (Cyrillic with additional characters). In this situation, it is important to strike a balance between an adequate reflection of dialectal features and the ability to correctly analyze the maximum number of word forms with an automatic parser. To resolve the situation, we can: a) edit the dictionary, introducing new units into it; b) add new affixes to the parser grammar table; c) change the rules of analysis of already entered affixes; d) bring the representation of dialectal word forms closer to the literary norm. After parsing, the analysis of word forms shall be corrected manually, therefore it is possible to e) add a literary analogue into the parser, subsequently correcting the presentation of the word form to the dialect one. When choosing an action option, the frequency of a particular phenomenon is considered.

Keywords: digital corpus; authomatic morphology analysis; Khakas dialects; sound and grammar pecularities.

ПРЕДСТАВЛЕНИЕ ДИАЛЕКТНЫХ ТЕКСТОВ В РЕЧЕВОМ КОРПУСЕ ХАКАССКИХ ДИАЛЕКТОВ

Анна Дыбо, Вера Мальцева
Институт языкознания, Российская академия наук,
Москва, Россия
Томский государственный университет,
Лаборатория лингвистической антропологии, Томск, Россия
adybo@iling-ru.ru

В статье рассматриваются принципы представления диалектных текстов для их дальнейшей обработки автоматическим парсером, разработанным для литературного хакасского языка. Парсер содержит

цифровой грамматический словарь хакасского языка. В словаре используются стандартные обозначения орфографии (кириллица с дополнительными символами). В этой ситуации важно найти баланс между адекватным отражением диалектных особенностей и способностью правильно анализировать максимальное количество словоформ с помощью автоматического парсера. Для разрешения ситуации мы можем: а) редактировать словарь, вводя новые единицы; б) добавлять новые парсеры в грамматическую таблицу парсера; в) изменить правила анализа уже введенных аффиксов; г) наиболее близко привести представление диалектных словоформ к литературной норме. После парсинга анализ словоформ будет откорректирован вручную, поэтому возможен еще один пункт д) добавить в парсинг литературный аналог, впоследствии изменяя представление словоформы на диалектную. При выборе варианта действия учитывается частота определенного феномена.

Ключевые слова: цифровой корпус; автоматический морфологический анализ; хакасские диалекты; звуковые и грамматические особенности.

Main text

0. The project «Digital corpus of the Khakas language» (khakas.altaica.ru) includes a sub-corpus of oral texts in Khakas dialects, collected and assembled in field expeditions from 2001 to the present. The texts were recorded first on a cassette recorder, then on electronic media in .mp3 format; since 2011 on electronic voice recorders in .wma and .wav formats. All of them are deciphered using the help of native speakers, transcribed and translated into Russian. The current version of the oral dialect corpus is located at: https://linghub.ru/oral_khakas_corpus/search; right now, only a small part of the available materials has been introduced there; significant replenishment of the corpus and additional functions are planned. Presently available representations are: audio file in .flac / .wav format, orthographic representation, parsed morphophonemic representation, glossing, Russian translation. We plan to add phonetic (allophonic) transcription in IPA and an ability to display the word-formation and lexical-semantic glossing within the texts (available currently in the grammatical dictionary; see Dybo, Krylov, Sheimovich 2015).

1. The main problem is the presentation of dialectal texts for their further processing by an automatic parser designed for the literary Khakas language. The parser works with the grammar dictionary of the Khakas literary language, based on the Khakas-Russian dictionary

(KhRS 2006) (22 thousand articles). See about it in (Dybo, Sheimovich 2016). The Khakas texts are analyzed in a standard spelling notation (Cyrillic with additional characters). The problem of automatic processing of dialect texts was discussed, in particular, in relation to the presentation of Russian dialectal texts in the National Corpus of the Russian Language and other dialectal corpora of the Russian language (Sichinava, Kachinskaya 2014; Kachinskaya 2011; Yurina 2011; Kachinskaya 2009; Letuchy 2009). Despite the difference in the sociolinguistic status and age of the Khakas and Russian literary languages (but not in the organization of their dialect systems), we came here to largely similar decisions.

2. In this situation, it is important to strike a balance between adequate reflection of dialectal features and the ability to correctly analyze the highest possible number of word forms with an automatic parser. To resolve the situation, in principle, one can a) edit the dictionary, introducing new units into it; b) add new affixes to the parser grammar table; c) change the rules of the analysis of already entered affixes; d) bring the representation of dialectal word forms introduced into the analysis closer to the literary norm. In this case, the possibility of manual post-editing should be considered. It should also be considered that modern informants allow code switching and mixing between literary and dialect versions in their texts, and therefore it is senseless to process dialect texts with separate dialect modules of the parser, and therefore method c) is practically inapplicable (since it is impossible to select texts in which the parser should use only dialect rules). When choosing an action option, the frequency of a particular phenomenon is considered.

Thus, the layer of presentation of the dialect text which is fed into the input of automatic marking, is not a phonetic transcription, but instead a «normalization» of the dialect text, which only uses the characters of the Khakas alphabet. In general, normalization is a record of a morph-to-morph literary analogue of dialect word forms, calculated on the basis of morphophonological recount.

3. Based on the needs of parsing, the following specific decisions were made.

3.1. The parser must give correct morphological parsing of the word forms of a dialect text that have a morphophonological representation that matches the literary one (the morphophonemic transcription used in parsing is based on Khakas spelling. For more details on the

morphophonemic representation adopted in the Khakas corpus, see Dybo et al. 2019). Let us consider some typical cases[1].

3.1.1. The dialectal and literary forms have the same morphophonemic and phonemic representations; the phonetic realizations of phonemes do not coincide. In these cases, differences in phonetic realizations are reflected in phonetic transcription and phonetic dialectological commentary, but not in the normalization itself[2]. Examples of such a difference: a) Beltyr and Shor [ĕ] correspond to the literary and Sagai *i* [ɘ]: spelling and normalization *минің пилім* (Beltir *[minĕŋ pilĕm]*, lit. *[minɘŋ pilɘm]*) 'my back'. b) spelling and normalization: *хол* (Kyzyl *[x̲ɔl]*, Kachin *[χɔl]*, Staro-Iyus *[qɔl]*, Sagai *[χol]*, lit. *[χɔl]*) 'road'; c) spelling and normalization: *парчадыр (*Kachin *[parɕe̲a̲dɨr]*, lit. *[partɕá̲dɨr]*) 'he goes'.

3.1.2. The phonological differences between the forms are removed if they relate to the same morphophonemic representation. Examples:

a) On the border of the Shor and Beltyr dialects, there are systems in which, the morphophonemes of the first syllable {*o*}, {*u*}, {*ö*} are realized as alternating phonemes /*o*/ ~ /*u*/, /*e*/ ~ /*i*/, /*ö*/ ~ /*ü*/, with the following positions: if the vowel of the next syllable is a reflex of a historical narrow vowel, then a narrow phoneme is chosen; else, in the absolute position and before a historically wide vowel, a wide one: /*em*/ 'medicine' – /*imɘ*/ 'its medicine' , /*emner*/ 'medicines' – /*emnirɘ*/ 'its medicines'. By the normalization the phonemic alternation is removed (= spelling form): *им*, *ими*, *имнер*, *имнері*.

b) The phonemic realization of the morphophoneme {*ч*} in a number of dialects is a separate phoneme with various phonetic realizations; in Kyzyl it coincides in Anlaut with the phoneme /*ʂ*/, which is absent in the literary language, and in the middle and end of the word it is represented by the phoneme /*ɕ*/, opposed to the phoneme

[1] It can be noted that dialect consonantism ceteris paribus remains unreduced to a literary analogue more often than vocalism. The reason, first of all, is that the spelling of the Khakas consonants is oriented more likely to a phonemic representation, and vowels to a morphophonemic representation.

[2] It should be noted that in the RNC Dialect subcorpus, the procedures for «orthographization» of dialect texts are carried out automatically using the programs Detranscriptor-1 and Detranscriptor-2 (Kachinskaya 2009, 3–4). In our corpus, on the contrary, the decoding of a sounding dialect text by a linguist – native speaker is initially performed in a nearly orthographic form; but transcription in IPA is done by a familiar with the language linguist, specialist in phonetics.

/ʂ/ in the back-vowel words (/parɕádɨr/ vs /taʂ/, Domozhakov 1948, 7). Both phonetic and phonological differences in the normalization are removed: spelling and normalization *чол* (Kyzyl /ʂɔl/, Kachin /ɕɔl/, Staro-Iyus /tɕɔl/, Sagai /ɕol/, lit. /tɕɔl/) ‘road’; *парчадыр* (Kyzyl /parɕádɨr/, Kachin /parɕeadɨr/, lit. /partɕádɨr/) ‘he is coming’.

c) The phonemic realization of the morphophoneme {*u*} in the first syllable in some dialects is the phoneme /e/, the same as the realization of the morphophoneme {*A*} in inflectional affixes. This dialectal difference is removed by normalization: *искі*, Kyzyl /eskі/ ‘old, ancient’, *ип*, Kyzyl /eb/ ‘home’. See also examples in b).

Exactly in the same way, well-known correspondences that are regular for certain morphological forms, such as contractions and prolapses, or, conversely, the absence of acquisitions occurring in the literary norm, are removed: *парыбысты* (Sagai /parɨ:stɨ/) ‘he left’, *парчадырлар* (Kachin /parɕeadɨrla/) ‘they leave.’

3.1.3. In a number of dialects, the morphophoneme {*H*} receives phonemic realization /d/ after the morphophonemes {*л*} and {*р*}; this happens both on the border of the affix, and in alternating stems. In this case, the morphophonemic composition of morphemes is identical in different dialects, but another rule of phonemic realization applies. Since we abandoned the special dialect analysis module, we had to use two different approaches in these types of cases. In rarer cases, the end of the stem alternates, so we write the corresponding literary form in square brackets in the dialectal text submitted to the parser for analysis, and the parser successfully analyzes it. By further manual processing, the form in brackets is deleted, and its analysis is transferred to the dialect form. Example: *ирд-і* (lit. *ирн-і*) ‘his lip’. For more frequent cases of alternation in affixes, separate morphemes marked as *«dial.»* are introduced into the grammar table. (i.e., approach 3.2.3 is applied, see below). Examples: *ир-ді* (lit. *ир-ні*) ‘man-Acc’, *ир-дің* (lit. *ир-нің*) ‘man-Gen’’.

3.2. There are also dialectal features appearing in different morphemes constituting the word forms. In particular, we consider to be different the morphemes with different morphophonemic composition. In these cases, we have to supplement the lexical and grammatical inventories.

3.2.1. Supplement of lexical inventory in regular cases.

a) In some Khakas dialects (Shor, Kyzyl), the inventory of consonant morphophonemes is larger than in the literary Khakas: the morphophonemes {*ш*} and {*с*} are distinguished, while in the literary

language they coincide in {*с*}. This difference is manifested in root and derivational morphemes, but not in inflectional ones. Accordingly, if in the literary language the form *астар* receives two analyzes, 1. *ас*-Pl ‘barley’ and 2. *ас*-Pl ‘ermine’, then the Shor forms *аштар* and *астар* should each receive a single analysis. At least in relation to one of the forms, the only analysis is achieved by entering into the dictionary a dialect version of the lexeme *аш* ‘barley’ with the mark «Shor, Kyzyl».

b) in Kyzyl at the end of the stem, three different morphophonemes are distinguished, {*с*}, {*ш*} and {*ч*} (the phonemic realization of the latter is /*ɕ*/), respectively, the base *аҕач* ‘tree’ /*аʁаɕ*/ with a mark «Kyzyl», synonymous with literary *аҕас*[1]. Note that here we are acting differently than is usual in KhRS; in KhRS this phenomenon is not processed quite systematically. We find there not only cases of bringing the Kyzyl lexical dialecticisms to a «literary» phonetic appearance (examples: *ибес* (*-зі*) Kachin, Kyzyl ‘small; piece’ (< **ebeč* VEWT 34); *эгес* (*-зі*) Kyzyl ‘saw, file’ (< **ējke-č*), but also cases of phonetic recording of varying degrees of accuracy (*омащ* Kyzyl ‘spoon’; *пӱрӱщ* Kyzyl ‘pepper’, lit. *пӱрӱс*; *кертещ* Kyzyl ‘hill’; but *ыстырҕаш* Kyzyl ‘tongs’) and our morphophonological method (*мыҷыҕач* Kyzyl ‘cat’; *саруч* Kyzyl ‘ermine’; *чэч* Kyzyl ‘hair’, lit. *сас*). The same three different ways KhRS (inconsistently) uses in relation to the previously mentioned dialectical features, compare *сіс* Kyzyl ‘impenetrable forest (taiga)’, lit. *чыс*; *сыс* Kyzyl ‘smell’, lit. *чыс*; *шыс* (*-зы*) Kyzyl ‘smell’, lit. *чыс*.

c) Through additions to the lexical inventory, the issue of alternations of the initial consonant in auxiliary verbs as part of analytical verb forms has also been resolved. This alternation is caused by an external sandhi with the end of the form of the main verb, cf., for example, the Beltyr *тоҕын мар-* «continue to work» (lit. *тоҕын пар-*). As a matter of fact, for such dialects in which these alternations are observed, one should write in the dictionary the corresponding roots beginning with alternating morphophonemes and introduce a rule for choosing the degree of alternation depending on the last non-zero preceding segment before the Grenzsignal): {*тоҕын-*∅}#{*пар-*}. But, as

[1] However, it was not possible in any way to provide a simple automatic processing for the case of the Kachin positional alternation *аҕас* ‘tree’ – *аҕачы* ‘tree-Pos3’, *сас* ‘hair’ – *чечі* ‘hair-Pos3’ and similar. Here the result of parsing is corrected manually.

mentioned above, we decided against creating a dialect module for the parser, having solved the problem by introducing the dialect auxiliary verbs *мар-*, *мир-* etc. into the dictionary.

3.2.2. The less orderly differing lexical morphemes, stems and derivational affixes, can also be reflected in the dictionary (while ignoring those differences that do not matter for parsing, and avoiding the use of special morphophonemic symbols in the dictionary – this principle also is applied to literary stems). Examples: Sagai, Shor, Beltyr *ник*/*них* (lit. *інек*) ‘cow’, *малты* (lit. *палты*) ‘axe’, Beltyr *уннуң* ‘very’ (lit. analogue is unclear); dial. *аалда-* (lit. *аалла-*) ‘to visit’.

3.2.3. Additional inflectional morphemes are entered into the grammar table with a mark on the dialect. In particular, these are grammatical morphemes that are of the same origin as the literary ones, but have a different morphophonemic composition in the dialect. These, for example, are cases when in a dialect some of the affixes do not obey the row vowel harmony, as the dial. {*че*} Pres., {*ох*} Ass., lit. {*чА*} and {*ОК*}. A more common, but also rarer case: Sagai *парахча* ‘wants to go’ contains the dialectal affix of prospective {*АК*}, absent in the literary language; it is added to the grammar table marked as «*dial*.»

It should be noted that the dialectal (Beltyr) forms of the converbп as a part of an analytic word form like /*товіпт odir-*/ ‘continue to work’ (lit. *тоғынып одыр-*) morphophonologically do not differ from the literary representation: {*тоғын-ЫП*} {*одыр-*}. . /m/ at the end of the converb is the standard representation of the morphophoneme {*П*} after a nasal consonant. The phonemic representation here, however, is translated into the morphophonemic one manually, since the rules for disclosing the results of external sandhi were not introduced into the parser, and the position of {*П*} immediately after the nasal here is due to a dialect external sandhi, which requires the fall of a narrow vowel of the converb marker by a vocalic beginning of the next word form.

4. We try to take into account all morphological phenomena. The layout of affixes, their semantics and morphological variants are adjusted based on the analyzed data. At the moment, about two dozen dialect markers have been entered into the parser

5. Rarely occurring phenomena are left for manual analysis. If later, with an increase of the volume of the analyzed texts, some of them are recognized as frequent regular non-phonetic phenomena, an operational adjustment will be made to the parser.

Acknowledgements

Materials were collected in the framework of the project «Language and ethno-cultural variability of Southern Siberia in synchrony and diachrony: language and culture interaction» (the RF Government grant No. 14.Y26.31.0014); the analysis was made in the framework of the project «Digital Dialectologic Atlas of Turkic languages in Russia» (the Russian Science Foundation project 18-18-00501); the software maintenance of the analysis was made in the framework of the Basic Research Program of the Presidium of the RAS «Monuments of material and spiritual culture in the modern information environment» (the project «Digital parallel corpora of minor Turkic languages and dialects in the Russian Federation»).

REFERENCES

Domozhakov, N. G. (1948). Opisanie kyzyl'skogo dialekta khakasskogo jazyka: avtoref. dis. ... kand. filol. nauk. Abakan.

Dybo, A. V., Krylov, S. A., Sheymovich, A. V. (2015). Nekotorye vozmozhnosti semanticheskoi i etimologicheskoi razmetki dl'a korpusov t'urkskix jazykov (rasstanovka semanticheskix tegov v eelektronnom xakassko-russkom slovare) // Sbornik trudov mezhdunarodnoi konferencii TurkLang 2015. Kazan. Pp. 304–327.

Dybo, A. V., Sheymovich, A. V. (2016). Apparat avtomaticheskogo morfologicheskogo analiza dl'a korpusa xakasskogo jazyka // Rodnoi jazyk, № 2(5). Pp. 9–39.

Dybo, A. V., Krylov, F. S., Maltseva, V. S., Sheymovich, A. V. (2019). Segmentnye pravila v avtomaticheskom parsere Korpusa xakasskogo jazyka. // Ural-Altaic Studies. № 32(2). Pp. 48–69.

Jurina E . A. (2011).Tomskij dialektnyj korpus: v nachale puti // Vestnik Tomskogo gosudarstvennogo universiteta. Filologija, 2, p. 58–63, access at: http: //cyberleninka.ru/article/n/tomskiy-dialektnyy-korpus-v-nachale-puti.

Kachinskaja, I . B. (2009), Korpus dialektnyh tekstov v Natsional'nom korpuse russkogo jazyka: sostojanie i perspektivy // Leksicheskij atlas russkih narodnyh govorov (Materialy i issledovajia), St. Petersburg, p. 57–68, access at http: //www.philol.msu.ru/~ruslang/pdfs/kachinskaya. i.b/19. pdf.

Kachinskaja, I. B. (2011), Dialektnyj podkorpus NKRJA. Novyj standart podachi. Novoe rabochee mesto // Russkaja ustnaja rech'. Materialy mezhdunarodnoj nauchnoj konferencii «Barannikovskie chtenija. Ustnaja rech': russkaja dialektnaja i razgovorno-prostorechnaja kul'tura obschenija». Mezhvuzovskoe soveshchanie «Problemy sozdanijai ispol'zovanija dialektnyh korpusov», Saratov University, Saratov, p. 245–255.

KhRS (2006): Subrakova, O. V. (ed.) Xakassko-russkii slovar'. Novosibirsk.

Letuchij, A . B. (2009), Dialektnyj korpus: sostav i osobennosti razmetki //Natsional'nyj korpus russkogo jazyka: 2006–2008: Novye rezul'taty i perspektivy, Nestor-Istorija, Saint-Petersburg, p. 114–128. http://www.ruscorpora.ru/old/sbornik2008/06.pdf.

Oral Khakas dialectal corpus: https://linghub.ru/oral_khakas_corpus/search.

Sitchinava, D. V., Kachinskaya, I. B. (2014). The dialectal subcorpus within the Russian national corpus: today and tomorrow// Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). – М.: Изд-во РГГУ. С. 620–628, access at http://www.dialog-21.ru/digests/dialog2014/materials/pdf/SitchinavaDVKachinskayaIB.pdf.

VEWT (1959): Räsänen, M. Versuch eines etymologisches Wörterbuchs der Türksprachen. Helsinki.

УДК 81’33

LINGUISTIC TAGGING AND ONTOLOGICAL MODEL OF KAZAKH LANGUAGE FREE PHRASES

G. K. Yelibayeva[1], A. S. Mukanova[2], A. A. Sharipbay[3]
[1 2 3]Scientific-Research Institute «Artificial intelligence»
L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan
gaziza_y@mail.ru, [2]asel_ms@bk.ru, [3]sharalt@mail.ru

This article presents the linguistic tagging of phrases in the UniTurk metalanguage, namely, noun and verb adjoinments that are often found in the Kazakh language sentences. They were created on the basis of the English names of the noun and verb adjoinment’s parts of the Kazakh language, and are used to create a knowledge base of the Kazakh language syntax in the Protégé environment. There are also presented ontological models of formations of these phrases. The results of the work will be applied in semantic searches, question-answering systems and in the development of software applications for knowledge extraction, as well as for training and evaluation of knowledge of the Kazakh language syntax in the e-learning system.

Keywords: Kazakh language, linguistic tagging, metalanguage, ontology, knowledge representation, syntax rules, phrase.

ЛИНГВИСТИЧЕСКАЯ РАЗМЕТКА И ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ СВОБОДНЫХ СЛОВОСОЧЕТАНИЙ КАЗАХСКОГО ЯЗЫКА

Г. К. Елибаева[1], А. С. Муканова[2], А. А. Шарипбай[3]
[1 2 3]Научно-исследовательский институт «Искусственный интеллект», Евразийский национальный университет им. Л. Н. Гумилева,Нур-Султан, Казахстан
gaziza_y@mail.ru, [2]asel_ms@bk.ru, [3]sharalt@mail.ru

В этой статье представлены лингвистические разметки словосочетаний (именных и глагольных примыканий) в метаязыке UniTurk, которые часто встречаются в предложениях на казахском языке. Они созданы на основе английских названий частей именных и глагольных примыканий казахского языка, и используются при создании базы знаний синтаксиса казахского языка в среде Protégé. Также представлены онтологические модели образований указанных словосочетаний. Результаты работы будут применены в семантических поисках, вопросно-ответных системах и в разработке программных приложений для

получения знаний, а также для обучения и оценки знаний по синтаксису казахского языка в системе электронного обучения.

Ключевые слова: казахский язык, лингвистическая разметка, метаязык, онтология, представление знаний, синтаксические правила, словосочетание.

1. Введение

Бурное развитие современных информационных технологий и расширение сферы их применения увеличило потребность в обработке текстов на естественном языке для взаимодействия человека с компьютером. Увеличение количества текстовой информации на естественном языке и сложность их структур требуют создания удобных для пользователя систем, которые размечивают и анализируют тексты на естественном языке. Для реализации таких систем, сперва необходимо создать единый метаязык, который, предназначен для разметки текста на естественном языке и базы знаний предметных областей.

В настоящее время интенсивно развиваются проекты по созданию электронных корпусов для разных языков (в том числе тюркских). Лингвистическая разметка является ключевым компонентом таких корпусов и представляет из себя задание информации о лингвистических единицах непосредственно в тексте в форме разметки на специальном языке. Использование лингвистической разметки очень удобно для задач обработки естественного языка. Они позволяют легко анализировать результаты обработки пользователем или разработчиком, игнорировать не относящуюся к задаче разметку и использовать стандартные инструменты для обработки. Например, такие структуры могут возникнуть вследствие неоднозначности анализа текста на одном из этапов обработки. Представление таких структур значительно усложняет используемую схему разметки, что сводит на нет преимущества от использования стандартных программ и анализируемости человеком [1]. Следовательно, необходимо создать систему лингвистической разметки для обработки естественных языков. Многие существующие метаязыки в основном охватывают понятие романо-германских и славянских групп. Такие метаязыки не могут быть адаптированы к описанию тюркских языков, которые отличаются от упомянутых языков. Поэтому создание метаязыка UniTurk [2, 3], особенно для казахского языка, является ключе-

вым вопросом для лингвистической разметки текстов тюркских языков.

Метаязык необходим для создания общего ресурса, на котором могут работать все разработчики электронных корпусов тюркских языков. Этот ресурс может служить справочной системой как для разработчиков, так и для пользователей тюркских языков, а также для унификации лингвистических разметок, упрощения их понимания и использования общего программного обеспечения. Онтологические модели грамматики тюркских языков можно использовать как наиболее оптимальный компонент, отвечающий необходимым условиям такого ресурса.

Онтология – это иерархически структурированный набор множеств, которые описывают любую предметную область, а также терминологию, которая может использоваться в качестве основной структуры базы знаний, возможности их взаимосвязи и изменения. В основе онтологии лежит объекты в виде набора значений, которые позволяют представлять: свойства, классы, объекты и ограничения. Эти значения взаимосвязаны и объединяются в классы с помощью четких признаков (свойств и ограничений). В результате полного описания объектов и их свойств он может быть представлен как сложная иерархическая база знаний, которая может использоваться в качестве «интеллектуальных» операций, таких как семантический поиск объекта, определение целостности и надежности данных [4-6].

2. Лингвистическая разметка именных и глагольных примыканий казахского языка

Лингвистическая разметка – это разметка текстов и их компонентов специальными тэгами. Такая разметка позволяет идентифицировать тексты по различным параметрам, позволяя осуществлять осмысленный поиск по корпусу [7, 8].

Чтобы автоматизировать обработку текста на казахском языке, нам нужно сделать разметку данного текста с определенными тэгами. В частности, для формализации именных и глагольных примыканий казахского языка прежде всего необходимо ввести специальные лингвистические разметки – тэги для описания синтаксических особенностей. В таблице 1 представлены тэги именных и глагольных примыканий казахского языка в системе UniTurk.

Таблица 1. Тэги именных и глагольных примыканий казахского языка в системе UniTurk

Tag	Name_English		Name_Kazakh				
1	2		3				
	Dependent	Head	Түрі	Бағыныңқы	Басыңқы	Анықтамасы	Мысалы
NP	Noun Phrase		Есімді тіркес			Басыңқы бөлімі есімдер болатын сөз тіркесі	қызық өмір, үш кісі, келген қыз, ағаштың бұтағы, менің ақылдым, үштің екісі, саған қиын
NA	Noun Adjoinment		Есімді қабысу			Басыңқы сыңарлары есімдер болып келетін және ешқандай жалғаусыз, іргелес тұрып байланысқан есімді сөз тіркестері	алтын қасық, үш кітап, өзге әңгіме, алыс үй
NA1	Noun	Noun	Зат есімді сөз тіркестері	Зат есім	Зат есім	Қабыса байланысқан екі зат есімнің біріншісі анықтауыштық қатынаста жұмсалады	Темір күрек, ағаш күрек, жел диірмен, түлкі тымақ
NA2	Adjective	Noun	Сын есімді сөз тіркестері	Сын есім	Зат есім	Сын есім қабыса байланысқан есімді сөз тіркесінің құрамында анықтауыштық қатынаста жұмсалады	Қызыл алма, атты адам, биік тау, қызық өмір, жақсы талап
NA3	Numeral	Noun	Сан есімді сөз тіркестері	Сан есім	Зат есім	Сан есім қабыса байланысқан есімді сөз тіркесінің құрамында анықтауыштық қатынаста жұмсалады	Үш кісі, мың қой, бесінші бригада
NA4	Pronoun	Noun	Есімдікті сөз тіркестері	Есімдік	Зат есім	Есімдік қабыса байланысқан есімді сөз тіркесінің құрамында анықтауыштық қатынаста жұмсалады	Бұл қала, мына бала, осы ауыл, сол табыс

Продолжение таблицы 1

NA5	Participle	Noun	Есімшелі сөз тіркестері	Есімше	Зат есім	Есімше қабыса байланысқан есімді сөз тіркесінің құрамында анықтауыштық қатынаста жұмсалады	Айтылған сөз, келген кісі, келетін бала, оқылатын тапсырма
NA6	Adverb	Noun	Үстеулі сөз тіркестері	Үстеу	Зат есім	Үстеу қабыса байланысқан есімді сөз тіркесінің құрамында анықтауыштық қатынаста жұмсалады	бүгін бала, ертең ата-ана, бүгін қыз, ертең ана
VP	Verb Phrase		Етістікті тіркес			Басыңқы бөлімі етістік болатын сөз тіркесі	пішен ору, ағаш кесу, жақсы оқиды, бүгін келдім, сылқ-сылқ күледі
VA	Verb Adjoinment		Етістікті қабысу			Басыңқы бөлімі етістік болып келетін және ешқандай жалғаусыз байланысқан етістікті сөз тіркесі	таңертең ашылады, қарқ-қарқ күлді, күлімсіреп сөйлеу,майда турады, бес-алты рет жөндейді,арыз бермекші, не алды?
VA1	Adverb	Verb	Үстеулі сөз тіркестері	Үстеу	Етістік	Үстеулер етістікті сөз тіркесінің құрамында пысықтауыштық қатынаста жұмсалады	кеше келді, төмен қарады, бүгінше тоқтай тұр, қазақша сөйлеу, көліксіз келу, амалсыздан тоқтады
VA2	Imitative words	Verb	Еліктеуішті сөз тіркестері	Еліктеуіш сөз	Етістік	Еліктеуіш сөздер көмекші не негізгі етістіктермен тіркескен синтаксистік құрамда қолданылады	жалт қарау, кілт тоқтады, тарс-тұрс жарылды
VA3	Adverbial Participle	Verb	Көсемшелі сөз тіркестері	Көсемше	Етістік	Көсемшелі сөз тіркестері пысықтауыштық қатынастарды, қимылдың амалын, мезгілін, мақстатын білдіреді	ұшып тұрды, сызылып отыру, толтыра асау, шашып жеу, мүлгіп тыңдау

Продолжение таблицы 1

VA4	Adjective	Verb	Сын есімді сөз тіркестері	Сын есім	Етістік	Кісінің сезімін, күйін, қабылдау қабілетін білдіретін сабақты етістіктер сын есіммен байланыста болады, сапалық сын есім бағыныңқы сыңар болады	кең пішті, жаман жазды, оңай түсінді, жақсы сөйледі, ұзын кесті
VA5	Numeral	Verb	Сан есімді сөз тіркестері	Сан есім	Етістік	Сан есімдер етістікті сөз тіркесінің құрамында пысықтауыштық қатынаста қимыл процесін сандық сапа тұрғысынан пысықтайды, «есе» сөзімен тіркесіп те етістікті сөз тіркесін құрайды	жаңбыр бір жауса, терек екі жауады; бес есе орындады, үш есе ұлғайды, екі келіп, екі кетті
VA6	Noun	Verb	Зат есімді сөз тіркестері	Зат есім	Етістік	Зат есімдер сабақты етістіктермен әрі мағыналық, әрі грамматикалық тығыз байланыста сөз тіркесін құрайды	мылтық атты, жер жырту, бала оқыту, қымыз ішу, сөз сөйлеу
VA7	Pronoun	Verb	Есімдікті сөз тіркестері	Есімдік	Етістік	Есімдіктер етістіктермен тіркескенде екі түрлі жағдайда кездеседі, бірінші ден таза атау тұлғада етістіктермен тіркеседі, екіншіден түрлі көмекші сөздер арқылы етістікпен тіркеседі	түгел келді, түгел жиналды, әлдеқашан орналасты

Первый столбец таблицы содержит обозначения тэга, второй – обозначение понятий в виде названия на английском языке, а третий – название тэгов на казахском языке. Текстовая часть описания синтаксических понятий казахского языка в системе

UniTurk (метаязык) состоит из определений и примеров этих понятий, как показано в таблице выше.

3. Онтологическая модель именных и глагольных примыканий казахского языка

Как правило, чтобы автоматизировать процесс создания текстового корпуса, мы сначала делим большие единицы в этом предложении, пока не дойдем до отдельных слов. В нашем случае этими единицами будут словосочетания. Словосочетания содержит как минимум два слова или может быть больше, которые дают одно семантическое значение. Для того чтобы создать текстовый корпус мы должны сперва найти такие единицы, которые имеют одно семантическое значение. Чтобы их найти, нам нужно построить структурированную семантическую модель таких единиц, другими словами онтологию. Онтология является структурно-семантической моделью. Можно также сказать, что онтология – это база знаний, потому что если вы добавите интерпретирующие функции к структурно-семантической модели, то она станет базой знаний.

Словосочетания на казахском языке делятся на устойчивые и свободные. В этой работе рассматриваются свободные словосочетания, на примере именных и глагольных примыканий. На рисунке 1 представлен виды связи слов казахского языка, которые были созданы в Protégé [9, 10].

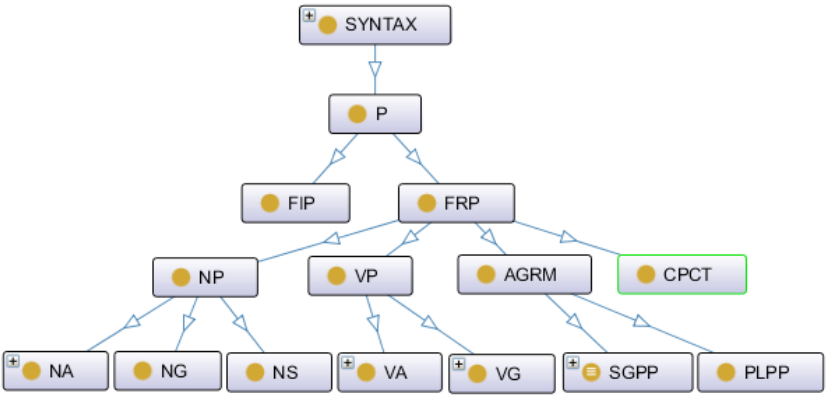

Рис. 1. Виды связи слов казахского языка

где:

- P – Phrase – Сөз тіркесі;
- FIP – Fixed phrase – Тұрақты сөз тіркесі;
- FRP – Free phrase – Еркін сөз тіркесі;
- NP – Noun Phrase – Есімді тіркес;
- NA – Noun Adjoinment – Есімді қабысу;
- NG – Noun Government – Есімді меңгеру;
- NS – Noun Subordination –Есімді матасу;
- VP – Verb Phrase – Етістікті тіркес;
- VA – Verb Adjoinment – Етістікті қабысу;
- VG – Verb Government – Етістікті меңгеру;
- AGRM – Agreement – Қиысу;
- SLPP – Singular personal pronouns –Жекеше жіктеу есімдіктері;
- PLPP – Plural personal pronouns – Көпше жіктеу есімдіктері;
- CPCT– Complicity – Жанасу.

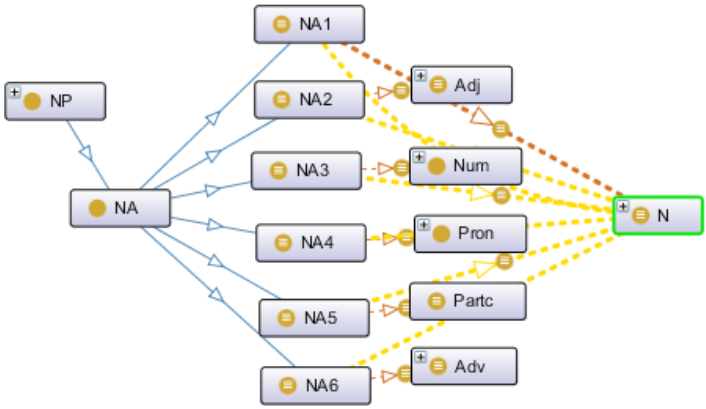

Рис. 2. Онтологическая модель именных примыканий казахского языка

При построении формальной модели словосочетаний учитываем, что словосочетание – это соединение двух или нескольких самостоятельных слов, связанных по смыслу [11, 12]. В словосочетании одно слово главное, а другое – зависимое. Например, в именных примыканий казахского языка главным словом являются именные части речи (в большинстве случаев существительные), а в глагольных примыканий главным словом является сам глагол.

Примыкание – вид связи, при котором зависимость слова выражается лексически, порядком слов и интонацией, без применения служебных слов или морфологического изменения [13-15]. На рисунках 2 и 3 показаны онтологические модели именных и глагольных примыканий казахского языка.

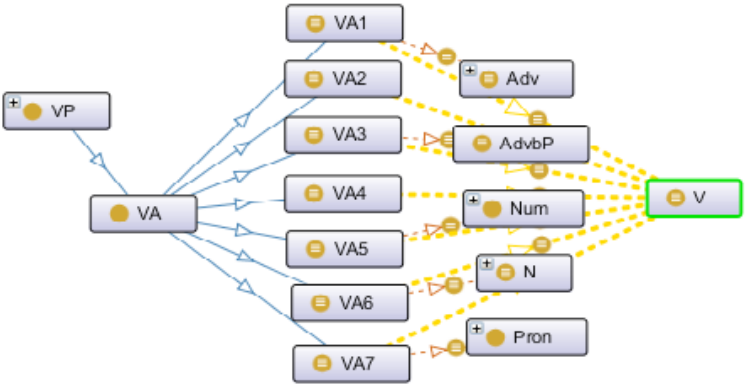

Рис. 3. Онтологическая модель глагольных примыканий казахского языка

На рисунке 4 представлен реализация именного примыкания (зависимое слово – имя прилагательное, главное слово – имя существительное) в среде Protégé.

Например, этому правилу в казахском языке соответствуют словосочетания *«қызыл алма»*, *«атты адам»*. Для этого должны быть выполнены следующие необходимые и достаточные условия:

$$NA2 \equiv \exists hasDependent(Adj) \cap \exists hasHead(N) \qquad (1)$$

где:

- NA2 – именное примыкание (имя прилагательное + имя существительное);
- hasDependent – имеется зависимое слово;
- Adj – имя прилагательное;
- hasHead – имеется главное слово;
- N – имя существительное.

Это позволяет определять словосочетания *«қызыл алма»*, *«атты адам»* которые являются индивидами концепта P (словосочетание) в категорию NA2.

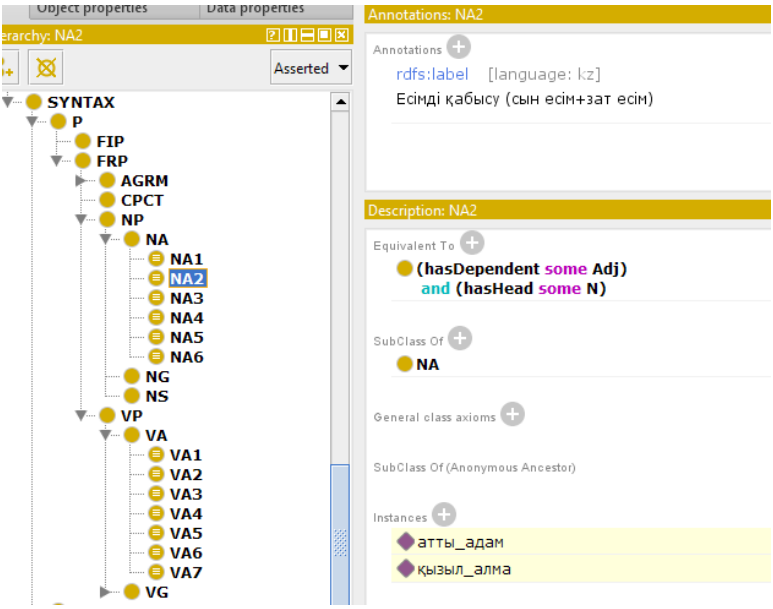

Рис. 4. Именное примыкание
(имя прилагательное + имя существительное)

Кроме того, для синтаксической категории VA1 (примыкание глагола с наречием) должны быть выполнены следующие необходимые и достаточные условия:

$$VA1 \equiv \exists hasDependent(Adv) \cap \exists hasHead(V) \qquad (2)$$

где:

- VA1 – глагольное примыкание (наречие + глагол);
- hasDependent – имеется зависимое слово;
- Adv – наречие;
- hasHead – имеется главное слово;
- V – глагол.

При запуске резонера в среде Protégé, словосочетания *«балаша күлді»*, *«кеше келді»* которые являются индивидами концепта P (словосочетание) определяется как глагольные примыкания в категорию VA1 (рисунок 5).

Такие условия были выполнены в 6 категориях именных примыканий, и в 7 категориях глагольных примыканий как показано в таблице 1. Таким образом, онтологическая модель именных

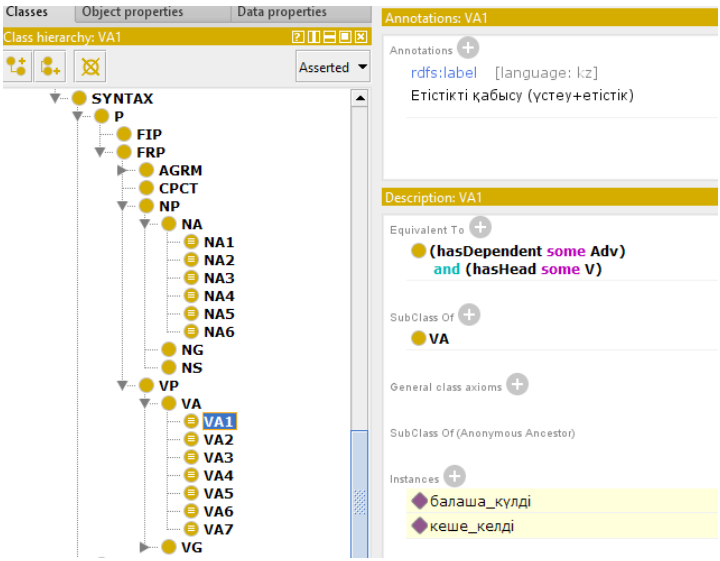

Рис. 5. Глагольное примыкание (наречие + глагол)

и глагольных примыканий казахского языка (рисунок 6), включает в себя все понятий и взаимосвязи этих словосочетаний.

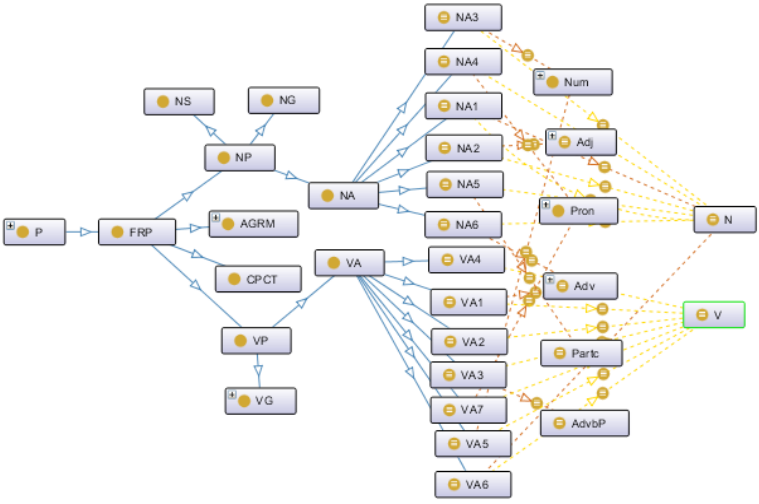

Рис. 6. Онтологическая модель именных и глагольных примыканий казахского языка

4. Заключение

На основе изучения методов создания метаязыков и онтологических моделей сформулирован метаязык UniTurk и разработан онтологическая модель именных и глагольных примыканий казахского языка.

В будущем метаязык UniTurk также будет включать в себя и другие словосочетания казахского языка, и будут созданы онтологические модели других словосочетаний.

Такие лингвистические ресурсы будут использоваться в машинном переводе, в системах многоязычного поиска, а также как базы знаний используемые в системах обучения. Это также является одним из задач автоматизации создания текстовых корпусов.

Благодарности

Статья подготовлена в рамках проекта APS05132249 «Разработка электронных версий тюркских языков для создания многоязычных поисковых и основанных на знаниях систем» по контракту № 132 от 12 марта 2018 года.

ЛИТЕРАТУРА

1. Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. Пособие. – М.: МИЭМ, 2011. – с. 272.

2. Yelibayeva G., Mukanova A., Sharipbay A., Zulkhazhav A., Yergesh B., Bekmanova G. Metalanguage and Knowledgebase for Kazakh Morphology // Lecture Notes in Computer Science. №11619 – 2019. – p. 717–730.

3. Шарипбай А. А., Гатиатуллин А. Р., Ергеш Б. Ж., Қажымұхан Д. А. Разработка единого метаязыка морфологии тюркских языков // Вестник КазНУ – Алматы. №4 (100) – 2018. – с. 78–87.

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

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

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

7. Толдова С. Ю., Логинова Е. А., Попова Д. П. Разметка (лингвистическая) // Режим доступа: http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127221:article [28.06.2019]

8. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. – СПб.: Филологический факультет, 2013. – с. 148.

9. A free, open-source ontology editor and framework for building intelligent systems // Access mode: https://protege.stanford.edu/ [25.06.2019]

10. Горшков С. В., Кралин С. С., Муштак О. И., Гумеров С. З., Мирошниченко М. Г., Гребешков А. Ю., Шебалов Р. Ю. Онтологическое моделирование предприятий: методы и технологии: монография. – Екатеринбург: Изд-во Урал. ун-та, 2019. – с. 236.

11. Разахова Б. Ш., Мусайф М., Жабаева Г. Қазақ тіліндегі сөз тіркестерінің онтологиясы // «Қоғамды ақпараттандыру» III Халықаралық ғылыми-практикалық конференция еңбектері, Астана, Қазақстан, 2012. – 577–580 б.

12. Смирнов И. В. Введение в анализ естественных языков. Российский университет дружбы народов, 2014. – с. 85.

13. Арпабеков С. Қазақ тілі: Анықтамалық. – Алматы: Дәуір, 2004. – 120 б.

14. Балақаев М. Қазіргі қазақ тілі: Сөз тіркесі мен жай сөйлем синтаксисі. – Астана: Л. Н. Гумилев атындағы ЕҰУ, 2006. – 237 б.

15. Әуелбекова А. Ө., Бескемпірова Г. К. Қазіргі қазақ тілінің синтаксисі. – Алматы: «Эпиграф», 2016. – 176 б.

УДК 81’27

DYNAMIC LANGUAGE AS A RESULT OF THE DIFFERENTIATION OF THE SOCIAL STRUCTURE OF THE SOCIETY

E. Isayev, O. Isayeva
The Crimean Federal V. I. Vernadsky University,
Simferopol, the Russian Federation
eduard.krim@mail.ru

The paper is carried out in line with the basic principles of the scientific school of sociophonetics and phonostylistics of Professor A. D. Petrenko. The methodology and methodological principles of sociolinguistic and sociophonetic study of language material that were formulated in the works of W. Labov and A. D. Petrenko are taken as a basis. The problem of linguistic variability is viewed through the lens of historical and social conditions of life of the communicants. Particular attention is paid to the analysis of the relationship of the theory of variability and the concept of a language norm, as well as to peculiarities of speech variants of individual social groups representatives analysis. The basic parameters of sociophonetic analysis of language material are illustrated by an example of possible models of correlation between pronunciation and social structures.

Keywords: variability; language norm; sociolect; sociolinguistic variable.

ДИНАМИЧНОСТЬ ЯЗЫКА КАК СЛЕДСТВИЕ ДИФФЕРЕНЦИАЦИИ СОЦИАЛЬНОЙ СТРУКТУРЫ ОБЩЕСТВА

Э. Ш. Исаев, О. В. Исаева
Крымский федеральный университет им. В. И. Вернадского,
Симферополь, Российская Федерация
eduard.krim@mail.ru

Данная работа выполнена в русле основных принципов научной школы социофонетики и фоностилистики профессора А. Д. Петренко. За основу приняты методология и методологические принципы социолингвистического и социофонетического изучения языкового материала, сформулированные в работах У. Лабова и А. Д. Петренко. Проблема языковой вариативности рассматривается с учетом исторических и социальных условий жизни коммуникантов. Особое внимание уделяется анализу взаимосвязи теории вариативности с понятием языковой нормы, а также особенностям исследования речевых вариантов представителей отдельных социальных групп. Базовые параметры

социофонетического анализа языкового материала рассматриваются на примере возможных моделей корреляции между произносительными и социальными структурами.

Ключевые слова: вариативность; языковая норма; социолект; социолингвистическая переменная.

Развитие языка происходит параллельно с изменением общества. Поэтому объективное изучение проблемы вариативности невозможно без учета социальных и исторических условий жизни коммуникантов. Кроме этого, вариативность – это незаменимое свойство любого литературного языка и возможность варьирования заложена в самой природе языка. На разных этапах развития лингвистической науки специалисты пытались объяснить причины динамичности языковой системы.

В 19-м веке известный младограмматист Г. Пауль [3, с. 52] причину изменений искал в неких индивидуальных колебаниях, внутренних особенностях человека. Во второй половине 20-го века наметилась тенденция к поиску социальной природы языковых изменений. А. Д. Петренко [7, с. 6] главной целью лингвистических исследований называет углубленное изучение форм связи с реальной действительностью, «главным направлением социально-лингвистических исследований является установление общественно значимых вариантов языковых знаков и анализ употребления этих вариантов в социологически релевантных группах, в определенных ситуациях и при определенной целевой направленности речевой коммуникации».

К. Нарингс [15] подчеркивает, что постановка вопроса об истоках и причинах существования языковой вариативности неизбежно приводит к выяснению характера взаимосвязи языка и общества. В любом языковом сообществе существуют различные группы, которые, отличаются друг от друга не только по социальным признакам, но и по языковым маркерам. Кроме этого, характер коммуникативной ситуации и уровень взаимоотношений между коммуникантами также влияют на выбор языковых вариантов. Внутренняя дифференциация языка является следствием и выражением социальной структуры общества.

Вопросы вариативности языковой системы необходимо рассматривать в тесной связи с понятием языковой нормы. В первой половине 20-го века вопрос о степени правильности или неправильности высказывания решался путем сравнения с нормами

письменного языка, которые устанавливались писателями и лингвистами. К примеру, в первом словаре немецкого произношения, подготовленном под редакцией Т. Зибса [17], за основу немецкого литературного произношения было взято сценическое произношение. Создавая первые в истории Германии нормы произношения немецкого литературного языка, Т. Зибс ставил не только задачу унификации языка актеров, но и предлагал руководство для образцового немецкого произношения.

Литературный язык (кодифицированный стандарт) рассматривался Т. Зибсом и его последователями не как форма проявления того или иного национального языка, а как искусственное образование, как идеал, не допускающий никаких отступлений.

Представители Пражского лингвистического кружка Б. Гавранек [8], А. Едличка [8] и др. впервые предложили разграничить понятие объективно существующей языковой нормы (общепринятое речевое употребление) и понятие кодификации (совокупность правил).

Понятие нормы неотделимо от понятия системы языка, существующего в данную историческую эпоху. В современной лингвистике одной из наиболее популярных является теория уровней Э. Косериу [1], в которой различаются три уровня: система – норма – речь. Под системой автор понимает схему, которая охватывает идеальные формы реализации определенного языка. Норма – это система обязательных реализаций тех возможностей, которые заложены в языке. Речь – это индивидуальная реализация нормы. Согласно этой схеме на уровне нормы происходит обязательная реализация, а на уровне речи – нетрадиционная.

А. Д. Петренко [5, с. 30–31] считает вполне очевидным утверждение, «что норма включает в себя различные формы языкового поведения, как на уровне предписания, так и на уровне употребления, и находится в тесной связи с комплексом экстралингвистических факторов, коммуникативной ситуацией общения, коллективом носителей языка, различающихся своим территориальным происхождением и социальным положением».

Отличительной чертой языкознания второй половины 20-го века является учет динамичности языка и признание того факта, что кодификация языковых норм всегда в какой-то степени не соответствует объективно существующей норме. Она обычно отстает от общепринятого речевого употребления, представляя

собой норму прошедшего времени. И. Эрбен [12], Г. Крех [13], Г. Майнхольд [14] допускают существование общепризнанных отклонений от эталона, вариантов норм. Вариантность норм является, по их мнению, одним из основных признаков, из которых складывается понятие литературной нормы.

Х. Штегер [18] связывает понятие нормы с социальными факторами и отмечает, что каждая группа языкового коллектива обладает собственной языковой нормой, а во многообразии социально обусловленных вариантов ярко отражается социальная неоднородность общества. В предшествующие столетия, в эпоху становления немецкого литературного языка, проблема нормирования состояла в преодолении региональных различий. В настоящее время в языке наблюдается конкуренция социальных вариантов. И проблема нормирования на современном этапе состоит в достижении такого соотношения между вариантами, при котором все они самостоятельно, но в едином направлении способствовали бы повышению уровня языковой нормы.

Г. Виллигер [19, с. 42–46] предложил следующую иерархию норм современного немецкого языка (Рис. 1):

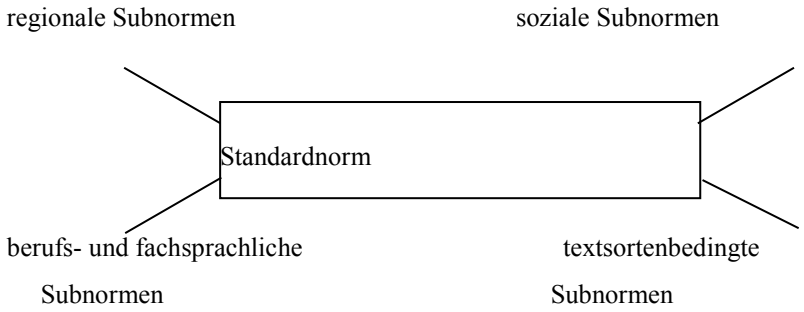

Рис.1. Normenbaum der modernen deutschen Sprache

На Standardnorm замыкаются т.н. Subnormen с ограниченными областями применения: regionale (Северная Германия, Южная Германия, Швейцария, Австрия, Люксембург, Лихтенштейн, Восточная Бельгия, Южный Тироль и т.д.), soziale und standessprachliche (возвышенная, разговорная, вульгарная), berufs- und fachsprachliche (язык охотников, моряков и т.д.; языки науки:

химии, медицины, психологии, языкознания и т.д.), а также textsortenbedingte (реклама; сообщения о свадьбе, о смерти и т.д.).

В рамках такого подхода, варианты, реализуемые в речи, определяются условиями общения и не трактуются как отклонения от нормы.

Э. Косериу [9, с. 232] полагает, что существуют два типа различий между языковыми вариантами: 1) географические (диатопические) различия; 2) социально-культурные (диастратные) различия. «В любом языковом ареале и на всех языковых уровнях существуют определенные географические и социально-культурные отличия. И в этом смысле язык представляет собой некую «сумму языковых систем» (eine Summe von Sprachsystemen)».

Дальнейшее развитие идея Э. Косериу получила в исследованиях Б. Шлибен-Ланге [16, с. 87–94], предложившей для анализа теорию о трех уровнях языковых отличий:

1) диатопические (региональные);

2) диастратные (социолектные);

3) диафазные (стилистические).

При этом особый акцент автор делает на отличиях, проявляющихся в языковых вариантах отдельных социальных групп. Социально маркированные варианты, в таких случаях, выступают не только в роли средства коммуникации, но и идентификации: специфические языковые знаки представляют собой средство распознания «своих».

В 90-е годы 20-го века на факультете иностранной филологии Таврического национального университета им. В.И. Вернадского (г. Симферополь) под руководством профессора А. Д. Петренко начало формироваться новое направление лингвистических исследований – изучение социально обусловленной фонетико-фонологической вариативности национальных языков как целостной структуры. Основные положения данного направления изложены в автореферате диссертации А. Д. Петренко «Социофонетическая вариативность современного немецкого языка в Германии» [6] и паспорте научной школы социофонетики и фоностилистики профессора А. Д. Петренко [2]. Детальный анализ разрабатываемой проблематики представлен в монографиях и публикациях А. Д. Петренко и представителей научной школы (Л. С. Бор, Т. В. Бридко, О. И. Гладких, Э. Ш. Исаев, Э. В. Лихачев, К. А. Мележик, С. Е. Перепечкина, Д. А. Петренко, А. В. Пономарева, Е. А. Устинович, Ю. Б. Федотова, Д. М. Храбскова и др.). В рам-

ках многоаспектного изучения социально обусловленной языковой вариативности исследуются актуальные проблемы социальной стратификации языка, социальные аспекты нормы, стиля и речевой ситуации, определяются базисные единицы фоностилистики и социофонетики, демонстрируется комплексная методика проведения социофонетических исследований.

Результаты данных исследований позволяют проследить динамичность языка вследствие дифференциации социальной структуры общества.

Одним из наглядных примеров изучения взаимосвязи вариантов языковой структуры с элементами общественной структуры может служить анализ произношения немецких школьников. В исследовании вариативности произношения немецких школьников практический материал собирался с помощью комплекса методов для получения информации об одних и тех же информантах при помощи нескольких способов. Такая методика сбора информации позволяет проверить адекватность, качество и репрезентативность данных практически сразу, еще на этапе их обработки.

Социолингвистическая информация о социолекте немецких школьников собиралась следующими методами: метод полустандартного интервью; метод анкетирования, как составная часть интервью; метод скрытой записи; метод «включенного наблюдения».

Особое внимание было уделено попытке смоделировать ситуацию общения таким образом, чтобы максимально уйти от официальной беседы и подтолкнуть говорящего к неофициальному общению, так как вариативность наиболее четко проявляется при смене степени официальности общения. Нормированное литературное произношение зафиксировано в словарях немецкого произношения. При анализе вариативности произносительных форм особый интерес представляет именно ситуация непринужденного общения, в которой повышается вероятность проявления социально обусловленных особенностей.

Целью анализа практического материала было определение средних индексов реализации для каждой фонологической переменной и ее вариантов в трех различных по степени официальности ситуациях общения, что дало возможность для выведения переменных правил реализации фонологических переменных и их вариантов в социолекте старшеклассников, которые могут

рассматриваться в качестве нормативных для данного социолекта. Одним из типичных является пример реализации дифтонгов [a͡i], [a͡u] и [o͡y].

За последние десятилетия изменились нормативные требования, касающиеся правил реализации дифтонга [a͡i] в стандартном немецком произношении. Так, в работах Г. Майнхольда [14], а также в словаре стандартного немецкого произношения GWdDA [11] содержится единственно возможный вариант реализации данного дифтонга – [a͡e].

Однако в ряде исследований по фонетике немецкого языка, проведенных в конце 20-го века отмечается, что все большее распространение в речи получает другая форма – [a͡i]. В частности, А. Д. Петренко [4, с. 89] отмечает, что данный вариант реализации преобладает в произношении немецких студентов не только в непринужденной речи, но и в официальной ситуации. А в новой редакции словаря немецкого произношения DUDEN [10] форма [a͡i] была закреплена в качестве нормативной.

Варианты реализации дифтонга [a͡i] в речи немецких старшеклассников тесно взаимосвязаны с такими параметрами, как официальность / неофициальность ситуации, темп речи, ударность / безударность позиции, величина смысловой нагрузки. Вариативность реализаций дифтонга [a͡i] наиболее четко проявилась в слабых формах, которые, как правило, безударны, несут малую смысловую нагрузку и, следовательно, в неофициальной ситуации при высоком темпе речи подвергаются редукции.

Экспериментальный материал показывает наличие следующих вариантов реализации дифтонга [a͡i] в речи немецких школьников:

[a͡i] → [a͡j] → [a] → «0»-форма.

Вариант [a͡j] наиболее часто отмечен в сочетаниях «неопределенный артикль + имя существительное»:

ein Mann – [a͡j′man], einen Menschen – [a͡jn′mɛnʃn],

einem Bekannten – [a͡jmbɛ′kantn].

В неофициальных ситуациях общения при высоком темпе речи широкое распространение получил вариант [a] в наречиях:

ein biβchen – [an′bisçn], eigentlich – [′agntlix].

Частые случаи полного выпадения дифтонга [a͡i] явились типичными для ситуаций неофициального общения при минимальном внимании к речи со стороны информантов в сочетаниях «предлог + неопределенный артикль»:

auf einer – [aufɑ], in einer – [inɑ].

Что касается особенностей реализации дифтонга [a͡u], то здесь в первую очередь следует также отметить изменение кодифицированной нормы произношения данного дифтонга. В словарях произношения, подготовленных представителями Лейпцигской фонетической школы WdDA [20] и GWdDA [11] в качестве нормативного зафиксирован вариант [a͡o]. В последней же редакции словаря немецкого произношения DUDEN [10] нормативным признан вариант [a͡u]. Изменения были внесены согласно экспериментальным данным, полученным фонетистами за последние десятилетия 20-го века.

В ходе исследования фоностилистической вариативности в произношении немецких старшеклассников были отмечены следующие варианты реализации дифтонга [a͡u]:

[a͡u] → [a͡o] → [o],

aus – [a͡os], auf – [a͡of],

Australien – [ostʹRa·liən], auch – [ox].

Повышение темпа речи в неофициальной ситуации общения, безударность позиции в слове явились причинами качественной редукции составляющих дифтонга [u]→[o], а в некоторых случаях привели к монофтонгизации [a͡u]→[o], причем, с позицией дифтонга в слове явления количественной и качественной редукции оказались мало связанными. Основной причиной выбора информантами того или иного варианта, как свидетельствуют результаты «тестов самооценки», можно назвать степень внимания, уделяемого информантами собственной речи. В быстрой речи при минимальном внимании основным вариантом реализации дифтонга [a͡u] был монофтонг [o]. В ситуации полустандартного интервью при повышенном темпе речи во время обсуждения вопросов, которые особенно волновали информантов также реализовывался краткий [o]. В ситуации полустандартного интервью, но при обсуждении вопросов, которые хорошо знакомы информантам, преобладал вариант [a͡o]. А при чтении списков слов в подавляющем большинстве случаев был использован вариант [a͡u], являющийся на данный момент нормативным.

Анализируя особенности произношения третьего немецкого дифтонга [o͡y], необходимо также отметить изменение кодифицированной нормы произношения для этого дифтонга:

Г. Майнхольд [14], WdDA [20], GWdDA [11] – [o͡ø],

DUDEN [10] – [o͡y].

При реализации данного дифтонга в речи немецких стар-

шеклассников ключевым параметром, существенно влияющим на особенности его произношения, можно признать ситуацию общения. Ибо в одних и тех же словах, у одних и тех же информантов варианты реализации колеблются от полного дифтонга [o͡y] (низкий темп речи + внимание к речи) до монофтонга [o] (высокий темп + отсутствие внимания к речи):

[o͡y] → [o(i)] → [o],

Deutschland – [ˈdo(i)tʃlant], Leute – [ˈlo(i)tə],

Deutschland – [ˈdotʃlan], Leute – [ˈlote].

Степень внимания к речи, в первую очередь, определяла характер реализации и выбор конкретного варианта.

Таким образом, следует отметить, что немецкие школьники владеют как нормативной формой произношения дифтонгов, так и употребляют различные варианты. Распределение этих вариантов связано, главным образом, с параметром «внимание». Чем выше внимание информанта к собственной речи, тем ниже число вариантов, несовпадающих с произносительной нормой. И если при рассмотрении первого дифтонга [a͡i] можно говорить и о позиционной, и о ситуативной зависимости вариантов реализации, то варианты реализации дифтонгов [a͡u] и [o͡y], в основном, зависят от ситуации общения и от связанных с ней характеристик.

Выводы

Наиболее наглядно варианты произношения проявляются преимущественно в неофициальных ситуациях общения и выражаются в различных формах ассимиляции согласных, редукции гласных, выпадении и слиянии сегментов, в отсутствии твердого приступа в начале корневых гласных, а также в отсутствии придыхания у глухих смычных. Существенное отличие в распределении произносительных вариантов отмечено для двух ситуаций общения: официальной и неофициальной. Данный факт еще раз доказывает верность позиции, согласно которой, в исследованиях фоностилистической вариативности произношения достаточно ограничиться анализом речевого поведения в двух диаметрально противоположных типах ситуаций общения.

Полученные результаты служат доказательством возможности проведения социолингвистического исследования на фонетико-фонологическом уровне, т.е. с учетом специфики произношения представителей разных социальных групп в различных ситуациях

общения. Исследование употребления социально значимых вариантов в социологически релевантных группах, в заданных ситуациях речевой коммуникации позволяет проследить отдельные этапы развития языковой системы вследствие дифференциации социальной структуры общества.

СПИСОК ЛИТЕРАТУРЫ

1. Косериу Э. Синхрония, диахрония и история / Э.Косериу // Новое в лингвистике, вып.3. – М., 1963. – С. 175.

2. Научная школа социофонетики и фоностилистики профессора Петренко А. Д. [Электронный ресурс] – Режим доступа: https://iif.cfuv.ru/nauchno-issledovatelskaya-deyatelnost/nauchnaya-shkola-sociofonetiki-i-fonostilistiki-professora-petrenko-a-d/ (дата обращения: 03.10.2019)

3. Пауль Г. Принципы истории языка / Г. Пауль. – М.: Изд-во иностр. лит., 1960. – 500 с.

4. Петренко А. Д. Тенденции развития немецкого произношения в студенческой среде ГДР: Дис... канд. филол. наук: 10.02.04. – К., 1986. – 203 с.

5. Петренко А. Д. Социофонетическая вариативность современного немецкого языка в Германии / А. Д. Петренко. – К., 1998. – 254 с.

6. Петренко А. Д. Социофонетическая вариативность современного немецкого языка в Германии: Автореф. … дис. д-ра филол. наук / А. Д. Петренко. – К., 1998. – 36 с.

7. Петренко А. Д. Актуальные проблемы языковой вариативности в аспекте мировой интеграции и глобализации / Петренко А. Д., Петренко Д. А., Храбскова Д. М., Исаев Э. Ш. – Симферополь, 2011. – 274 с.

8. Пражский лингвистический кружок: Сб. науч. тр. – М.: Прогресс,1967. – 558 с.

9. Coseriu E. Structure lexicale et enseignement du vocabulaire / E.Coseriu // Actes du Premier Colloque International de Linguistique Applique∈Duden. Aussprachewörterbuch. – 3. völlig neu bearb. und erw. Aufl. –Mannheim; Wien; Zürich: Dudenverl., 1990. – 794 S.

10. Großes Wörterbuch der deutschen Aussprache // Hrsg. von E.-M. Krech, E. Kurka, H. Stelzig u.a. – 1.Aufl. – Leipzig: Bibliograph. Inst., 1982. – 600 S.

12. Erben J. Gesetz und Freiheit in der deutschen Hochsprache der Gegenwart / J. Erben. – Stuttgart, 1961. – H. 5.

13. Krech H. Beiträge zur deutschen Ausspracheregelung / H.Krech. – Berlin, 1961. – 133 S.

14. Meinhold G. Deutsche Standardaussprache: Lautschwächungen und Formstufen / G. Meinhold. – Jena: Friedrich-Schiller-Universität, 1973. – 147 S.

15. Nahrings K. Sprachliche Varietäten / K. Nahrings. – Tübingen, 1981. – 281 S.

16. Schlieben-Lange B. Soziolinguistik: eine Einführung / B. Schlieben-Lange. – Stuttgart; Berlin; Köln: Kohlhammer, 1991. – 165 S.

17. Siebs Th. Deutsche Hochsprache (Bühnenaussprache) / Th. Siebs. – Bonn, 1915. – 252 S.

18. Steger H. Sprachverhalten-Sprachsystem-Sprachnorm / H. Steger // Bd.1. Soziolinguistik. Aufsätze zur soziolinguistischen Theorienbildung. – Darmstadt, 1982. – S. 353–372.

19. Villiger H. Sprachnorm und Sprachrealität / H. Villiger // Sprachspiegel. – Bern, 1997. – № 53 (Heft 2). – S. 42–45.

20. Wörterbuch der deutschen Aussprache // Hrsg. von H. Krech u.a. – Leipzig: Bibliograph. Inst., 1964. – 454 S.

УДК 81’33

ONTOLOGICAL MODELS OF MORPHOLOGICAL RULES OF THE KYRGYZ LANGUAGE

N. A. Israilova, P. S. Bakasova
Kyrgyz State Technical University named after I. Razzakov
inela.kstu@gmail.com, bakasovap@mail.ru

The article is based on the results obtained during the implementation of the project on the development of electronic thesauruses of the Turkic languages for the creation of multilingual search and knowledge extraction systems. In particular, ontological models of verbs, adjectives and numerals based on the morphological rules of the Kyrgyz language are considered. The created resource is intended for processing and analysis of the text in the Kyrgyz language.

Keywords: computer linguistics; ontological model; morphological rules; patterns of morphology; Kyrgyz language; knowledge base.

ОНТОЛОГИЧЕСКИЕ МОДЕЛИ МОРФОЛОГИЧЕСКИХ ПРАВИЛ КИРГИЗСКОГО ЯЗЫКА

Н. А. Исраилова, П. С. Бакасова
Кыргызский государственный технический университет им. И. Раззакова

Статья основана на результатах, полученных в ходе реализации проекта по разработке электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний. В частности, рассматриваются онтологические модели глаголов, имен прилагательных и имен числительных на основе морфологических правил кыргызского языка. Создаваемый ресурс предназначен для обработки и анализа текста на кыргызском языке.

Ключевые слова: компьютерная лингвистика, онтологическая модель, морфологические правила, кыргызский язык, база знаний.

Введение. Автоматизация обработки текстов на естественном языке является основной задачей компьютерной лингвистики, что в первую очередь требует описания естественного языка адекватными математическими моделями. Эффективное решение поставленных задач возможно с применением онтологического подхода и семантической сети.

В данной статье представлены некоторые онтологические модели морфологических правил кыргызского языка реализованные в виде онтологических графов построенных в среде разработки онтологий Protégé. Работа выполнена в рамках проекта ИРН: AP05132249 «Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний».

Среда разработки онтологий Protégé позволяет упростить работу с онтологией, а также предоставляет общий доступ к онтологии для удобного совместного проектирования и работы. Проектирование онтологии в данной среде позволяет развернуть иерархическую структуру классов. Более того, возможности среды позволяют генерировать формы извлечения знаний для ввода экземпляров классов на основе спроектированной онтологии. Необходимо отметить, что среда поддерживает OWL 2 Web Ontology Language, язык описания онтологий для семантической паутины. Среда Protégé позволяет работать с несколькими онтологиями в одном рабочем пространстве, а инструменты визуализации обеспечивают удобное перемещение по связям.

На основе созданного единого метаязыка понятий морфологических правил тюркских языков, в частности, на основе разработанного метаязыка понятий морфологических правил кыргызского языка были разработаны онтологические модели морфологических правил кыргызского языка. Рассмотрим онтологические модели морфологических правил некоторых частей речи кыргызского языка

Глагол

Общая характеристика глагола.

Глагол (Этиш) – часть речи, которая обозначает действие или состояние лица, предмета. В кыргызском языке глагол отвечает на вопросы: эмне кылып жатат? (что делает?), эмне кылды? (что сделал? что делал?), эмне кылат? (что сделает? что будет делать?). Различают спрягаемые и неспрягаемые формы. К неспрягаемым формам относятся причастия (атоочтук), деепричастия (чакчылдар), имена действия (кыймыл атоочтор). Спрягаемые формы в предложении выступают в роли сказуемого, неспрягаемые – в роли любого члена предложения. Глагол имеет грамматические категории времени, лица, числа, наклонения, залога. Форма 2-го

лица единственного числа повелительного наклонения является исходной формой, т.е. от нее образуются все остальные глагольные формы.

По составу глаголы делятся на простые (жөнөкөй этиштер) и сложные (татаал этиштер). Простые глаголы состоят из одного слова (корня) и могут быть непроизводными или корневыми (уңгу этиштер) и производными (туунду этиштер).

Ниже в таблице 1 приведены тэги глагола кыргызского языка.

Таблица №1. Таблица тэгов глагола кыргызского языка

Tag	**Name_Russian**	**Name_Kyrgyz**
V	Глагол	Этиш
V_Simp	Простые	Жөнөкөй этиштер
V_Undr	Непроизводные	Уңгу этиштер
V_Drvt	Производные	Туунду этиштер
V_Comp	Сложные	Татаал этиштер
V_Intrs	Непереходный глагол	Өтпөс этиштер
V_Trns	Переходный глагол	Өтмө этиштер
voice	Залог	Мамиле категориясы
V_main	Основной	Негизги мамиле
V_Reflv	Возвратный залог	Өздүк мамиле
V_Passv	Страдательный залог	Туюк мамиле
V_Caust	Понудительный залог	Аркылуу мамиле
V_Recp	Взаимный залог	Кош мамиле
V_Partc	Причастие	Атоочтуктар
V_AdvbV	Деепричастие	Чакчылдар

Продолжение таблицы №1

V_Mood	Наклонение	Ыңгай категориясы
V_Dsrbl	Изъявительное наклонение	Баяндагыч ыңгай
V_Imprv	Повелительное наклонение	Буйрук ыңгай
V_Condt	Условное наклонение	Шарттуу ыңгай
V_Opttv	Желательное наклонение	Каалоо-тилек ыңгайы
	Намерения	Максат ыңгай
TENSE	Время	Чак
PstTense	Прошедшее время	ӨТКӨН ЧАК
V_PastPart	Давно прошедшее время	Жалпы (белгисиз) өткөн чак
V_PastDef	Недавно-прошедшее время	Айкын өткөн чак
V_PastTrans	Переходное-прошедшое время	Капыскы өткөн чак
	Прошедшее длительное	Адат өткөн чак
V_PrstTense	Настоящее время	Учур чак
FtrTense	Будущее время	Келер чак
V_FtrTrans	Переходно -будущее время.	Татаал келер чак
V_FtrIndef	Предположительно-будущее время	Арсар келер чак
V_FtrDef	Определенное будущее время	Айкын келер чак.

Онтологическая модель глагола состоит из морфологических и семантических характекристик глагола кыргызского языка.

На рис.1 приведена разработанная в среде Protege онтологическая модель глагола кыргызского языка:

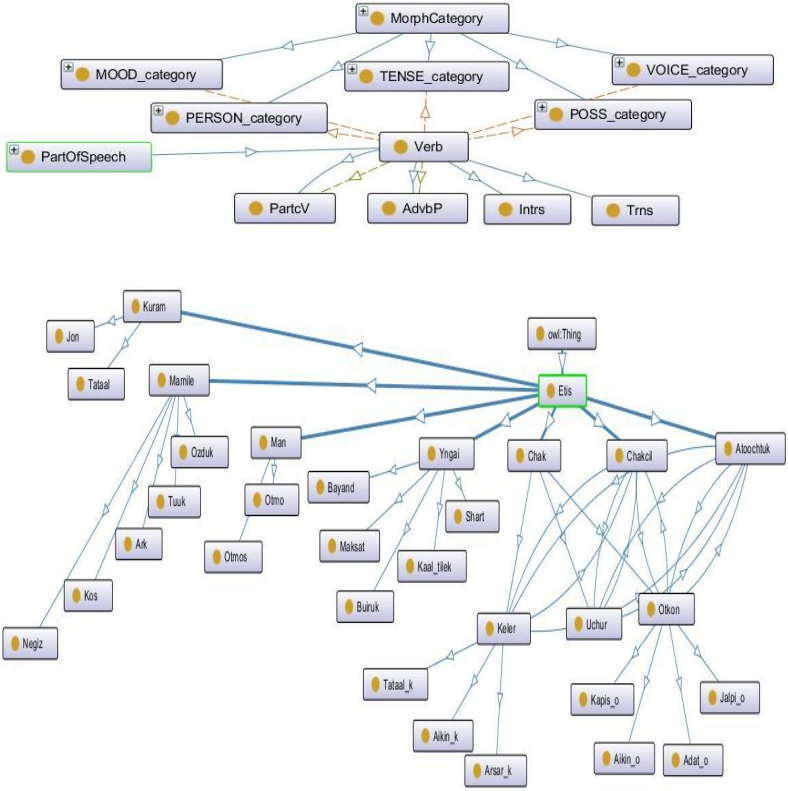

Рис. 1. Онтологическая модель глагола кыргызского языка

Имя прилагательное

Общая характеристика имени прилагательного.

Именем прилагательным (Сын атооч) в кыргызском языке называется часть речи, объединяющая слова, обозначающие признаки, качества, свойства предмета. Имя прилагательное отвечает на вопросы: ***кандай? кайсы? (какой? какая? какое? какие?).*** Имена прилагательные в кыргызском языке в отличие от имен прилагательных в русском языке не имеют категории рода и числа, они примыкают к определяемым словам, а не согласуются, как в русском языке.

По лексическому значению и грамматическим свойствам прилагательные делятся на качественные (сапаттык сын атоочтор) и относительные (катыштык сын атоочтор).

Ниже в таблице 2 приведены тэги имени прилагательного кыргызского языка.

Tag	Name_Russian	Name_Kyrgyz
Adj	Имя прилагательное	**Сын атооч**
Adj_Simp	Простые	Жөнөкөй
Adj_Compl	Сложные	Татаал
Adj_Pair	парные прилагательные	Кош сын атоочтор
Adj_Comp	составные прилагательные	Кошмок сын атоочтор
Qual	качественные имена прилагательные	Сапаттык сын атоочтор
Rel	относительные имена прилагательные	Катыштык сын атоочтор
DegComp	Степени сравнения	Сын атоочтун даражалары
PositDeg	Положительная степень	Жай даража
CompDeg	Сравнительная степень	Салыштырма даража
EnhanDeg	Уменьшительная степень	Басаңдатма даража
SuperDeg	Превосходная степень	Күчөтмө даража

Онтолгическая модель имени прилагательного состоит из морфологических и семантических характекристик имени прилагательного кыргызского языка. На рис.2 приведена разработанная в среде Protege онтологическая модель имени прилагательного кыргызского языка:

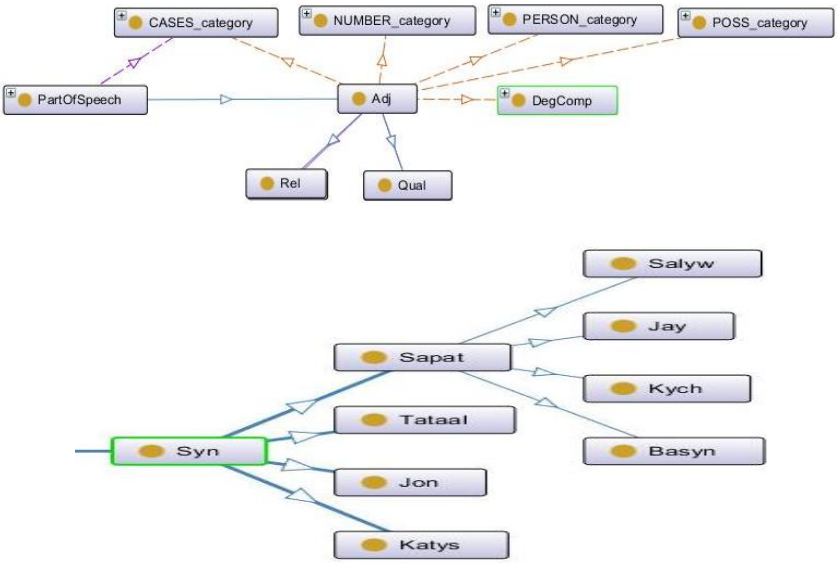

Рис. 2. Онтологическая модель имени прилагательного кыргызского языка

Имя числительное

Общая характеристика имени числительного.

Имя числительное (Сан атооч) – часть речи, которая обозначает количество или порядковый номер предмета и отвечает на вопросы: канча?, нече? –сколько?; канчанчы?, неченчи? – который (по счету)?; канчоо?, нечөө? – сколько (в совокупности)?

По лексико-семантическим и грамматическим особенностям делится на 6 разрядов:

1) количественные (эсептик сан атооч);
2) порядковые (иреттик сан атооч);
3) собирательные (жамдама сан атооч);
4) неопределенно-количественные или числительные приблизительного подсчета
(чамалама сан атооч);
5) дробные (бөлчөктүк сан атооч);
6) разделительные числительные (топ сан атооч).

Ниже в таблице 3 приведены тэги имени числительного кыргызского языка.

Таблица 3. Таблица тэгов имени числительного кыргызского языка

Tag	Name_Russian	Name_Kyrg
NUM	Имя числительное	**САН АТООЧ**
Quan	Количественное	Эсептик
Ord	Порядковое	Иреттик
Coll	Собирательное	Жамдама
Grou	Группирующее	Топ сан
Con	Предположительное	Чамалама
Fra	Дробное	Бөлчөк

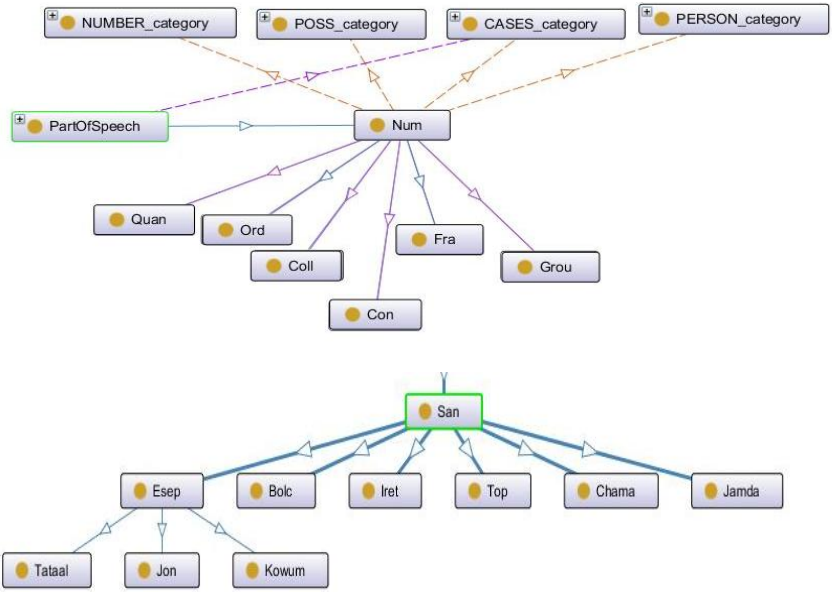

Рис. 3. Онтологическая модель имени числительного кыргызского языка

Заключение

Полученные онтологические модели морфологических правил кыргызского языка обеспечили оптимизацию алгоритмов словообразования глаголов, имен прилагательных и имен числительных, которые реализованы в виде модулей морфологического анализа и синтеза морфологического анализатора кыргызского языка.

Проведено тестирование, в результате которого морфологический анализатор показал устойчивую работу на множестве слов с различными морфохарактеристиками.

Построенные онтологические модели морфологичексих правил и оптимизированные алгоритмы словообразования кыргызского языка могут быть применены в разработке необходимого программного обеспечения по компьютерной обработке кыргызского языка.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ:

Абдувалиев И., Садыков Т. (1997). *Азыркы кыргыз тили: Морфология*. Бишкек.

Th.Gruber. *What is an Ontology*. URL: http://www-ksl.stanford.edu/kst/what- is-an-ontology.html.

Бакасова, П. С., Исраилова, Н. А.(2016). А*лгоритм образования словоформ для автоматизации процедуры пополнения базы данных словаря*. Бишкек: Известия КГТУ им. И. Раззакова.Т. 38. № 2. С. 23–27.

Исраилова, Н. А.(2012). *Организация морфологического анализа в трансляторах.* Усть-Каменогорск: Вестн. Вост.-Казахст. гос. техн. ун-та им. Д. Серикбаева.-.- №1.– С. 97–101.

Бакасова, П. С., Исраилова, Н. А.(2018). *Онтологическая модель морфологических правил кыргызского языка*. Бишкек: Мат. № 60 МНТК

http://protege.stanford.edu/ (accessed 10.09.2019).

Zhetkenbay L., Sharipbay A.,.Bekmanova G,.Kamanur U. (2016). Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages..Journal ofTheoretical and Applied Information Technology. Vol.91.N.2.2016. P. 257–263.

УДК 81’366.54

NOMINAL COORDINATION IN TATAR POSTPOSITIONAL CONSTRUCTIONS AND CASE VARIATION

Ekaterina Lyutikova [a,b]***, Anastasia Gerasimova*** [a,b]
[a]*Lomonosov Moscow State University, Moscow, Russian Federation*
[b]*Pushkin State Russian Language Institute, Moscow, Russian Federation*
lyutikova2008@gmail.com

The paper develops a formal model for case variation in Tatar postpositional constructions with coordinated noun phrases. In Tatar postpositional constructions, the choice of the case form of the dependent nominal is determined by its morphological and syntactic class. Moreover, for postpositions synchronically associated with relational nouns marking of the pronominal argument affects the make-up of the postpositional phrase itself and governs the choice between the non-possessive form of the postposition and the possessive agreement. The study of coordinated noun phrases allows us to establish syntactic mechanisms that determine this variation. In particular, configurations where two coordinated arguments belong to different morphological and syntactic classes make it possible to establish a relation between the choice of case marking of conjuncts and the choice of possessive agreement on the postposition.

Two linguistic experiments were conducted to investigate the interaction of the two types of variation. In experiment 1, respondents filled in gaps in sentences, choosing the most natural form for the components of the postpositional construction. In experiment 2, respondents evaluated sentences with different configurations of postpositional phrases on a Likert scale from 1 to 5. The results of experiments show that the first argument of the coordinated construction shows a clear tendency to be used in the unmarked (caseless) form. However, if the first argument is a 1st/2nd person pronoun, respondents choose the genitive in half of the cases. In addition, respondents prefer the 3rd person possessive agreement marker, regardless of whether one of the conjuncts is 3rd person or not, which can be considered as a default agreement pattern.

The experimental data reveal a number of syntactic properties that postpositional constructions share with nominal ezafe constructions. In particular, the correlation between the choice of the genitive case marking of the argument and the agreeing possessive form of the postposition, on the one hand, and the choice of the unmarked form of the argument and the non-agreeing possessive form of the postposition, on the other hand, reflects the opposition of ezafe constructions 3 and 2, respectively. Thus, the process of grammaticalization of denominal postpositions is not yet complete. Our results allow us to characterize the synchronic variability, as well as provide

data for improving morphological and syntactic processing for Tatar, so that they fully cover the actual linguistic behavior of native speakers.

Keywords: postpositional phrase, differential case marking, case, agreement, nominal coordination, Tatar, experimental syntax.

ИМЕННОЕ СОЧИНЕНИЕ В ТАТАРСКИХ ПОСЛЕЛОЖНЫХ КОНСТРУКЦИЯХ И ПАДЕЖНОЕ ВАРЬИРОВАНИЕ

***Е. А. Лютикова*[a,b], *А. А. Герасимова*[a,b]**
[a]*МГУ им. М. В. Ломоносова, Москва, Россия*
[b]*Гос. ИРЯ им. А.С. Пушкина, Москва, Россия*
lyutikova2008@gmail.com

В статье строится формальная модель для варьирования падежного оформления в татарских послеложных конструкциях с сочиненными именными группами. В татарских послеложных конструкциях выбор падежной формы зависимой именной синтагмы определяется ее морфолого-синтаксическим классом. Причем для послелогов, синхронно связанных с существительными, способы маркирования местоименного дополнения оказывают влияние на оформление самой послеложной группы – выбор непритяжательной формы послелога или притяжательного согласования. Исследование сочиненных именных групп позволяет установить синтаксические механизмы, которые определяют два названных варьирования. В частности, случаи, когда два аргумента сочинения принадлежат различным морфолого-синтаксическим классам, дают возможность установить связь между выбором падежного маркирования конъюнктов и выбором притяжательного согласования для послелога.

Для исследования взаимодействия двух типов варьирования было проведено два лингвистических эксперимента. В эксперименте 1 респонденты заполняли пропуски в предложениях, выбирая наиболее естественную форму составляющих послеложной конструкции. В эксперименте 2 респонденты оценивали предложения с различными конфигурациями послеложных конструкций по шкале Ликерта от 1 до 5. Результаты экспериментов показывают, что первый аргумент сочиненной конструкции демонстрирует явную тенденцию к употреблению в немаркированной форме. Тем не менее, если первым аргументом является личное местоимение 1-2 лица, в половине случаев носители выбирают для него форму генитива. Кроме того, носители отдают предпочтение притяжательному согласованию по 3 лицу, вне зависимости от того, демонстрирует ли значение 3 лица один из конъюнктов, что позволяет говорить о выборе дефолтной модели согласования.

Экспериментальные данные выявляют целый ряд синтаксических свойств, которые объединяют послеложные конструкции с именными изафетными конструкциями. В частности, корреляция между выбором генитива зависимого и согласуемой притяжательной формой послело-

га, с одной стороны, и выбором немаркированной формы зависимого и несогласуемой притяжательной формой, с другой стороны, отражает противопоставление изафетных конструкций 3 и 2, соответственно. Таким образом, процесс грамматикализации послеложных слов еще не завершен. Полученные результаты позволяют охарактеризовать вариативность в синхронном срезе языка, а также предоставляют материал для усовершенствования систем морфологической и синтаксической обработки текстов на татарском, чтобы они наиболее полно охватывали актуальное языковое поведение носителей.

Ключевые слова: послеложная группа; дифференцированное маркирование аргумента; падеж; согласование; сочинение; татарский язык; экспериментальный синтаксис.

1. Введение

В последние годы отмечается существенный рост интереса описательной и теоретической лингвистики к проблеме согласования с сочиненными конструкциями (см., например, (Bošković, 2009, 2010; Willim, 2012; Franks, Willer-Gold, 2014; Marušič, Nevins, Badecker, 2015; Nevins, Weisser, 2018, Citko, 2018)). Известно, что выбор согласовательной модели определяется многими факторами, среди которых, в частности, тип мишени согласования (глагол, адъектив, имя), препозиция / постпозиция мишени согласования по отношению к контролеру, линейное расстояние между контролером и мишенью, тип сочинения (симметричное / асимметричное). Выделяются следующие формальные типы согласования: разрешающее (resolved) согласование – по признакам сочиненной конструкции (*Петя и Маша пришли* (мн.ч.), *я и ты придем* (1-е л. мн.ч.)), по признакам первого конъюнкта, по признакам последнего конъюнкта[1]. Предполагается также возможность более сложных иерархий не только для вычисления признаков при разрешающем согласовании (1-е + 2-е лицо = 1-е лицо, мужск. род + женск. род = мужск. род), но и для выбора контролера согласования (например, контролером оказывается конъюнкт с более маркированным признаком лица: *я и он* / *он и я приду*).

В этой статье мы предполагаем исследовать татарскую симметричную сочиненную конструкцию с союзом *һәм* ‘и’ в контек-

[1] В (Nevins, Weisser, 2018) высказывается идея, что по меньшей мере часть случаев согласования с первым или последним конъюнктом в действительности представляют собой согласование с ближайшим конъюнктом, что может быть установлено, если позиция мишени не фиксирована относительно контролера грамматическими правилами.

сте послеложной группы. Своеобразие послеложной конструкции в татарском языке состоит в том, что имена разных морфолого-синтаксических классов выступают в послеложной конструкции в разных падежных формах: личные местоимения – в форме генитива, а существительные – в форме номинатива (основного падежа). В этой связи, если конъюнкты относятся к разным грамматическим разрядам, возникает проблема разрешения падежного конфликта в сочиненной конструкции: получает ли вся конструкция единое падежное оформление, или соответствующим падежом оформляется каждый компонент сочиненной конструкции. Дополнительные сложности возникают, если послеложную конструкцию образуют послеложные слова, имеющие позицию для согласования с зависимым именем. Выбор контролера согласования предположительно должен коррелировать со способом разрешения падежного конфликта. Таким образом, мы имеем возможность не только изучить согласование с сочиненной конструкцией в нетривиальном для данной проблематике контексте – послеложной группе, но и установить связь между внутренней структурой конструкции (единое падежное оформление / собственный падеж у каждого конъюнкта) и согласованием.

Дальнейшее изложение построено следующим образом. В разделе 2 мы описываем основные характеристики послеложных конструкций татарского языка. В разделе 3 представлены результаты двух экспериментов, в которых при помощи разных методик исследовался выбор стратегии оформления послеложной конструкции носителями татарского языка. В разделе 4 подводятся итоги обсуждения.

2. Послеложные конструкции в татарском языке

Послеложные конструкции в татарском языке образует гетерогенная категория лексических единиц, в разной степени грамматикализованных в качестве функциональных. Татарская грамматика (Закиев, 1993: 309; Закиев 1995: 47) выделяет собственно послелоги и послеложные слова, которые «...отличаются от послелогов тем, что имеют живые словообразовательные отношения и лексико-семантические связи с знаменательными частями речи...». Соответственно, послеложные слова имеют знаменательный эквивалент, обладающий лексическим значением: «В предложении *Язу өстендә озак утырды* ‘Долго сидел над письмом’ *өстендә* – послеложное слово; в предложении *Су өстендә көймә*

йөзә ‘Лодка плывет на поверхности воды’ то же самое *өстендә* – имя существительное» (Закиев 1993: 11).

Послелоги и послеложные слова управляют именной группой в определенной падежной форме. Так, например, послелог *таба(н)* ‘к’ и послеложное слово *күрә* ‘ввиду, из-за, по’ управляют дативом, а послелоги *башка* ‘кроме’ и *соң* ‘после’ – аблативом.

Большой класс послелогов и послеложных слов употребляются с именной группой в немаркированной форме, совпадающей с номинативом (татар. *баш килеш*) – падежом подлежащего. В этот класс входят как собственно послелоги, например, *белән* ‘с’, *өчен* ‘для’, так и послеложные слова, соотносимые с существительными с локативным или абстрактным значением, например, *өстендә* ‘над’ (ср. *өс* ‘верх’), *янына* ‘рядом, к’ (ср. *ян* ‘бок’), *урынында* ‘вместо’ (ср. *урын* ‘место’), *ярдәмендә* ‘благодаря’ (ср. *ярдәм* ‘помощь’). Зачастую послеложные слова образуют группы, объединенные общей основой и различающиеся падежными аффиксами, напр. *янына* ‘к’ (датив), *янында* ‘около’ (локатив), *яныннан* ‘мимо’ (аблатив); *артына* ‘за, вслед’ (датив), *артында* ‘за, позади’ (локатив), *артыннан* ‘из-за’ (аблатив).

Татарская грамматика (Закиев, 1993: 253) указывает, что с послелогами и послеложными словами этого класса местоимения-существительные используются в форме генитива. Таким образом, в послеложной конструкции возникает падежное варьирование: именные группы на основе существительного (или субстантивированного атрибута, отглагольной номинализации и т.п.) демонстрируют немаркированную форму, в то время как местоимения-существительные – форму генитива: «… *аның белән (абый белән) әшләү* ‘делать с ним (с братом)’, *аның өчен (дустым өчен) тырышу* ‘стараться ради него (друга)’». К местоимениям-существительным Татарская грамматика относит личные местоимения 1-2 лица *мин* ‘я’, *син* ‘ты’, *без* ‘мы’, *сез* ‘вы’, местоимения 3 лица *ул* ‘он’ и *алар* ‘они’, возвратное местоимение *үз* ‘сам’, вопросительные местоимения *кем* ‘кто’, *нәрсә* ‘что’, а также образованные на основе вопросительных местоимений серии неопределенных и универсальных местоимений. Дифференцированное маркирование дополнения в послеложной конструкции, таким образом, лицензируется формальным фактором – его морфосинтаксической категорией.

Отличие личных местоимений от прочих субстантивов проявляется также в том, что послеложные слова демонстрируют

лично-числовое (притяжательное) согласование с генитивным местоименным дополнением. В примерах (1a-b) показаны согласованные по лицу и числу с местоименным дополнением формы послелога *алдында* ‘перед’; в (1c) – тот же послелог в дефолтной форме 3 лица.

(1) a. *Юк, сез-нең* ***минем*** *алд-**ым**-да*
нет вы-GEN я.GEN перед-1SG-LOC
бер гаеб-егез дә юк.
один вина-2PL EMPH нет
‘Нет, вы передо мной ни в чем не виноваты.’ [TT][11]

b. *Әфәнде-ләр, хәзер мин* ***сез-нең****алд-**ыгыз**-да*
господин-PL сейчас я вы-GEN перед-2PL-LOC
бик зур эш ача-чак-мын.
очень большой дело открывать-FUT-1SG
‘Господа, сейчас я открою вам (=перед вами) одну очень важную вещь.’ [TT]

c. ***Кыз-лар*** *алд-**ын**-да ясалма*
девушка-PL перед-3-LOC искуственно
йөр-гән-не ярат-м-ый-м.
ходить-PART-ACC любить-NEG-PRS-1SG
‘Не люблю выпендриваться (= лживо ходить) перед девушками.’ [TT]

Причиной вариативной структуры конструкций с послеложными словами является, безусловно, их диахронический источник. Послеложная конструкция представляет собой результат грамматикализации именной синтагмы, возглавляемой локативным или абстрактным существительным в одной из падежных форм. Зависимое в послеложной конструкции, соответственно, является приименным зависимым и образует с вершиной посессивную конструкцию.

Посессивная конструкция в татарском языке также демонстрирует дифференцированное маркирование аргумента-посессора. Традиционная грамматика выделяет два типа посессивных конструкций – изафетную конструкцию 2 и изафетную конструкцию

[1] Примеры с пометой [TT] получены из Татарского национального корпуса «Туган тел» (http://www.tugantel.tatar/).

3[1]. В изафетной конструкции 2 зависимое выступает в немаркированной форме, вершина несет на себе изафетный (посессивный) показатель (2a). В изафетной конструкции 3 зависимое имеет форму генитива, вершина несет на себе изафетный показатель (2b).

(2) a. *укычы* *дәфтәр-е*
ученик тетрадь-3
‘ученическая тетрадь, тетрадь ученика’
b. *укычы-ның* *дәфтәр-е*
ученик-GEN тетрадь-3
‘тетрадь ученика’

Татарская грамматика (Закиев, 1993: 32-37) отмечает, что местоимения-существительные образуют только изафетную конструкцию 3, но не 2, ср. (3a-b); если посессором выступают местоимения 1-2 лица, в разговорной речи возможно опущение изафетного показателя в изафетной конструкции 3 (3c). Структуры, в которых изафетный показатель на вершине возникает в отсутствие выраженного генитивного посессора, естественно рассматривать как изафетную конструкцию 3, в которой в позиции посессора находится нулевое местоимение *pro* соответствующего лица и числа, контролирующее согласование изафетного показателя (3d).

(3) a. *без-нең* *мәктәб-ебез*
мы-GEN школа-1PL
‘наша школа’
b. * *без* *мәктәб-ебез*
мы школа-1PL
c. *без-нең* *мәктәп*
мы-GEN школа
‘наша школа’
d. [*pro*] *мәктәб-ебез*
pro.1PL.GEN школа-1PL
‘наша школа’

[1] В изафетной конструкции 1, образованной соположением двух существительных (таш йорт ‘каменный дом’), зависимое не может ветвиться, присоединять показатели числа и принадлежности, то есть является вершиной. Ее свойства нерелевантны для обсуждения послеложной конструкции.

Еще один класс посессоров, которые возможны только в изафетной конструкции 3, но не 2, – это именные группы, сами представляющие собой изафетную конструкцию 3. Это, во-первых, собственно посессивные конструкции, такие как в (3), а также субстантивированные кванторные и атрибутивные конструкции, такие как в (4):

(4) a. *(а-лар-ның)* *күб-есе*
он-PL-GEN много-3
‘многие из них’

b. (без-нең) иң акыл-лы-быз
мы-GEN самый ум-ATR-1PL
‘самый умный из нас’

c. кыз-лар-ның кайсы-сы
девушка-PL-GEN который-3
‘которая из девушек’

Остальные типы именных групп могут быть зависимыми в обеих изафетных конструкциях. Выбор между генитивом и немаркированной формой определяется как определенностью зависимой именной группы, так и семантическим отношением между вершиной и зависимым. Так, например, имена собственные обычно выступают в изафетной конструкции 3 (*Марат**(*-ның*) *ата-сы* ‘отец Марата’), однако при обозначении наименования используется изафетная конструкция 2 (*Марат урам-ы* ‘улица Марата’).

В (Pereltsvaig, Lyutikova, 2014; Лютикова, Перельцвайг, 2015; Lyutikova, Pereltsvaig, 2015; Lyutikova, 2017) показано, что различия между двумя конструкциями не сводятся к падежному оформлению, но затрагивают различные характеристики посессоров – их структурную позицию в именной группе, их собственный категориальный статус, возможные интерпретации и способность к выражению различных тематических отношений с именной вершиной. Указанные кластеры свойств возникают не случайно, но выводятся из категориального статуса именной группы-посессора. DP-посессоры насыщают аргументные позиции, выражают тематические отношения, получают конкретно-референтную или обобщенно-кванторную интерпретацию, получают падеж, контролируют посессивное согласование показателя изафета и располагаются в крайней левой позиции в именной группе (Spec, DP). По-

сессоры малой структуры вводятся особой функциональной вершиной Poss, выражают широкий спектр отношений между двумя именными группами, уточняемых на основе энциклопедических знаний, имеют предикатную интерпретацию, не нуждаются в падеже, неспособны контролировать посессивное согласование и располагаются в своей базовой позиции, правее атрибутивных модификаторов (Spec, PossP). Таким образом, дифференцированное маркирование посессора в татарском языке существенно отличается от дифференцированного маркирования дополнения послелога: в послеложной конструкции генитивом оформляются только местоимения-существительные, в то время как в посессивной конструкции – любые именные составляющие, имеющие статус DP, в том числе – изафетная конструкция 3, имена собственные, другие определенные именные группы. Ср. примеры в (5)-(6), демонстрирующие послеложные группы, и (7) с именными группами.

(5) a. *минем / *мин өчен*
я.GEN / я для
'для меня'

b. *ата-м / *ата-м-ның өчен*
отец-1SG / отец-1SG-GEN для
'для моего отца'

c. *Марат / *Марат-ның өчен*
Марат / Марат-GEN для
'для Марата'

(6) a. *минем / *мин урын-ым-да*
я.GEN / я вместо-1SG-LOC
'вместо меня'

b. *ата-м / *ата-м-ның урын-ын-да*
отец-1SG / отец-1SG-GEN вместо-3-LOC
'вместо моего отца'

c. *Марат / *Марат-ның урын-ын-да*
Марат / Марат-GEN вместо-3-LOC
'вместо Марата'

(7) a. *минем / *мин мәктәб-ем-дә*
я.GEN / я школа-1SG-LOC
'в моей школе'

b. **ата-м / ата-м-ның мәктәб-ен-дә*
отец-1SG / отец-1SG-GEN школа-3-LOC
'в школе моего отца'

c.	**Марат / Марат-ның*	*мәктәб-ен-дә*
	Марат / Марат-GEN	школа-3-LOC
	‘в школе Марата’	

Таким образом, послелоги и послеложные слова, присоединяющие именные группы в немаркированной форме или генитиве, могут стать источником особенно интересного материала для изучения вариативности в оформлении сочиненной конструкции. Во-первых, грамматические свойства конструкции включают в себя выбор между генитивным или немаркированным оформлением зависимой именной группы. Во-вторых, сосуществование в языке послеложных слов и омонимичных им существительных, деривационная связь между которыми, по свидетельству авторов Татарской грамматики, ощущается носителями татарского языка, позволяет предположить, что способы оформления именной синтагмы будут оказывать влияние и на оформление послеложной группы. Случаи, когда два аргумента сочинения принадлежат различным морфолого-синтаксическим классам, дают возможность установить связь между выбором падежного маркирования конъюнктов и выбором притяжательного согласования для послелога.

Для проверки этих предположений было проведено экспериментальное исследование управления послелогов и послеложных слов с сочиненной конструкцией в качестве аргумента. В следующем разделе мы опишем эксперименты и полученные нами результаты.

3. Экспериментальное исследование

Для исследования оформления послеложной конструкции с сочиненной группой в качестве аргумента было проведено два лингвистических эксперимента. В первом эксперименте респонденты заполняли пропуски в предложениях, выбирая наиболее естественную форму составляющих послеложной конструкции. Зависимое и послелог находились в скобках в словарной форме (8). В этом эксперименте мы изучали конструкции с сочинением имени собственного и личного местоимения первого лица (пр. *мин һәм Марат* ‘я и Марат’). Дизайн эксперимента подразумевал два фактора: фактор «тип вершины в послеложной конструкции» (два уровня: послелоги и послеложные слова), а также фактор «порядок конъюнктов в сочиненной конструкции» (2 уровня: лич-

ное местоимение предшествует имени собственному или следует за ним).

(8) Танылган җырчы (син һәм Әхмет, янәшәсендә) ___ ___ ___ ___ утырды.
‘Известный певец сидел рядом с тобой и Ахметом.’

В эксперименте участвовало 109 респондентов (средний возраст 23; SD = 7; мин. возраст 17, макс. возраст 61; 85 женщин и 23 мужчины). Чтобы установить уровень владения татарским языком перед началом эксперимента мы просили респондентов ответить на вопросы социолингвистической анкеты. Данные анкетирования показали, что 90 из 109 респондентов родились и проживают в Республике Татарстан и ежедневно общаются по-татарски в семье и в месте учебы или работы.

Результаты эксперимента на заполнение пропусков были обработаны с помощью логлинейного анализа, который показал значимое взаимодействие экспериментальных факторов ($p = 0{,}001$). Результаты эксперимента показывают, что в подавляющем большинстве случаев имя собственное имеет немаркированную форму вне зависимости от позиции относительно личного местоимения. При этом личное местоимение значимо часто маркируется генитивом, также вне зависимости от своей позиции в сочиненной конструкции. Ситуация, в которой имя собственное получает генитив, а личное местоимение имеет немаркированную форму, практически недопустима, при этом маркирование двух конъюнктов генитивом возможно.

Таблица 1. Результаты эксперимента на порождение

	NOM NOM	NOM GEN	GEN NOM	GEN GEN
Марат и я	41	335	0	13
я и Марат	180	7	142	25

Таблица 2. Результаты эксперимента на порождение для послелогов без позиции для лично-числового согласования

	NOM NOM	NOM GEN	GEN NOM	GEN GEN
Марат и я	19	179	0	0
я и Марат	116	0	39	0

Таблица 3. Результаты эксперимента на порождение для послеложных слов

Марат и я	NOM NOM	NOM GEN	GEN NOM	GEN GEN
без согласования	1	36	0	3
контроль со стороны первого конъюнкта / дефолтное согласование	21	111	0	9
контроль со стороны второго конъюнкта	0	7	0	1
контроль со стороны сочиненной конструкции	0	2	0	0
я и Марат	NOM NOM	NOM GEN	GEN NOM	GEN GEN
без согласования	0	0	0	0
контроль со стороны первого конъюнкта / дефолтное согласование	0	0	0	0
контроль со стороны второго конъюнкта	64	7	102	25
контроль со стороны сочиненной конструкции	0	0	1	0

Экспериментальные данные для послеложных слов позволяют утверждать, что носители чаще всего выбирают притяжательное согласование по 3 лицу. Эта форма, с одной стороны, может означать контроль со стороны имени собственного, а с другой – выбор стратегии дефолтного согласования. Чтобы разрешить данную омонимию, а также подробнее выяснить предпочтения носителей татарского языка мы провели второй эксперимент, в котором наравне с рассмотренными сочиненными конструкциями были исследованы сочиненные конструкциями с двумя личными местоимениями.

Во втором эксперименте респонденты оценивали предложения с различными конфигурациями послеложных конструкций по шкале Ликерта от 1 до 5. В этом эксперименте к конструкциям с сочинением из первого эксперимента мы добавили конструкции с сочинением двух личных местоимений (*мин һәм син* ‘я и ты’), причем эти конструкции использовались только с послеложными словами. Также добавился фактор «ПАДЕЖНОЕ МАРКИРОВАНИЕ» (4 уровня: оба конъюнкта маркированы генитивом; оба конъюнк-

та маркированы номинативом; первый конъюнкт маркирован генитивом, второй – номинативом; первый конъюнкт маркирован номинативом, второй – генитивом). Кроме того, на каждое послеложное слово теперь приходилось до четырех форм: непритяжательная, притяжательная согласуемая для субстантивов с граммемой 1-2 лица едиснтвенного, непритяжательная несогласуемая, притяжательная согласуемая для субстантивов с граммемой 1 лица множественного числа (в случае контроля согласования со стороны сочинительной конструкции). В эксперименте участвовало 38 респондентов (средний возраст 24; SD = 8; мин. возраст 17, макс. возраст 62; 30 женщин и 8 мужчин); из них 31 респондент участвовал также в первом эксперименте.

Результаты эксперимента показывают, что в случае конструкции с именем собственным и личным местоимением при послелогах наиболее высокие оценки получают стратегии маркирования, при которых имя собственное имеет немаркированную форму, а личное местоимение либо также имеет немаркированную форму, либо демонстрирует форму генитива (*t-критерий Стьюдента*, $p << 0{,}001$). В случае послеложных слов наибольшие оценки получает стратегия, при которой имя собственное имеет немаркированную форму, а личное местоимение – форму генитива.

Для сочиненных конструкций с двумя личными местоимениями наиболее предпочтительной оказалась стратегия падежного маркирования, при которой оба местоимения получают генитив. При этом различные варианты лично-числового согласования имеют равный уровень приемлемости: так, равно допустимы отсутствие согласования, согласование по третьему лицу, согласование с первым конъюнктом и согласование с сочиненной конструкцией. Обратим внимание, что наименее приемлемым оказалось согласование со вторым конъюнктом. Это наблюдение позволяет по-новому интерпретировать результаты эксперимента на порождение. Получается, порядок конъюнктов оказывается значимым для выбора контроллера согласования: носители отдают предпочтение согласованию с первым конъюнктом. Следовательно, видимый контроль согласования со стороны второго конъюнкта для сочиненных конструкций с именем собственным и личным местоимением на самом деле является реализацией дефолтного согласования по третьему лицу. Таким образом, результаты двух экспериментов позволяют заключить, что носители отдают предпочтение притяжательному согласованию по третьему

лицу, вне зависимости от того, демонстрирует ли значение третьего лица один из конъюнктов, что говорит о выборе дефолтной модели согласования.

Таблица 4. Результаты эксперимента на оценку для сочинения имени собственного и личного местоимений (ненормализованные и нормализованные оценки)

		NOM NOM	NOM GEN	GEN NOM	GEN GEN
Послелоги без позиции для лично-числового согласования	*Марат и я*	1,5	**3,84**	1,32	2,08
	я и Марат	**3,05**	2,11	**3,11**	1,84
Послеложные слова	*Марат и я*	2,18	**2,58**	1,82	1,95
	я и Марат	1,84	2,21	1,84	1,79
Послелоги без позиции для лично-числового согласования	*Марат и я*	-0,7	**0,86**	-0,79	-0,29
	я и Марат	**0,38**	-0,21	**0,45**	-0,39
Послеложные слова	*Марат и я*	-0,19	**0,07**	-0,43	-0,33
	я и Марат	-0,44	-0,13	-0,43	-0,44

Таблица 5. Результаты эксперимента на оценку для сочинения двух личных местоимений (ненормализованные и нормализованные оценки)

я и ты	NOM NOM	NOM GEN	GEN NOM	GEN GEN
Без согласования	1,42	2,08	1,89	**2,89**
Второй конъюнкт	1,63	1,79	1,53	1,55
Дефолтное (3Sg)	1,71	1,66	2	**3,05**
Первый конъюнкт	1,71	1,55	1,53	**3**
Сочинение	1,34	1,97	1,76	**3,5**
я и ты	NOM NOM	NOM GEN	GEN NOM	GEN GEN
Без согласования	-0,69	-0,21	-0,36	**0,29**
Второй конъюнкт	-0,54	-0,47	-0,63	-0,7
Дефолтное (3Sg)	-0,53	-0,54	-0,3	**0,37**
Первый конъюнкт	-0,48	-0,6	-0,6	**0,38**
Сочинение	-0,79	-0,32	-0,42	**0,73**

4. Выводы

Таким образом, результаты экспериментов показывают, что первый аргумент сочиненной конструкции демонстрирует явную тенденцию к употреблению в немаркированной форме. Тем не менее, если первым аргументом является личное местоимение 1-2 лица, в половине случаев носители выбирают для него форму генитива. Кроме того, носители отдают предпочтение притяжательному согласованию по 3 лицу, вне зависимости от того, демонстрирует ли значение 3 лица один из конъюнктов, что позволяет говорить о выборе дефолтной модели согласования.

Экспериментальные данные выявляют целый ряд синтаксических свойств, которые объединяют послеложные конструкции с именными изафетными конструкциями. В частности, корреляция между выбором генитива зависимого и согласуемой притяжательной формой послелога, с одной стороны, и выбором немаркированной формы зависимого и несогласуемой притяжательной формой, с другой стороны, отражает противопоставление изафетных конструкций 3 и 2, соответственно. Таким образом, процесс грамматикализации послеложных слов еще не завершен. Полученные результаты позволяют охарактеризовать вариативность в синхронном срезе языка, а также предоставляют материал для усовершенствования систем морфологической и синтаксической обработки текстов на татарском, чтобы они наиболее полно охватывали актуальное языковое поведение носителей.

Благодарности

Исследование выполнено в рамках проекта РНФ № 18-18-00462 «Коммуникативно-синтаксический интерфейс: типология и грамматика», реализуемого в Гос. ИРЯ им. А.С. Пушкина. Авторы выражают искреннюю благодарность Д. А. Зариповой и А. М. Галиевой за неоценимую помощь в подготовке материалов для экспериментального исследования.

БИБЛИОГРАФИЯ

Лютикова, Е. А., & Перельцвайг, А. М. (2015). Структура именной группы в безартиклевых языках: универсальность и вариативность. *Вопросы языкознания, 3*, 52–69.

Закиев, М. Ф. (1993). *Татарская грамматика. Т. 2: Морфология.* Казань: изд-во Казанского гос. ун-та.

Закиев, М. Ф. (1995). *Татарская грамматика. Т. 3: Синтаксис.* Казань: изд-во Казанского гос. ун-та.

Bošković, Ž. (2009). Unifying first and last conjunct agreement. *Natural Language and Linguistic Theory*, 27, 455–496. DOI: https://doi.org/10.1007/s11049-009-9072-6

Bošković, Ž. (2010). Conjunct-sensitive agreement: Serbo-Croatian vs Russian. In G. Zybatow, P. Dudchuk, S. Minor & E. Pshehotskaya (Eds.), *Formal studies in Slavic linguistics: Proceedings of formal descriptions of Slavic languages,* 7(5) (pp. 31–48). Frankfurt am Main: Peter Lang.

Citko, B. (2018). Complementizer agreement with coordinated subjects in Polish. *Glossa: a journal of general linguistics*, 3(1), 124. DOI: https://doi.org/10.5334/gjgl.588

Franks, S., & Willer-Gold, J. (2014). Agreement strategies with conjoined subjects in Croatian. In J. Witkoś & S. Jaworski (Eds.), *New insights into Slavic linguistics* (pp. 91–115). Frankfurt am Main: Peter Lang. DOI: https://doi.org/10.3726/978-3-653-04359-4

Lyutikova, E. (2017). Agreement, case and licensing: Evidence from Tatar. *Урало-алтайские исследования,* 25:2, 25–45.

Lyutikova, E., & Pereltsvaig, A. (2015). The Tatar DP. Canadian Journal of Linguistics, 60:3, 289–325.

Marušič, F. L., Nevins, A., & Badecker, W. (2015). The grammars of conjunction agreement in Slovenian. Syntax, 18, 39–77. DOI: https://doi.org/10.1111/synt.12025

Nevins, A. & Weisser, P. (2018). Closest conjunct agreement. *Annual Review of Linguistics,* AA, 1–25.

Pereltsvaig A., & Lyutikova E. (2014) Possessives within and beyond NP: Two ezafe–constructions in Tatar. In by A. Bondaruk, G. Dalmi & A. Grosu (Eds.), *Advances in the syntax of DPs: Structure, agreement, and case* (pp. 193–219). Amsterdam: Benjamins.

Willim, E. (2012). Concord in Polish coordinate NPs as agree. In M. Ziková & M. Docekal (Eds.), *Slavic languages in formal grammar: Proceedings of FDSL 8.5, Brno 2010* (pp. 233–253). Frankfurt am Main: Peter Lang.

УДК 372.881.1

SOME APPROACHES TO THE ASSESSMENT OF AZERBAIJAN LANGUAGE PROFICIENCY

Pirdas H. Muradova
Institute of Information Technology of ANAS, Azerbaijan, Baku
pirdas.davudova@gmail.com

One of the most important issues of the modern period is the acquisition of language skills according to all rules. The article is devoted to studying the current station and problems of the Azerbaijani language. Also, the learning and assessment of the Azerbaijani language have been investigated. At the same time, have been analyzed some approaches to language learning and assessment. In a time when hegemonic languages like the English language were suppressing the languages of the states, some proposals have been made for protection of the Azerbaijani language and for its recognition in the international arena. Due to the influence of ICT in the globalizing world there are opportunities for the use of the Azerbaijani language in an electronic world. The perspectives of the application of information and communication technologies in language learning are demonstrated. The main purpose of this paper is to give suggestions for learning, teaching and assessment of the Azerbaijani language based on the international standard.

Keywords: Azerbaijani language, language standard, language skills, assessment, teaching, learning, language ability, ICT.

НЕКОТОРЫЕ ПОДХОДЫ К ОЦЕНКЕ УРОВНЯ ВЛАДЕНИЯ АЗЕРБАЙДЖАНСКИМ ЯЗЫКОМ

Пирдас Мурадова
Институт информационных технологий, Национальная академия наук Азербайджана, Азербайджан, Баку
pirdas.davudova@gmail.com

Одним из важных вопросов современного периода является приобретение языковых навыков в соответствии с правилами того или иного языка. Статья посвящена проблемам современного положения азербайджанского языка. Были рассмотрены вопросы и подходы к изучению и оценке азербайджанского языка. Так как влияние языков международного общения, таких как английский язык, на национальные языки является огромным, в статье приведены некоторые предложения по защите азербайджанского языка и признанию его на международной арене. В глобализирующемся мире благодаря влиянию информационных компьютерных технологий создаются возможности для использования азербайджанского языка в информационном простран-

стве, а также применение их для изучения языка. Основная цель данной статьи – представить новые подходы к изучению, преподаванию и оценке азербайджанского языка на основе международных стандартов.

Ключевые слова: азербайджанский язык; языковой стандарт; языковые навыки; оценка; преподавание; изучение; языковая способность; информационные и компьютерные технологии.

1. Introduction

In a globalized world, influenced by the Information and Communications Technology (ICT), when the economy, culture, science and technology are improving, there have been creating opportunities for the use and development of the Azerbaijani language. One of the reasons for the relevance of issues related to the Azerbaijani language at the time when hegemonic languages such as English squeezed the languages of the nations is to make certain moves for the protection of the Azerbaijani language and at the same time to be known internationally. In the article is studied some issues related to the Azerbaijani language, the reasons for the violation of the rules and styles of the literary language, as well as language training and assessment. There also were proffered recommendations for the teaching of the Azerbaijani language to foreign immigrants, the assessment and certification of Azerbaijani language proficiency as a whole using ICT.

2. The current situation and problems in the Azerbaijani language

In recent years, some reforms have been carried out in the education system of Azerbaijan, and the most important result is the National Curriculum, which is a new educational concept. The philosophy of the National Curriculum formed on new standards are based on the activeness of the language learner, aimed at the formation of an independent, creative and free-thinking personality (Milli Kurrikulum, 2006). However, if today Azerbaijani citizens' knowledge of the native language is evaluated according to international standards, the results may be not as expected. Today, linguistics of Azerbaijani language not only meets problems during scientific research, but it also faces language problems in the linguistic community. One of the issues that are harmful to Azerbaijani linguistics is the use of foreign words in the names of service industries, advertisements and posters in the

country. For example, often used foreign words such as boutique, palace, saloon, etc. It also leads to a violation of the literary language of Azerbaijan. Even though the state language is Azerbaijani, giving more preference to foreign language proficiency in many government agencies and it affects the compression of the Azerbaijani language by other languages.

First of all, for the protection and recognition promotion of the Azerbaijani language between many European languages, the approach to educating our language must be changed. If we study the results of teaching the Azerbaijani language today, we will see that the level of language skills in the country is very low. Researches show that one of the key states is the use of new approaches to improve the level of language skills in the country. In a global environment, different approaches are recommended for the improvement of proficiency of the Azerbaijani language. Of course, these methods are based principally on European studies of language learning.

The first approach is related to the science of linguistics in Azerbaijan, where it is proposed to study the language not as humanities, but as a social science. In general, language is a part of social society, that is, a means of communication between people in society. If there is a dialogue between people, there is language and society is unimaginable without speech. So, language is the source of the formation of a social community. At the same time, language is associated with all spheres of social life and science. For this reason, first of all, linguistics in Azerbaijan should be studied as social science and not as humanities that is why the language should be separated from literature.

One of the most important issues in the context of the globalization of our modern age is the acquisition of language skills through all the rules of the language. Generally, the reason for the difficulties and problems in the use of written or spoken language is that the language is not properly taught and inadequately evaluated. Another approach is to change the philosophy of language teaching, that is, teaching language skills rather than linguistics in the classroom. For providing efficiency of communication between the native citizens of Azerbaijan, in other words, for each member of society to be able to easily receive information from others and convey their feelings and thoughts to others, they must own four skills of the language. Receiving and transmitting information occurs both in written and oral forms. By listening and reading information is receiving, by speaking and writing

information is transmitted. Thus, when communicating in a language, four language skills are used: listening, reading, speaking and writing. Therefore, language teaching should be built on these skills. It is also necessary to pay special attention to the development of textbooks and content. According to Western educators, anyone with a high level of language skills can succeed in any field of science. However, in Azerbaijan language teaching is not concentrated on language skills, but on teaching linguistics. For this reason, each skill must be taken into account when teaching the language.

In another approach, it is suggested that language teaching topics should be separated from literature and shaped according to real-world requirements. Today, Azerbaijani language education is not designed to facilitate communication for the citizen in the community. As mentioned earlier, the language should be separated from the literature. While explaining any rules of the language should not be given examples from works of poets or writers, rather, examples should be based on real-life situations. Because both of written and oral practice of the language should be formed on the realities of life. One of the main goals in teaching modern languages is to develop the communication skills of young people so that they can effectively use the language in real-life situations and see that acquiring excellent language skills can give them valuable social and working skills.

At the same time, language learning becomes easier if language learning is carried out by the levels, that is, teaching according to the needs of any age or social status (pupil, student, etc.). This approach provides for a staged study of the language from simple to compound. Step-by-step language learning from easy to hard plays an important role in better acquisition and retention of language knowledge. In other words, as you move from the beginning to a more advanced level and, as it progresses, the learner's motivation and interest for language learning increases as he sees his success. It is suggested that the division of language competence by levels is required. After the level division is carried out, there should be defined descriptions of the language skills corresponding to each level. As an example of for this, the Council of Europe introduced to the world «Common European Framework of Reference for Languages: Learning, Teaching, Assessment» (Council of Europe, CEFR, 2001). Of course, textbooks and additional tutorials (audio-video lessons, special supplementary programs, etc.) should be prepared following the requirements of each level.

It should be noted that language problems in the republic exist not

only in teaching but also in assessing language skills. For example, when applying for a job in the country, in the language skills section request information about our language skills based on four skills: reading, speaking, writing and listening. Principally, all four skills in this section are evaluated as excellent. However, no sources or documents confirming this assessment are provided. This is because there is no system in the republic to test the level of proficiency of the Azerbaijani language based on four skills (reading, writing, listening and speaking).

If the teaching of the Azerbaijani language is based on four skills, as well as textbooks and additional resources prepared on this principle, assessment of the language can also be organized on these four skills. It seems that if we can speak and write in the Azerbaijani language, we have excellent language skills. In public life, this is probably true. However, the requirements for education and the business community are higher. Assessment should also help increase the interest, motivation and desire of language learners to develop language skills in extracurricular areas. Through the assessment of the modern Azerbaijani language, children and young people or teenagers can learn a language and see the development of their skills as listening, speaking, reading and writing and making progress in the use of these skills in real life. The main principle in the assessment of language proficiency should be based on the teaching language by the levels. That is, as a result of the assessment, each language learner can determine which level of division a person belongs to.

Generally, modern language assessment implies the ability of children and young people to expand their vocabulary and use of vocabulary, to develop their understanding of written and oral form of words, language structure and rules, and to apply them all in real social life, education and work. In many foreign countries, one of the main prerequisites for obtaining an educational, immigration or work visa is the acquisition of language skills and an internationally recognized certificate (IELTS, TOEFL, etc.) for a short period. For example, IELTS is one of the prerequisites for applying for a visa and immigration services for the UK. IELTS (International English Language Testing System) is a high-level English language test for training, migration and work. There are two types of IELTS: general and academic IELTS. The IELTS General Exam is designed for those who register for higher education, work experience, study programs, or residence permits. Academic IELTS is an exam for those who

want to get higher education and professional work. These two types of exam systems are built on 4 skills: reading, listening, writing and speaking. Countries that require IELTS for migration include the United Kingdom, Australia, New Zealand, and Canada. Academic IELTS is a prerequisite for admission to recognized universities in the USA, Canada, Great Britain, Australia and even non-English-speaking countries (Belgium, Finland, Germany, France, Japan, gin, Korea, etc.) (IELTS, www.ielts.org). In general, exams in many languages can be given as a sample, for example, TOEFL (English), ESOL (English), DaF (German), CCE (Czech), TOCFL (Chinese), TRKI (Russian), TYS (Turkey), etc (Wikipedia, List of language proficiency).

The language work should be carried out at the national level for the formation of the language policy in the country, further increasing the prestige of the Azerbaijani language and its international recognition. The migration of foreigners from foreign countries to Azerbaijan has increased in recent years. Hundreds of foreign citizens come to our country every year to study and establish trade relations. It is necessary to organize the teaching of Azerbaijani language for introducing them Azerbaijani culture, make easily adapt to our community and at least to overcome language difficulties during the education of foreign students. So, it is intended to conduct language teaching in two directions. In other words, there are two objects of Azerbaijani language learning:

1. Citizens of Azerbaijan (national educational model)

2. Non-Azerbaijani citizens or foreigners (language teaching model for foreigners)

The reason for two objects of language teaching can be explained by the fact that, firstly, both target objects do not need the same level of language skills. That is, Azerbaijani language proficiency expected from Azerbaijani citizen is higher than for foreigners. The issue of teaching Azerbaijani to migrants has already been paid at the state level. President of Azerbaijan Ilham Aliyev has declared order in 2018 for «Safeguard the Azerbaijani language and to improve the use of the state language». The decree contains the questions as, «Increasing the use of the Azerbaijani language in the electronic space, facilitating access and learning for those interested in this language» (2018, president.az). One of the difficulties about the Azerbaijani language is that there is a minority of Internet resources, electronic and interactive textbooks in this language. Today, the researcher or reader often has to make use of physical resources (books, scientific journals) to get information related to the language.

Today, a unified exam system should be established to assess the language skills of each person (4 skills). This exam system can be used for many purposes: during entrance examinations in higher educational systems, for employment, and so on. This exam system can be used for many purposes: during entrance exams for higher education, for employment, and so on. A proper assessment can determine how well children and young people can apply these skills in education, everyday life and the business world.

3. The role of ICT in teaching and promoting the Azerbaijani language

In general, assessment in modern Azerbaijani language will focus on children and young people's progress in developing and applying their skills in listening, speaking, reading and writing. To solve all these problems, information and communication technologies should be used. One of the most important issues in language teaching is the involvement of ICTs. World experience shows that ICT opportunities can be successfully applied in both language teaching and language assessment. ICT for the first time began to be used for learning and teaching foreign languages in the 60s of the 20th century (Wikipedia, Educational Technology). In general, the most important achievements of ICT in teaching linguistic research are as follows (Ghasemi B., Hashemi M., 2011):

- *increase the quality of learning and teaching*
- *ease of access to a very high volume of information and knowledge available in the world*
- *rapid and timely access to information in very little time*
- *reduction of some educational expenses*
- *improve the quality, accuracy and scientific texts for academic disciplines*
- *indirect creation of learning experiences*
- *create an exact relationship*
- *create an interest in learning*
- *increase learning opportunities*
- *educators can evaluate students, they have collected*
- *the necessary information and appropriate feedback to students are presented* (Ghasemi B., Hashemi M., 2011).

Modern teaching methods using ICTs can encourage language learners to be more creative in language learning, develop their

ability to discuss and judge, while at the same time providing a more enthusiastic and enthusiastic approach to the learning process. The use of ICT in language teaching, as in other areas of learning, contributes to better language learning, which means achieving better results by simplifying the learning process. The role of the teacher has already changed as a result of the innovations in ICT teaching language. Previously, the teacher was the only source of information, now playing the role of an instructor helps the student develop skills such as selecting, receiving, evaluating, and remembering information. However, it should be noted that ICT is not a new teaching method, but a tool for implementing new methods.

Conclusion

The following suggestions were made for the protection of the state language status of the Azerbaijani language, the organization of language training for foreigners for the recognition of the Azerbaijani language in the international community and, in general, using of European experience (CEFR) in teaching and assessment of the Azerbaijani language:

1. Modern Azerbaijani language teaching should be conducted in two directions (for Azerbaijani citizens and foreigners);
2. Develop a four skill level based curriculum for both directions and the development of textbooks and additional teaching aids (audio, video, software, etc.) based on this curriculum;
3. Development of new evaluation criteria based on international standards (by levels);
4. Formation of the modern Azerbaijani language level examination system (at the same time creating a «Foreign Language Examination Test» for foreigners);
5. Suggestions for areas where you can use the Azerbaijani language exam:
   - when applying for higher education by foreign students;
   - during employment a government agency;
   - during employment for radio and television;
   - at the request of migrants for study or living in Azerbaijan.
6. Creation of an electronic database supported by ICT for the implementation of the above proposals, as well as additional resources for acquiring language skills;

The article also emphasizes the importance of ICT in the

development of new teaching materials based on foreign experience, that is, the development of separate textbooks and additional teaching materials at each level and language teaching.

REFERENCES

Azərbaycan Respublikasında Təhsil Konsepsiyası (Milli Kurrikulum), 2006, http://portal.edu.az/index.php?r=article/item&id=194&ut=&lang=az

Council of Europe, Common European Framework of References for Languages: Learning, Teaching, Assessment, 2001, 260p. www.coe.int/lang-CEFR

www. ielts.org

https://en.wikipedia.org/wiki/List_of_language_proficiency_tests

«Azərbaycan dilinin saflığının qorunması və dövlət dilindən istifadənin daha da təkmilləşdirilməsi ilə bağlı tədbirlər haqqında» Azərbaycan Respublikası Prezidentinin Fərmanı (01 noyabr 2018) http://www.prezident.az.

https://en.wikipedia.org/wiki/Educational_technology

Ghasemi B., Hashemi M., ICT: Newwave in English language learning/teaching. Procedia Social and Behavioral Sciences 15, 2011, pp. 3098–3102

УДК 81’33

A SYSTEMATIC ANALYSIS INTO THE MORPHOTACTICS OF PERSON, QUESTION AND MULTIPLE TENSES IN TURKISH

Berke Özenç, Ercan Solak
Işık University, İstanbul, Turkey
ercan.solak@isikun.edu.tr

In Turkish verb paradigm, Tense, Person and Question (TPQ) morphemes combine in interesting ways. While the rest of the Turkish morphotactics is regular, TPQ sub-paradigm is comparatively quite complex. Although there have been a few studies in the literature to analyze patterns in their restricted frames, so far no systematic account has been attempted. In this paper, we present a complete analysis starting from a theoretically full matrix of morphotactic orders and systematically eliminate forbidden sequences to arrive at a computational model of TPQ. We present the final model as a finite-state transducer which can be used in larger morphological analyzers. Our systematic analysis covers all the combinations of base Tenses, 4 copula Tenses, Person paradigms and the Question particle. In addition to yielding a complete model, our analysis leads to a new interpretation of the interaction between the Question particle and its surrounding context. Another interesting result of our analysis is the semantic interpretation of the forbidden combinations of multiple Copula morphemes. This last point relates to the historical grammaticalization of the Copula morpheme in Turkish.

Keywords: Morphotactics; Turkish; Verb; Tense; Aspect; Modality; Question; Person suffix.

СИСТЕМНЫЙ АНАЛИЗ В МОРФОТАКТИКЕ ОБОЗНАЧЕНИЯ ЛИЦА, ВОПРОСИТЕЛЬНОСТИ И ВРЕМЕНИ В ТУРЕЦКОМ ЯЗЫКЕ

Берке Озенч, Эрджан Солак
Университет Ышык, Стамбул, Турция
ercan.solak@isikun.edu.tr

В парадигме турецкого глагола морфемы времени, лица и вопросительности (Tense, Person and Question (TPQ)) сочетаются (коррелируют) особым (уникальным) образом. Подпарадигма (TPQ) является довольно сложной, в то время как остальная часть турецкой морфотактики является регулярной. Несмотря на то, что в литературе было рассмотрено несколько исследований по анализу закономерностей в их ограниченных рамках, до сих пор не было проведено системного исследования. В данной статье мы представляем полный анализ,

начиная с теоретически полной матрицы морфотактических правил, последовательно исключая запрещенные последовательности, для того, чтобы прийти к вычислительной модели TPQ. Мы представляем окончательную модель в качестве конечного автомата, который можно использовать в более крупных морфологических анализаторах. Наш системный анализ охватывает все комбинации базовых времен, 4 сложных времени (выражаемых сложными морфемами), парадигмы лица и вопросительную частицу. В дополнении к полученной полной модели, наш анализ приводит к новой интерпретации взаимодействия между вопросительной частицей и ее окружающим контекстом. Также интересным результатом нашего анализа, на наш взгляд, является семантическая интерпретация запрещенных комбинаций множественных сложных морфем. Данный момент относится к исторической грамматизации сложных морфем в турецком языке.

Ключевые слова: морфотактика; турецкий язык; глагол; время; аспект; модальность; вопросительность; аффикс лица.

1. Introduction

Turkish has a complex verb inflection paradigm governed by a fairly regular morphotactics. A verb stem is followed by optional morphemes of Voice, Ability, Polarity, Probability and an obligatory Tense morpheme. The Voice part may comprise of multiple morphemes denoting Reflexive, Reciprocal and Causal voices. Up to and including the first Tense morpheme, the morphotactics is completely regular with a well-defined order, (Kornfilt, 1997).

After the first Tense morpheme, there are optional morphemes for Copula Tenses, an obligatory Person morpheme and an optional Question morpheme. The interaction among these three classes of morphemes are not immediately predictable. For example, in some cases, Question morpheme precedes the Person morpheme, in other it follows it. In the rest of the paper we refer to this part of the sub-paradigm as TPQ (Tense-Person-Question). Note that, although the Question particle -mI is written as a separate token in Turkish orthography, we treat it as part of the verb inflection paradigm because it interacts with the rest of the paradigm both in morphotactics and phonology.

In this paper, we provide a systematic analysis of TPQ and propose several semantic interpretations in terms of the Aspect and Modality of the TQP morphotactics.

The rest of the paper is organized as follows. Section 2 describes tenses and the copular tenses of verbal inflection in Turkish. Section 3 focuses on the valid orders among all combinations of tenses and

copula morphemes. Section 4 analyzes the placement of person morpheme. Section 5 examines the placement of the question morpheme. The paper finishes with concluding remarks.

2. Tense and Copular Tenses

Turkish verb stems are followed by an obligatory primary morpheme with either the temporal or mood semantics. In the simplest form, this morpheme is followed by the Person morpheme which makes up the finite verb form. As far as the morphotactics is concerned, the semantics of the primary morpheme play an interesting role which we examine in the rest of the paper. Table 1 list the primary morphemes in their canonic archmorpheme forms with their Tense/Aspect/Mood semantics.

Table 1. Primary morphemes and their semantics

Archmorpheme	Tense	Aspect	Mood
null			Imperative
-(y)A			Optative
-sA			Conditional
-mAlI			Necessitative
-DI	Past	Simple	Indicative
-(I)yor	Present	Imperfective	Indicative
-mAktA	Present	Imperfective	Indicative
-(y)AcAK	Future	Simple	Indicative
-(I/A)r	Aorist	Simple	Indicative
-mIş	Past		Evidential

As we can see from Table 1, the first 4 primary morphemes denote a mood and do not have associated Tense or Aspect semantics in an obvious sense. The rest of the morphemes, designate both Tense and Aspect of the action of the verb as well as Indicative mood.

The primary TAM morpheme may be followed by one or two optional copulas, (Sezer, 2002). Table 2 lists these and their associated semantics under isolation. When these copulas interact with their context, these semantics are modified.

Table 2. Copula morphemes

Archmorpheme	Tense	Aspect	Mood
-(y)sA			Conditional
-(y)DI	Past	Simple	Indicative
-DIr	Aorist	Simple	Indicative
-(y)mIş	Past		Evidential

3. Morphotactics of primary and copula morphemes

In this section, we provide a systematic analysis of the ways in which the primary morpheme is followed by copula morphemes. Since the Imperative mood does not interact with any copulas, we discard it in our analysis below.

When a single copula follows the primary morpheme we have four possible combinations. Two copula morphemes may combine in 16 different ways. Thus, in total, there are 20 distinct ways in which a primary morpheme may potentially be followed by a copula sequence. In Table 3, we marked which of these are valid combinations. For the sake of parsimony, the Table does not show the empty columns.

Table 3. Primary morpheme and copula combinations

Primary			Copula morphemes							
				ş-(y)sA		(y)DI	-(y)mIşş-(y)sA		ş -DIr	ş -(y)mIş
	-(y)sA	-(y)DI	-Dır	-(y)ml	-(y)DI	-(y)DI	-(y)DI	-(y)ml	-(y)ml	-(y)ml
-(y)A		X		X			X			X
-sA		X		X			X			X
-mAlI	X	X X		X	X X X			X	X X	
-DI	X	X					X			

Continuation of table No. 3

-(I)yor	X	X	X	X	X	X	X	X	X	X
-mAktA	X	X X		X	X X X			X	X X	
-(y)AcAK X		X	X	X	X	X	X	X	X	X
-(I/A)r	X	X		X	X	X	X	X	X	X
-mIş	X X X			XXXX				X		

Reviewing the valid combinations in Table 3, we observe the following:

The first part of the table which deals with the case where there is only a single copula that is denser than the rest. This confirms the intuition that acceptability of complex interaction among the primary and copula morphemes is low. In particular, the combinations with three copulas where the second copula is either Conditional – (y)sA or Aorist -DIr are not acceptable. A possible explanation for this consistency is that once the condition or certainty moods are introduced with the copulas, no further moods can be incorporated in to the semantics of the verb. In a way, condition and certainty are final when they are allowed.

The morpheme sequence -DI-DIr and –(y)DI-DIr are not acceptable anywhere, whether as a primary-copula or copula-copula. The certainty and finality expressed by the simple past -DI or (y)DI precludes any further aorist tense or reporting as expressed by -DIr copula.

4. Place of the Person morpheme

There are different Person paradigms in Turkish which are selected by the primary or copula morphemes that proceeds the Person morpheme, (Kornfilt, 1997). In this section, we analyze the patterns governing the placement of the Person morpheme.

In order to simplify the analysis and to easily identify the semantic patterns, we consider only the case of a single Copula. Table 4 summarizes the acceptable locations of Person morpheme. In our notation, PC means Person morpheme comes before the Copula and CP means that Person follows Copula.

Table 4. Locations of the Person morpheme relative to Copula

Primary	**Copula morphemes**			
	-(y)sA	-(y)DI	-DIr	-(y)mIş
-(y)A		CP		CP
-sA		CP		CP
-mAlI	CP	CP	PC	CP
-DI	CP	CP		
-(I)yor	CP	CP	PC	CP
-mAktA	CP	CP	PC	CP
-(y)AcAK	CP	CP	PC	CP
-(I/A)r	CP	CP		CP
-mIş	CP	CP	PC	CP

In Table 4, we observe that, except the Aorist copula -DIr, Person morpheme follows the Copula. In this way, the Aorist Copula follows a finite verb, adding a supposition/certainty/official voice to the whole action denoted by the verb.

(1) Kitab-ı oku-yacağ-ım-dır book-ACC read-FUT-1S-COP.AOR It is certain that I will read the book

The 3rd Plural Person morpheme -lAr has an interesting interaction with the surrounding copula. It can switch its position between before or after the copula without any change in the semantics as in (2) and (3).

(2) Kitab-ı oku-malı-lar-dı
Kitab-ı oku-malı-ydı-lar
(3) Kitab-ı oku-muş-lar-dır
Kitabı-ı oku-muş-tur-lar

Apparently, this is one of the rare instances where the usually strict morphotactics of Turkish allows a change of the order of morphemes without changing the semantics.

5. Place of the Question morpheme

Finally, we look at the placement of the Question morpheme -mI relative to the Copula and Person morphemes. The placements are summarized in Table 5.

Table 5. Locations of the Question morpheme relative to Person and Copula

Primary	**Copula morphemes**			
	-(y)sA	-(y)DI	-DIr	-(y)mIş
-(y)A		–		–
-sA		QCP		QCP
-mAlI	–	QCP	QPC	QCP
-DI	–	QCP		
-(I)yor	–	QCP	QPC	QCP
-mAktA	–	QCP	QPC	QCP
-(y)AcAK	–	QCP QPC QCP		
-(I/A)r	–	QCP		QCP
-mIş	–	QCP QPC QCP		

We observe that when the Copula is either the Conditional –(y) sA or the Optative –(y)A, no Question morpheme is acceptable. This is expected since the semantics of condition/optative and question are mutually exclusive: one cannot question a condition nor can one desire an action yet question it at the same time. However, when the condition is expressed in the primary morpheme, the question is possible as in

(4) Kitab-ı oku-sa-mı-ydı-nız?
Would it be better if you (had/were to) read the book?

In (4), condition morpheme -sA expresses an desirable alternative rather than a condition, i.e. the subject did not read the book and the speaker questions the possibility of a desirable alternative past where the correspondent had read the book.

For the rest of the second column in Table 5, we see that, when the primary morpheme is followed by the past Copula (y)DI, the order of morphemes are Question-Copula-Person.

The 3rd Plural Person morpheme -lAr once again has some freedom of movement. It can either precede the group of Question-Copula or follow it, independently of the internal order of Question-Copula group. This is illustrated by the examples in (5) and (6).

(5) Kitab-ı oku-malı-lar-mı-ymış?
Kitab-ı oku-malı-mı-ymış-lar?
(6) Kitab-ı oku-yacak-lar-mı-dır?
Kitab-ı oku-yacak-mı-dır-lar?

6. Conclusion

Turkish verbal inflection has an intricate interaction among first Tense, Copula, Person and Question morphemes. Generally, it is not a trivial task to predict the valid combinations of the first Tense and Copulas. Also locations of Person and Question morphemes relative to Tense and Copulas need to be worked out. In this paper we present an exhaustive analysis for valid combinations of first Tense and Copula among all combinations. We also present the locations of Person and Questions morphemes in the inflection. We highlight the exceptional freedom of movement of the 3rd plural Person morpheme which constitutes a contrast against the rest of the morphotactics in Turkish. For the disallowed combinations, we propose semantic arguments.

REFERENCES

Kornfilt, J. (1997). *Turkish.* Abingdon : Routledge.
Sezer, E. (2002). Finite inflection in Turkish. *The verb in Turkish*.

УДК 81’33

SYNTACTIC ANNOTATION OF TURKIC LANGUAGES

Berke Özenç, Ercan Solak
Işık University, İstanbul, Turkey
ercan.solak@isikun.edu.tr

The main difficulties in the analysis of Turkic languages lie in the morphology-syntax interaction and the scrambling of the word order. No consistent context free grammars have so far been written for Turkic languages. Several approaches have been developed for grammar modelling in the framework of dependency grammars. Because of its independence of the word order, however, dependency grammar is not generative. The contribution of this paper is three-fold. First, we propose a novel approach to grammar modelling which distinguishes between morphological and syntactic categories. In our approach, while the morphological categories specify the morphological behavior in inflection and derivation, the syntactic categories are context-dependent and are recursively percolated up the syntax tree. Second, we develop a consistent set of extensible morphological and syntactic tags specifically designed for the morpho-syntactic annotation of Turkic languages. Our tag-set has a key-value structure which makes it easier to process and generalize. Third, we introduce SynEdit, an open-source GUI tool to create and edit syntactic structures for Turkic languages, using an extended version of Penn Treebank format in the underlying textual storage. SynEdit has a service-level integration to morphological analysers and requires minimal amount of coding to configure it for use in the annotation of different languages.

Keywords: Morphosyntax, Parsing, Treebank, Syntax tree, Turkic languages.

СИНТАКСИЧЕСКАЯ АННОТАЦИЯ В ТЮРКСКИХ ЯЗЫКАХ

Берке Озенч, Эрджан Солак
Университет Ышык, Стамбул, Турция
ercan.solak@isikun.edu.tr

Основные трудности в анализе тюркских языков заключаются во взаимодействии морфологии и синтаксиса и расшифровке порядка слов. На данный момент для тюркских языков еще не было написано учебников контекстно-свободных грамматик. Было разработано несколько подходов для моделирования грамматики в рамках грамматики зависимостей. Однако из-за независимости порядка слов грамматика зависимости не является генеративной. Вклад данной статьи заключается в следующем: во-первых, предлагается новый подход к грамматическому моделированию, который различает морфологиче-

ские и синтаксические категории. В нашем подходе, в то время как морфологические категории определяют морфологическое поведение во флексии и деривации, синтаксические категории контекстно-зависимы и рекурсивно фильтруются по синтаксическому дереву. Во-вторых, мы разработали последовательный набор расширяемых морфологических и синтаксических тэгов, специально предназначенных для морфосинтаксической аннотации тюркских языков. Наш набор тэгов имеет структуру ключ-значение, которая облегчает обработку и генерализацию. В третьих, мы представляем SynEdit, инструмент с открытым исходным кодом GUI для создания и редактирования синтаксических структур для тюркских языков, используя расширенную версию формата Penn Treebank в базовом текстовом хранилище. SynEdit имеет интеграцию на уровне обслуживания с морфологическими анализаторами и требует минимального кодирования (количества кода) для того, чтобы настроить его для аннотирования разных языков.

Ключевые слова: морфосинтаксис; парсинг; Treebank; синтаксическое дерево; тюркские языки.

1. Introduction

There are two main approaches to model the syntax of sentences. The traditional approach is the constituency grammar where a sentence is recursively made up of smaller constituents, which, at the simplest level, corresponds to words. It is common to represent the constituents and their relations as a syntax tree. Finding the syntax tree of a given utterance is called syntactic parsing, (Martin & Jurafsky, 2009).

A more recent approach is the dependency grammar where the relations among words are viewed as categorized dependencies. Similar to a syntax tree, a dependency tree represents the whole graph of dependencies in a sentence, (Kübler, McDonald, & Nivre, 2009).

Although both the constituency and the dependency grammars have been studied for many languages, for Turkish, the emphasis has so far been on the dependency grammars (Eryiğit, Nivre, & Oflazer, 2008). Only recently, we have seen preliminary attempts at constituency grammars for Turkish, (Dönmez & Adalı, 2018).

In this paper, we focus on constituency parsing of Turkic languages. In the rest of the paper, the term parsing refers exclusively to constituency parsing.

Parsing requires a set of grammar rules which specifies in which combinations and orders the constituents may be combined to form new constituents. In older parsing systems, linguists used to write down grammar rules which were then embedded in parsing algorithms. Naturally, this requires a considerable expert effort.

In contrast, modern approaches define only the general methodology under which human annotators with moderate linguistic expertise construct the parse trees for a set of sentences, thus building a treebank. The rules of the grammar are then inferred by automatically analyzing the treebank. As a by-product of the latter approach, statistical distribution of the grammar rules are obtained only if the treebank is large enough. Using the grammar rules together with their likelihood in the treebank enables the automatic parser algorithms to associate likelihoods to any trees it generates over a given sentence, (Marcus, Santorini, & Marcinkiewicz, 1993).

Ideally, when constructing a treebank, the set of syntactic or frame categories need to be defined and remain fixed during the annotation before the annotators begin their construction of parse trees. In practice, however, the process is iterative, especially at the start. While constructing the parse trees, the annotators might realize the need to add new categories, modify existing categories or remove some categories and tags. Therefore, the policies and procedures need to be defined for the frequent revisions of the work. This is especially important for languages like Turkic family of languages for which previously no efforts to construct treebank exist.

Another difficulty for Turkic languages is that the categories and grammars constructed for Indo-European languages cannot be naturally and easily ported to the Turkic family. Two of the most difficult issues are the fluidity of word order in a sentence and the agglutinative nature of the word forms. The first issue requires an integration of semantic roles within the syntax incorporating the morphological cues. The latter issue requires approaching the grammar construction at the morphosyntactic level rather than solely at the word and syntax level.

In this paper, we propose a method and a set of categories for constructing a constituency treebank for Turkic family of languages. We extend the standard phrasal categories to include their semantic roles within a sentence. Furthermore, we use phrasal stems as an internal node in trees to accommodate late addition in agglutinative languages. We also introduce SynEdit, an open-source tool we developed to visually create and edit syntax trees to accommodate these peculiarities of Turkic languages.

The rest of the paper is organized as follows. In Section 2, we introduce the issues of word order in Turkish sentences as it is related to syntax. In Section 3, we present morphosyntactic categories to be used in Turkish syntax trees. Section 4 introduces SynEdit. In Section 5 we

propose a tagset for Turkic languages. Section 6 provides several SynEdit examples. The paper finishes with concluding remarks.

2. Word order and semantic roles

Turkic languages are scrambling languages where the word order usually serves pragmatic functions such as topic, focus and backgrounding, (Erguvanli & Taylan, 1984). Although not orthographically marked in writing, the marked word orders are usually associated with prosodic markers such as stress and short pauses in speech.

When listening to speech, humans use the syntax as well as the prosodic cues to parse the utterances. Lacking those cues, an automatic parser needs to use word order, morphological markers and statistical analysis to generate plausible syntax trees. The same is true when a human annotator constructs a tree over a sentence, except in this case, the statistics are stored in the mental lexicon and the world knowledge of the annotator.

The simplest approach for handling scrambling in constituency grammars is to create rules for each valid constituent order.

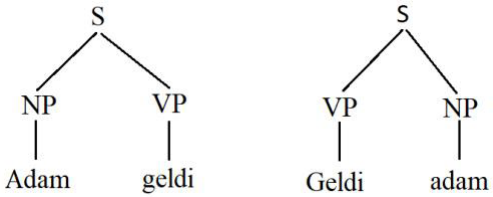

Figure. 1. Trees for sentences «Adam geldi» and «Geldi adam»

Consider the scrambled sentences «Adam geldi» and «Geldi adam» (the man came) and their parse trees in Figure 1. In order to incorporate the scrambling between these two sentences, we can add two rules

S: NP VP

S: VP NP

to the grammar. Or, if we want to distinguish the backgrounded subject NP in the second sentence from the one in the unmarked order of the first sentence, we can split the NP category into two, one for unmarked order and one for backgrounded constituent. In this case the rules become

S: NP VP

S: VP NP-BG

In both approaches, we need a large treebank to be able deduce significant probabilities for the rules. Yet, the second method generates the pragmatic roles in the parses.

Similarly, we can split the usual categories of constituents to indicate semantic roles. In Figure 2, the category NP is split into two categories in order to differentiate the subject and the direct object.

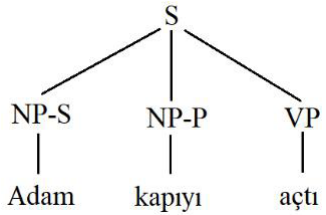

Figure. 2. Tree for «Adam kapıyı açtı.»

3. Morphosyntactic categories

In the traditional approaches to constituency grammars, the smallest constituent is a token which can be a word, a multi-word expression or a punctuation mark. All the other constituents are recursively generated by combining tokens. Thus, in the syntax tree representation, the tokens correspond to leaves.

Starting from the tokens as leaves is a practical approach for languages like English and Chinese which have simpler morphologies. In such languages, the small set of inflections of a word stem can be viewed as distinct tokens with appropriate categories. For example, in English, the number inflection of nouns generate two distinct forms at the leaf level, one for singular and one for plural.

This approach is insufficient when applied to languages with complex morphologies. For example, for Turkic languages, a large part of syntax tree is already contained within the morpheme sequence of a word form. The difference is illustrated in Figure 3 where two isomorphic sentences «I will not come» and «Gelmeyeceğim» correspond to completely different trees when leaves are taken to be tokens.

Instead, we propose a view of grammar that starts at the morpheme level. In this view, some of the morphosyntactic rules are incorporated into the grammar. The question of which morphemes need to be treated as leaves and which need to be grouped under a single leaf has to be worked out while constructing the treebank.

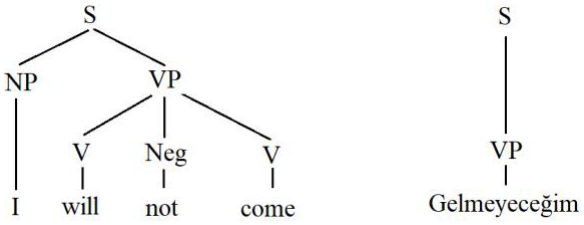

Figure. 3. Trees for «I will not come» and «Gelmeyeceğim»

With the morphemes or morpheme groups treated as leaves, the Turkish sentence in Figure 3 can be parsed as in Figure 4.

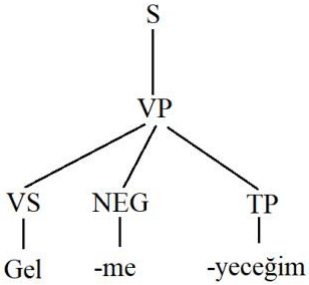

Figure. 4. Tree for «Gelmeyeceğim»

Note that, in Figure 4, we grouped Tense and Person morphemes together as they always occur in sequence. Looking at the new tree, we can infer two rules of the grammar as

S: VP

VP: VS NEG TP

where VP is the non-terminal symbol for finite verb and TP denotes the morpheme group composed of Tense and Person. VS denotes the Verb Stem.

A natural extension of our approach is that verb stems are treated as leaves which might combine with the constituents to their left to form phrasal stems. This extension is also consistent with the late attachment theories of Turkic grammars, (Göksel & Kerslake, 2004). Figure 5 illustrates this for the sentence «Bugün gelmeyeceğim» (I will not come today) where the adverb «bugün» is first attached to the verb stem «gel» to form a verb phrase stem, VPS, and then VPS is attached to the constituents to its right.

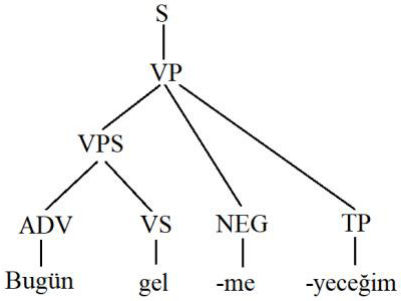

Figure. 5. Tree for «Bugün gelmeyeceğim»

4. SynEdit

In order to facilitate the construction of treebanks and the grammars for agglutinative languages in general and Turkic languages in particular, we created SynEdit, an open-source visual tool for editing syntax trees. Figure 6 shows a typical screenshot of SynEdit.

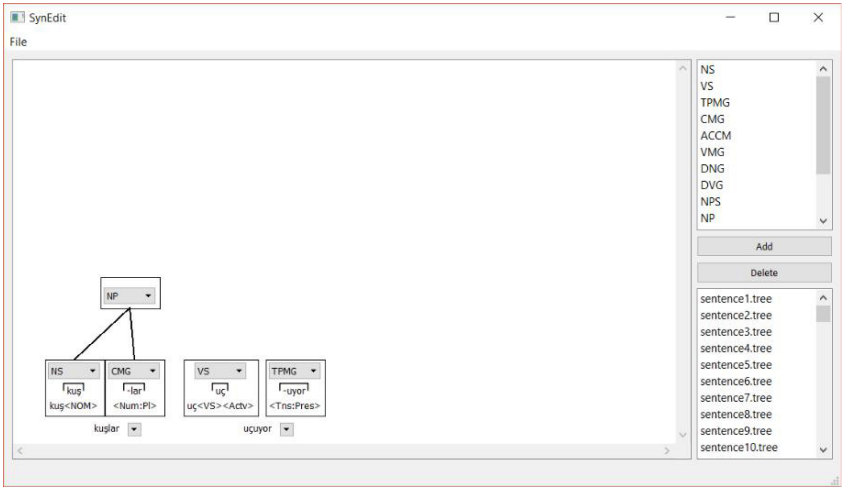

Figure. 6. Screenshot of SynEdit

When manually parsing a sentence, SynEdit uses a morphological analyzer through a web service protocol to obtain the analyses for all the tokens in the sentence. The construction starts by manually selec-

ting the correct morphological analysis for each token. Once selected, the morphemes in an analysis become the potential leaves of the syntax tree. The annotator can group sets of adjacent morphemes within a single token to create new leaves. In Figure 6, the Tense and Person morphemes are grouped together while all the other morphemes are left as singleton leaves.

After determining the leaves, the annotator begins constructing the tree. There are two basic operations. The first is selecting a set of nodes and creating a common parent for them. The second operation is connecting a node to an already existing parent node as its new child. Apart from these, deleting an internal node removes all the edges connected to it, leaving the rest of the tree intact. Similarly, changing the chosen analysis or the morpheme grouping of a leaf node removes all the edges connected to those affected groups.

SynEdit constrains the human annotator within allowable operations. A child node cannot be connected to more than one parent. There cannot be crossings among the edges of the tree.

SynEdit starts with a basic set of tags we propose for Turkic languages. We explain the rationale and the details of this tagset in the next section. The annotator might add new tags to the tagset and edit or remove the existing tags. If, upon an update of the tagset, some of the already parsed sentences become invalid, their files are indicated as such with a red color. The annotator can then visit those sentences again and update their trees as necessary. This feature enables an iterative development of the grammar as the annotators encounter various combinations of constituents.

Finally, SynEdit can generate the grammar as a set of tagsets and a set of rules inferred from the treebank together with their associated unsmoothed probabilities. These can then be fed into an automatic parser. We will provide the details of a parser for Turkic languages in a future paper.

5. A basic tagset for Turkic family

Here we propose a basic tagset for the internal nodes of syntax trees in Turkic languages. First, we outline the theoretical foundations on which the tagset is based.

1. The leaves are the morphemes or groups of morphemes. The root or stem of a derived word can be a leaf.

2. Semantic roles and POS tags are assigned to words or groups of words rather than to single words. This is in contrast to the traditional view of POS tags being exclusively assigned to single words.

3. Tags specify the behavior of constituents when combining with other constituents. If two constituents differ in their combinations in different contexts, then they must be labeled with distinct phrasal tags.

4. Every tree must have a single root node labeled S.

5. Adverb phrases that qualify verbs attach (usually from the left) to the verb stem, rather than the finite verb. Verb inflection morphemes are attached after these.

6. Nominal phrases that qualify a noun attach (usually from the left) to the nominal stem, rather than the infected noun. Number, possession and case inflections are attached after these.

Using these fundamentals, we propose two groups of tags.

Morpheme tags: These tags are used only for word stems, morphemes and morpheme groups.

NS: nominal stem
VS: verbal stem
TPMG: Tense and Person group
CMG: Nominal inflection group up to and including the case morphemes for dative, locative, ablative, genitive and instrumental cases.
ACCM: Accusative morpheme
VMG: Verbal inflection group including voice, ability, polarity, probability
DNG: Group of derivational morphemes deriving a nominal stem.
DVG: Group of derivational morphemes deriving a verbal stem

Phrase tags: These tags are used for stemmed and finite phrases.
NPS: Stemmed noun phrase
NP: Noun phrase
VPS: Stemmed verb phrase
VP: Verb phrase
NPC: Case (except accusative) marked noun phrase
NPACC: Accusative marked noun phrase

6. Examples

Here we provide some examples of syntax trees constructed using SynEdit and the tagset that we proposed.

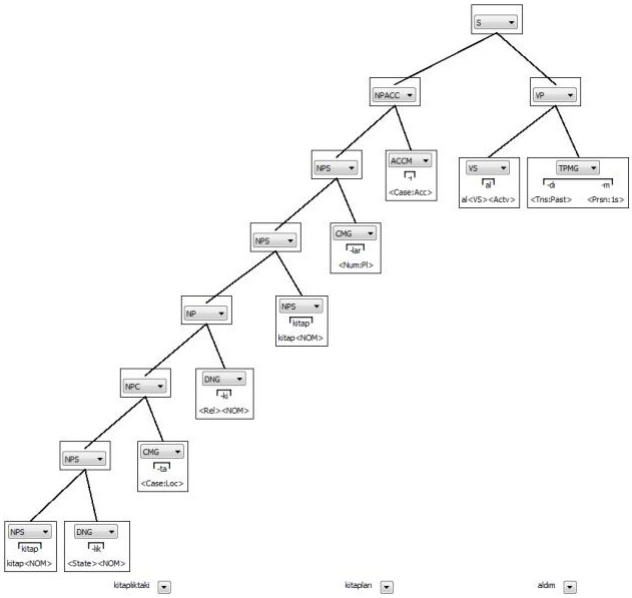

Figure. 7. The parse tree of the sentence «Kitaplıktaki kitapları aldım.»

The parse tree for the sentence,

Kitap-lık-ta-ki kitap-lar-ı al-dı-m book-TOOL-LOC-REL book-PL-ACC take-PAST-1S I took the books in the bookcase

is shown in Figure 7. The derivation of «kitap-lık» (bookcase) from the root «book» as a syntactic operation that generates an internal NPS node. Figure 7 also illustrates the left-attachment of qualifying NP «kitaplıktaki» to the noun stem «kitap» to form a stemmed noun phrase. The plural morpheme is attached after this.

Figure 8 illustrates the attachment of adverbs to verb stems before the rest of the verb inflection is attached on the sentence

Ben çabucak gel-eceğ-im
I quickly come-FUT-
1S I will come quickly

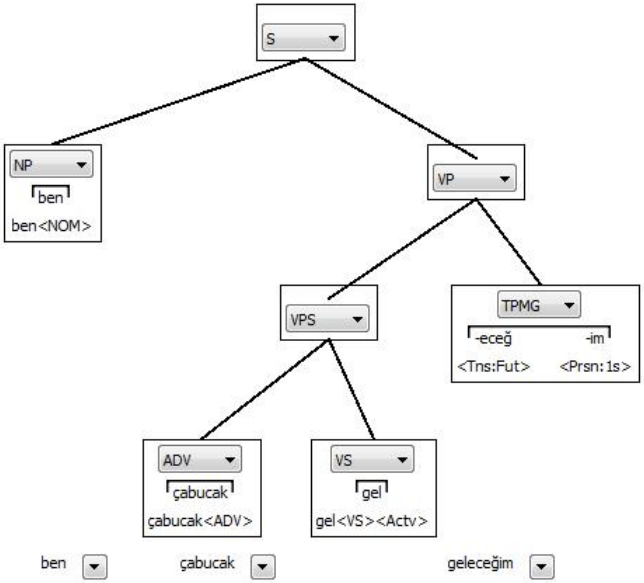

Figure. 8. The parse tree of the sentence «Ben çabucak geleceğim.»

7. Conclusion

In this paper, we propose a novel morphosyntactic approach to constituency grammars of Turkic languages. In our approach, syntax trees start from the level of morphemes. The stemmed phrases feature prominently in the syntax trees leaving the inflection constituents for late attachment. We introduce a novel set of basic tagset to be used in morphosyntactic parsing. We support our approach with SynEdit, an open-source tool we developed in order to facilitate the human annotation of parse trees for Turkic languages. We believe our approach to constituency grammar, supported by SynEdit, will lead to construction of treebank for Turkic languages.

REFERENCES

Dönmez, İ., & Adalı, E. (2018). Contex Free Grammer For Turkish. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22(2)*, 552–561.

Erguvanli, E. E., & Taylan, E. E. (1984). *he function of word order in Turkish grammar.* Los Angeles: Univ of California Press.

Eryiğit, G., Nivre, J., & Oflazer, K. (2008). Dependency parsing of Turkish. *Computational Linguistics, 34(3)*, 357–389.

Göksel, A., & Kerslake, C. (2004). *Turkish: A comprehensive grammar.* Abington: Routledge.

Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. G. Hirst içinde, *Synthesis Lectures on Human Language Technologies* (s. 1–127). San Rafael: Morgan & Claypool Publishers.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). *Building a large annotated corpus of English: The Penn Treebank.* Pennsylvania : University of Pennsylvania Department of Computer and Information Science.

Martin, J. H., & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River: Pearson.

УДК 81’33

VISUAL MODELLING OF MORPHOLOGY

Berke Özenç, Ercan Solak
Işık University, İstanbul, Turkey
ercan.solak@isikun.edu.tr

The most common way to computationally model a morphology is to use finite-state transducers (FST). This stage is followed by the modelling of phonology using a set of rewrite rules implemented using regular expressions. Modelling the morphology of a language thus is a manually intensive task which requires expertise as well as competence in using the FST tools. Furthermore, textual FST models are difficult to modify, maintain and adapt. On the other hand, graphical representations of FSTs are both intuitive and easy to maintain. In this paper, we introduce DiaMor, an open-source tool we created to bridge the gap between the graphical representations of morphology and its textual descriptions in the domain language of FSTs. DiaMor was created in python and PyQt graphical framework and previously tested to model the morphologies of Turkish and Azerbaijani. In this paper, we demonstrate the uses of DiaMor, in which a computational linguist can model a morphology in the popular diagramming tools draw.io. Once the modelling is complete, DiaMor automatically generates the first level FST specifications and combines them with the textual lexicon and the second-level phonology rewrite rules to generate a compiled FST, which can then be embedded in an application in either the generation or analysis mode.

Keywords: Morphology; Modelling; Finite-State Transducers; Turkic languages.

ВИЗУАЛЬНОЕ МОДЕЛИРОВАНИЕ МОРФОЛОГИИ

Берке Озенч, Ерджан Солак
Университет Ышык, Стамбул, Турция
ercan.solak@isikun.edu.tr

Наиболее распространенный способ вычислительной модели морфологии – использование конечных автоматов (FST). За этим этапом следует моделирование фонологии с использованием набора правил перезаписи, реализованных с использованием регулярных выражений. Таким образом, моделирование морфологии языка – это большая ручная работа, требующая большого опыта и навыков использования инструментов FST. Кроме того, текстовые модели FST сложно модифицировать, поддерживать и адаптировать. С другой стороны, графические представления FST являются интуитивно понятными и

простыми в обслуживании. В данной статье мы представляем DiaMor, инструмент с открытым исходным кодом, который мы создали, чтобы преодолеть разрыв между графическими представлениями морфологии и ее текстовыми описаниями на языке программирования FST. DiaMor был создан в Python и графической среде PyQt, будучи предварительно протестированным на моделировании морфологии турецкого и азербайджанского языков. В данной статье демонстрируется DiaMor, в котором лингвист, может моделировать морфологию того или иного языка в популярных инструментах построения диаграмм draw.io. По завершении моделирования DiaMor автоматически генерирует спецификации FST первого уровня и объединяет их с текстовым лексиконом (словарем) и правилами перезаписи фонологии второго уровня для создания скомпилированного FST, который затем может быть встроен в приложение в режиме генерации или анализа.

Ключевые слова: морфология; моделирование; конечные автоматы; тюркские языки.

1. Introduction

Modelling the rich morphologies of Turkic languages for computing purposes has always been a challenging and labor intensive task. Although there is a varying degree of similarity among the Turkic family of languages, each language has a distinct set of inflection and derivation paradigms which prevents a direct transfer of a model developed for a particular Turkic language to another. When creating a model for one of these languages, a researcher has to either painstakingly modify an existing model or start from scratch and try to arrive at a complete model of a new language.

Apart from these theoretical difficulties, there is the practical limitation of not having access to an easy-to-understand representation of existing models. Often, one needs to decipher a rather opaque code base in the fortunate case when the code is available.

As an illustration, below is an extract from the description of TrMorph, a morphological analyzer for Turkish (Çöltekin, 2014).

```
LEXICON NomPredQCont
#if (PREDICATE_WITHOUT_PAGR == 1)
%<cpl%:pres%>:@MB    ENDLEX;
#endif
%<q%>:@miSPCm^I@MB   NomPredPagr;
%<q%>:@miSPCm^I@MB   NomPredDir;
%<q%>:@miSPCm^I@MB   CplAll;
```

Obviously, even with the available comments dispersed at the start of different sections, this piece of code is difficult to grasp by someone other than the creator of the model. Even when the codebase is maintained by the same developer, evolving modifications and additions over the development cycle presents additional challenges.

On the other hand, pictorial representations of models are easier to create and maintain. Human perception is more aligned with pictorial visualizations compared to textual ones. Scanning a large diagram is much quicker than going back and forth among a number of text files. For example, the diagram in Figure 1 is a simplified version of noun inflection in Turkish. Even with a minimal background in its technical implementation and without any associated explanations, a linguist can immediately grasp the meanings of the states, the transitions and their associated outputs. By tracing the arrows, one can test the possible paths and combinations in the inflection paradigm.

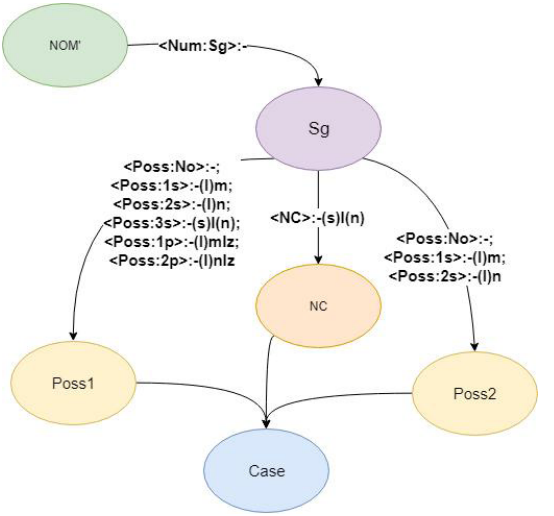

Fig. 1. A simplified noun inflection diagram

In this paper, we provide the details of DiaMor, a tool we created to bridge the gap between the diagrammatic representation of morphology and its two-level description as outlined above. Using DiaMor, computational linguists need not to spend time on the syntax of the lexical FST descriptions and instead concentrate on the linguistic

principles governing the morphology. Once the graphical description is complete, DiaMor generates the final analyzer as well as the intermediate lexical FST files. Thus, for the modelling of morphology, only a moderate competency in regular expressions is needed alongside the linguistic knowledge required in order to describe the morphotactics.

We chose `draw.io` as the diagramming tool for DiaMor for its simplicity of use and its open-source license. The underlying format for `draw.io` is XML. DiaMor parses this XML file and generates lexical FST files. Finally, the lexical files are compiled together with the phonological rules to generate the final analyzer.

The rest of the paper is organized as follows. In Section 2, we briefly describe the two-level morphology approach. Section 3 describes the representation morphotactics in diagrams. In Section 4, we introduce DiaMor in detail. The paper finishes with concluding remarks.

2. Two-level morphology

Since the introduction of `xfst` by (Koskenniemi, 1983), representing morphologies in two levels has become a standard approach. In this approach, the first level is a lexical Finite-State Transducer (FST) which represents the morphotactics and the second level is a series of re-write rules representing the rules of the phonology. The transitions in the FST also handle the conversion between the abstract morphemes into archmorphemes with special symbols and archphonemes. The second level converts the output of the first level to the orthographical surface form by replacing symbols and archphonemes to their surface level graphemes.

Both of the levels are described using their respective domain languages. The first level description starts with the root lexicon and lays out the possible transitions with their associated input/output pairs. Below is a short segment of an imaginary first level description stored in a `lexc` file.

```
Multichar_Symbols
%<NOM%>
%<Num%:Pl%>
%<Num%:Sg%>

LEXICON Nom'
%<Num%:Pl%>:%-l%{A%}r Pl;
```

```
%<Num%:Sg%>:%- Sg;
%<IsWhen%>:%-%(y%)ken Adv_CAT;

LEXICON    Case

%<Case%:Loc%>:%-%{D%}%{A%} Case2;
%<Case%:Gen%>:%-%(n%)%{I%}n Case2;
%<Case%:Acc%>:%-%(y%)%{I%} Case1;
CaseDeriv2;

LEXICON Case1
# ;
```

The second level description is a set of replacement rules written using regular expressions. When the context described by the regular expression is matched in the input to the second level, the replacement rule is applied. Below is a short segment from the second level description of vowel harmony in Turkish, stored in a twol file.

```
Alphabet
ABCÇDEFGĞHIİJKLMNOÖPRSŞTUÜVYZ a b c ç d e f g ğ h ı i j k
l m n o ö p r s ş t u ü v y z %{A%} %{I%} %{D%} %{C%}
%{K%} %{SPC%} %- %’ ;


Sets
Cons = B C Ç D F G Ğ H J K L M N P R S Ş T V Y Z
b c ç d f g ğ h j k l m n p r s ş t v y z %’ ;
Back = A I O U a ı o u ;
beforeBackA = B C Ç D F G Ğ H J K L M N P R S Ş T V Y Z b
c ç d f g ğ h j k l m n p r s ş t v y z
A I O U a ı o u %{A%} %{I%} %{D%} %{C%} %{K%} %- ;
Vow = A E I İ O Ö U Ü a e ı i o ö u ü ;

Rules
«A to a»
%{A%}:a <=> Back: [Cons: ]* [%-: ]* [beforeBackA:
]* _ ;
```

The first application of two-level morphology modelling to Turkish was undertaken by Oflazer (Oflazer, 1994). The same methodology have later been used for the morphologies of other Turkic languages, (Kessikbayeva, 2016), (Matlatipov & Vetulani, 2009), (Tantuğ, Adalı, & Oflazer, 2006), (Washington, Ipasov, & Tyers, 2012), (Özenc, Eh-

sani, & Solak, 2018). The availability of these analyzers ranges from completely open source to non-availability to the research community.

3. Morphotactics in diagrams

When working out a model of the morphology of a language, computational linguists seldom work directly in the domain language of lexical FST. Instead, they use pen and paper or graphics software to draw the morphotactics. When the model is mature enough, they manually transfer it to `lexc` and `twol` files and compile the resulting two-level description into a binary model of morphology.

When the model needs to be modified due to a bug or an extension to new sub-paradigms, the updates can be directly applied in the `lexc` files. If the developers maintain the original drawings or diagrams, they need to update those as well. One can imagine what kind of synchronization problems such a dual path approach would result in.

We developed DiaMor so that the developer can work with the diagrams and `twol` files only, not worrying about the `lexc` files. We chose `draw.io` as the diagramming tool due to its simplicity and availability as a web application as well as a desktop application in different platforms.

The `draw.io` tool is a general purpose diagramming tool but for DiaMor we use a small subset of its capabilities to draw circles to represent the states in a FST and arrows between those circles to represent the transitions between the states. The labels on the transitions contain both the triggers that invoke the transition and the outputs that the transition yields. The stylistic features of `draw.io` such as color and font can be used freely to obtain a better visualization without any modification on the underlying model of the morphotactics. In Figure 2, part of verbal inflection paradigm is shown where the stylistic features have been used to visually group or emphasize elements. This decoupling of FST representation and the visual style give the computational linguist a freedom in organizing the diagram.

There are of course some restrictions on the syntax of the symbols used in draw.io diagram that would be input to DiaMor. All the states must be labeled. Two states sharing the same label are treated as a single state in the FST. This becomes especially handy when the transitions intersect and the diagram becomes cluttered. By using a few replicas of a state in different locations, the transitions might be drawn without intersecting. Naturally, there is a trade-off in replicating the

states as, if used too often, one might lose the visual connection among different sub-paradigms in the diagram.

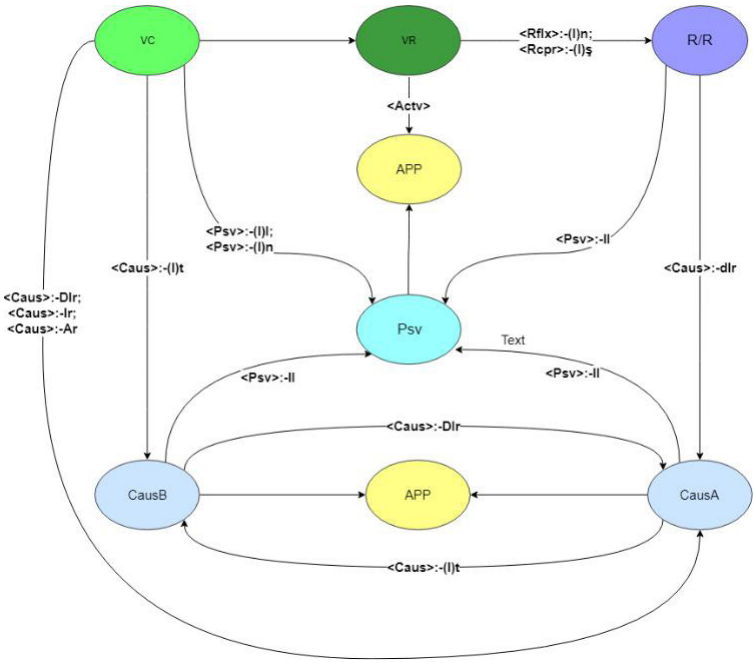

Fig. 2. Verb Inflection Diagram

The labels on the transitions denote the trigger/output pairs. The part that is left of ':' is the abstract morpheme. The general format of an abstract morpheme is either `<Key:Value>` or `<Key>` depending on its semantics. The part that is to the right is the first level output which might contain archphonemes and special symbols. Every symbol or archphoneme must be in the format `{Sym}` where `Sym` can be a multi-character string. The output string must start with a '-' to indicate the start of a morpheme. These symbols are either eliminated or converted to their surface forms in the second level. It is the responsibility of the modeler to write the appropriate replacement rules in the second level. It is possible to have multiple input/output pairs sharing the same edge in the diagram. In this case, different trigger/output pairs must be separate by a ';'.

When a transition corresponds to a derivation that updates the cat-

egory of the stem that it generates, the new category is indicated at the start of the abstract morpheme as in <NEWCAT><AbstractMorph>.

4. DiaMor

The `draw.io` tool stores its diagrams as XML files. DiaMor parses these in order to generate an internal data structure to represent the FST diagram, together with its list of transitions, states and special symbols. This interim structure is then written out in `lexc` files adhering to the syntax specified by `lexc` format.

DiaMor stores all its working files in a single project directory which also contains a project configuration file.

At the start, DiaMor expects to find in the project directory a `words.txt` file which contains a list of root lexemes and their category tags. This is essentially the root lexicon of the language.

DiaMor searches the project directory for any twol files that define phonological re-write rules. These are listed in the interface. The order of application of these rules can be changed or some rules might be disabled. This is an especially handy feature when testing the morphotactics independently of the phonology.

The interface of DiaMor is given in Figure 3.

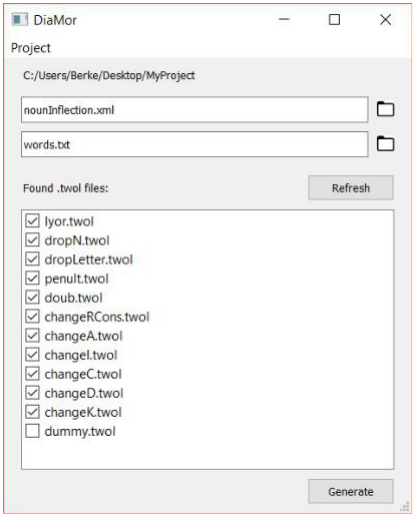

Fig. 3. DiaMor Interface

DiaMor generates two binary models, one for analysis and one for generation. The analysis model is frequently used while the generation is used for testing purposes.

We implemented DiaMor in python using PyQt framework for the interface. As the FST engine, we used HFST (Linden, Axelson, Hardwick, Silfverberg, & Pirinen, 2011). DiaMor is available as an open-source tool in its github repository (Özenç, 2019). The repository also contains instructions on how to install and use DiaMor for typical applications.

5. Conclusion

Modelling morphologies is labor intensive. In this paper, we described DiaMor, a tool we created to alleviate this burden and free the computational linguist from worrying about the implementation details of morphotactics using Finite-State Transducers. DiaMor also streamlines the process of combining morphotactics with phonology to generate a complete analyzer. With DiaMor, all the computational linguist needs to do is to provide a list of root lexemes, to visually draw the diagram of morphotactics and provide a set of phonology re-write rules.

REFERENCES

Çöltekin, Ç. (2014). A set of open source tools for Turkish natural language processing. *Ninth International Conference on Language Resources and Evaluation* (s. 1079–1086). Reykjavik: European Language Resources Association.

draw.io. (2019, August 25). draw.io: taken from https://www.draw.io/

Kessikbayeva, G. C. (2016). A RuleBased Morphological Analyzer and a Morphological Disambiguator for Kazakh Language. *Linguistics and Literature Studies*, 96–104.

Koskenniemi, K. (1983). *Two Level Mophology: A General Computational Model for Word-Form. Recognition and Production*. Helsinki: University of Heksinki.

Linden, K., Axelson, E., Hardwick, S., Silfverberg, M., & Pirinen, T. (2011). HFST–framework for compiling and applying morphologies. *Second International Workshop on Systems and Frameworks for Computational Morphology* (s. 77–85). Zurich: Springer.

Matlatipov, G., & Vetulani, Z. (2009). Representation of Uzbek Morphology in Prolog. M. M., & M. A. içinde, *Aspects of Natural Language Processing* (s. 83–110). Berlin: Springer.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 137–148.

Özenc, B., Ehsani, R., & Solak, E. (2018). Moraz: an open-source morphological analyzer for Azerbaijani. Turkish. *2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (s. 25–29). Brussels: Association for Computational Linguistics.

Özenç, B. (2019, August 15). *DiaMor*. GitHub: taken from https://github.com/berkeozenc/Diamor Tantuğ, A. C., Adalı, E., & Oflazer, K. (2006). Computer Analysis of the Turkmen Language. Morphology. *Advances in Natural Language Processing* (s. 186–193). Finland: Springer.

Washington, J., Ipasov, M., & Tyers, F. (2012). A finite-state morphological transducer for Kyrgyz. *Eighth International Conference on Language Resources and Evaluation* (s. 934–940). İstanbul: European Language Resources Association.

УДК 81’342

THE VOWEL SYSTEMS OF THE YORUBA LANGUAGE AND NIGERIAN ENGLISH: COMPARATIVE ASPECT

A. D. Petrenko, D. A. Petrenko, N. A. Vovk
The Crimean Federal V.I. Vernadsky University,
Simferopol, the Russian Federation
aldpetrenko@mail.ru, daniil.petrenko@list.ru, nick.wolf@mail.ru

The article deals with the sphere of the sociolinguistic researches. The main purpose of the work is in defining the influence of the phonetic characteristics of the mother tongue (Yoruba) on the peculiarities of pronunciation and formation of the vowel system of Nigerian English (further – NigE). In the article there is given the characteristic of the vowel phonemes of the Yoruba language, there is represented the vowel system of British English (further – BrE), there are analyzed NigE vowels. The authors compare all these three systems with each other.

The Yoruba language belongs to the group of the Benue Congo languages and is used in communication in the south western part of the Federative Republic of Nigeria. Yoruba has undergone the process of standardization on the basis of the Oyo dialect. In the standard Yoruba there are eighteen consonants, seven oral vowels, five nasal sounds what in general makes up twelve vowel phonemes. The letters which mean oral vowels are represented in the alphabet of the language. In BrE there are twenty vowels – twelve monophthongs and eight diphthongs. As a result of the Yoruba language vowel phoneme influence on the processes of learning BrE vowels by the representatives of the same-name ethnos there has been formed NigE vowel system of its own which differs from BrE one.

Keywords: sociolinguistics; sociophonetics; the Yoruba language; British English; Nigerian English; vowel system; vowel phoneme.

НАПРАВЛЕНИЕ ИССЛЕДОВАНИЯ: СОЦИОЛИНГВИСТИЧЕСКИЕ ИССЛЕДОВАНИЯ В КОНТЕКСТЕ НАУКИ XXI ВЕКА

СИСТЕМЫ ВОКАЛИЗМА ЯЗЫКА ЙОРУБА И НИГЕРИЙСКОГО ВАРИАНТА АНГЛИЙСКОГО ЯЗЫКА: СОПОСТАВИТЕЛЬНЫЙ АСПЕКТ

А. Д. Петренко, Д. А. Петренко, Н. А. Вовк
Крымский федеральный университет им. В. И. Вернадского
Симферополь, Российская Федерация
aldpetrenko@mail.ru, daniil.petrenko@list.ru, nick.wolf@mail.ru

Статья представлена в русле социолингвистических исследований. Основная цель работы состоит в выявлении влияния фонетических характеристик родного языка (йоруба) на особенности произношения и формирование системы вокализма нигерийского варианта английского языка (далее – НВАЯ). В статье дается характеристика гласных фонем языка йоруба, приводится система вокализма британского английского (далее – БА), рассматриваются гласные НВАЯ. Авторы сопоставляют все эти три системы между собой.

Язык йоруба принадлежит к группе бенуэ-конголезских языков и используется в коммуникации в юго-западной части Федеративной Республики Нигерия (далее – ФРН). Йоруба подвергался процессу стандартизации на основе диалекта ойо. В стандартном йоруба содержится восемнадцать согласных, семь ртовых гласных, пять назальных звуков, что составляет в целом двенадцать гласных фонем. Буквы, которые обозначают ртовые гласные, представлены в алфавите этого языка. В британском английском насчитывается двадцать гласных – двенадцать монофтонгов и восемь дифтонгов. В результате влияния гласных фонем языка йоруба на процессы освоения гласных британского английского языка представителями одноименного этноса сформировалась своя система вокализма НВАЯ, которая отличается от системы БА.

Ключевые слова: социолингвистика; социофонетика; язык йоруба; британский английский; нигерийский вариант английского языка; система вокализма; гласная фонема.

1. Введение

Актуальность статьи обусловлена рядом факторов. Статус нигерийского варианта английского языка всегда находился в центре внимания языковедов. Английскому языку (далее – АЯ) отводится важное место в нигерийском обществе. АЯ используется в деятельности Национального собрания (законодательного органа Федеративной Республики Нигерия) (статья 55 Конституции) и Национального собрания любого штата этого государства (статья 97 Конституции) (Constitution of the Federal Republic of Nigeria 1999, с. 24; Constitution of the Federal Republic of Nigeria 1999, с. 35). В той или иной степени английским владеет значительное большинство граждан этой страны. Однако, анализ научной литературы показывает, что на фонетико-фонологическом уровне присутствуют определенные сложности. В число последних входит отсутствие единой системы вокализма, которая подходила бы для всех носителей НВАЯ. Это объясняется историческими причинами (Jibril M.M., 1982) и тем фактом, что в ФРН насчитывается более 400 местных языков (The Oxford Companion to the English

Language, 1992, с. 700), которые располагают своими гласными фонемами и влияют на произношение тех, кто изучает АЯ. Среди таких языков представлен и йоруба.

2. Цель и задачи исследований

Основная цель исследования – выявить влияние фонетико-фонологического уровня языка йоруба на специфику произношения и образование системы вокализма НВАЯ. Эта цель обусловливает такие задачи: 1) дать характеристику гласных языка йоруба; 2) представить систему вокализма БА; 3) рассмотреть гласные нигерийского варианта АЯ; 4) сопоставить все эти три системы между собой.

3. Методика исследований

Анализ проводился с использованием аналитического и сопоставительного методов. Первый позволил выделить материал по системам вокализма языка йоруба, БА и НВАЯ. Сопоставительный метод применялся при сравнении гласных фонем упомянутых языков.

4. Результаты исследований

На текущем этапе социолингвистика разрабатывает ряд крупных тем. Анализируются вопросы двуязычия и многоязычия. В частности, ученые отмечают необходимость уточнения первого из двух терминов (Петренко А. Д., Петренко Д. А., 2018, с. 43). Много внимания уделяется вариативности языка и социальной принадлежности говорящих (Петренко А. Д., 1998, с. 3).

Социофонетика, как направление социолингвистики, также успешно исследует вопрос соотношения нормы и вариативности. Для этого используется значительный арсенал факторов, которые влияют на применение произносительных средств в речи (Петренко А. Д., 2018 б, с. 5). Фонетическая специфика речи в значительной мере определяется профессиональной принадлежностью, полом и возрастными особенностями (Петренко А. Д., 2018 б, с. 13). В качестве примера исследования, в котором профессиональный аспект играет очень важную роль, можно привести анализ особенностей произношения немецких политических деятелей (Петренко Д. А., 2000). В сферу интересов лингвистов также входит

и взаимодействие АЯ с местными языками в разных странах, в том числе, и в государствах африканского континента. Результаты интерференции выражаются и на фонетическом уровне.

4.1. Язык йоруба. Обзор

Язык йоруба – один из наиболее употребляемых в Африке. Также этот язык один из основных в Нигерии вместе с языками хауза и игбо. На нем разговаривает одна из самых больших этнических и культурных групп Нигерии – йоруба. Считается, что представители этноса мигрировали с востока к месту настоящего расположения на запад в район низовья реки Нигер более тысячи лет тому назад. Было создано несколько независимых королевств, которыми правили наследные короли. Центром каждого королевства была столица. Одной из них стал город Ифе, который, кроме политической значимости в прошлом, все еще имеет особую религиозную значимость как место сотворения земли согласно мифологии йоруба. Много представителей племени йоруба были взяты в качестве рабов в Северную и Южную Америки, а их верования и культура повлияли на общественный уклад Бразилии, Кубы и других стран. Йоруба – это тональный язык, в нем очень мало флективной морфологии и присутствует строгий порядок слов в виде модели подлежащее-сказуемое-дополнение.

На языке йоруба разговаривают около 29,4 миллиона людей, из которых 28,5 миллиона живут в Нигерии, а остальные в Бенине (800,000) и Того (100,000).

Самыми старыми письменными документами на языке йоруба являются: лексикон, составленный Томасом Боудичем в 1819; еще один лексикон, составленный Ханной Килхэм в 1828; обучающие буклеты Джона Рабана (1830–32); в 1843 был составлен самый первый словар языка йоруба, который был напечатан епископом Самюэлем Краузером, бывшим рабом из народности йоруба, который обосновался в Сьерра Леоне; в 1852 году Самюэлем Краузером была написана грамматика языка йоруба. В 1859–67 годах издавалось первое периодическое издание на языке йоруба, которое также стало и самым первым на местном языке в регионе Западной Африки (Yoruba. The Language Gulper).

Язык йоруба является самым наиболее изучаемым из всех западных бенуэ-конголезских языков. На йоруба говорят, главным образом, в юго-западной части Нигерии, а также в таких стра-

нах, как Республика Бенин, Того и Сьерра-Леоне в Африке. Также язык йоруба используется на Кубе, в Бразилии и многих странах, не находящихся в Африке. В число разнообразных диалектов языка йоруба входят игбомина, ондо, иджеша, оке-огун, иболо, ифе, йеуа, эгба, иджебу, авори, ойо и ибадан. Основой для стандартизации языка йоруба стал диалект ойо. Однако, К. Аденийи и О. Е. Бамигбаде (K. Adeniyi, O. E. Bamigbade) отмечают, что на текущем этапе стандартный йоруба (СЙ) так сильно отошел от диалекта ойо, что считается совсем другим языком. Считается, что, в основном, СЙ существует в письменной форме, а также как своего рода лингва франка в пределах территорий, которые говорят на йоруба. В ходе коммуникации в государственных учреждениях большинство говорящих на языке йоруба используют СЙ и дистанцируются от диалектов. Фактически, использование диалектов в общении в госучреждениях часто вызывает ироническую реакцию (Adeniyi K., Bamigbade O. E., 2019).

Прежде чем анализировать систему вокализма языка йоруба необходимо представить его алфавит, который состоит из двадцати пяти букв:

Aa [a] Bb [bi] Dd [di] Ee [e] Ẹẹ [ɛ] Ff [fi] Gg [gi] Gbgb [gb͡i] Hh [hi] Ii [i] Jj [dʒi] Kk [ki] Ll [li] Mm [mi] Nn [ni] Oo [o] Ọọ [ɔ] Pp [pi] Rr [ri] Ss [si] Ṣṣ [ʃi] Tt [ti] Uu [u] Ww [wi] Yy [ji] (ABD – Yoruba Alphabet Part 1; Yorùbá Alphabets (with Word Examples and Pictures) | Álífábẹ́ẹ̀tì Yorùbá; Yoruba (Èdè Yorùbá). Omniglot. The Online Encyclopedia of Writing Systems and Languages).

Основой алфавита языка йоруба стал латинский алфавит. Как видно, в алфавите йоруба отсутствуют буквы c, q, v, x, z (Yorùbá Alphabets (with Word Examples and Pictures) | Álífábẹ́ẹ̀tì Yorùbá). В целом, алфавит дает представление об орфографии языка.

В СЙ 18 согласных (b, t, d, k, g, kp, gb, j, m, n, f, s, ṣ, l, r, y, w, h), семь ртовых гласных (i, e, ẹ, a, o, ọ, u), пять назальных звуков (ĩ; ɛ̃; ã; ɔ̃; ũ), три уровневых тона (высокий, средний и низкий), а также два контурных тона (низкий восходящий и высокий нисходящий). Фонетические проекции диалектных разновидностей обычно выражаются разными степенями соответствия этим базовым компонентам звуковой системы СЙ, а также фонотактике СЙ. Например, в диалекте иджеша присутствует много дисиллабических существительных, которые начинаются с /u/, в том время как в СЙ это запрещено. Схожим образом, в диалекте ибадан нет *ṣ* (фонема /ʃ/), которая присутствует в СЙ и большинстве других диалектов. Мно-

гие люди, говорящие на диалекте ибадан, просто заменяют его с помощью s (/s/), даже если говорят на публике. Таким образом, это самый очевидный признак, показывающий, что люди разговаривают на диалекте ибадан (Adeniyi K., Bamigbade O. E., 2019).

4.2. Система гласных фонем языка йоруба

Как уже говорилось, в языке йоруба содержится семь ртовых гласных /i; e; ɛ; a; o; ɔ; u/ и пять назальных гласных /ĩ; ɛ̃; ã; ɔ̃; ũ/ (Yoruba (Èdè Yorùbá). Omniglot. The Online Encyclopedia of Writing Systems and Languages; Brief History of Yoruba Language. Nairature. Your all round Nigerian Literature resource). С. А. Эме и Д. Ю. Эбеле (C.A. Eme, D.U. Ebele) дают их фонетическое описание, орфографическую и фонемную репрезентацию, которые выглядят таким образом:

i /i/ закрытый неогубленный гласный переднего ряда

ìyá ‘mother’, orí ‘head’, ìta ‘outside’

e /e/ полузакрытый неогубленный гласный переднего ряда

ewé ‘leaf’, ejò ‘snake’, ikólè ‘dustpan’

ẹ /ɛ/ полуоткрытый неогубленный гласный переднего ряда

ẹpà ‘groundnut’, ẹgẹ ‘cassava’, ẹwà ‘beans’

a /a/ открытый неогубленный гласный смешанного ряда

àgè ‘kettle’, abà ‘hut’, adè ‘chair’

ọ /ɔ/ полуоткрытый лабиализованный гласный заднего ряда

ọpọlọ ‘frog’, ọbe ‘knife’, ọsàn ‘orange’

o /o/ полузакрытый лабиализованный гласный заднего ряда

okó ‘hoe’, ikókó ‘pot’, ólógbó ‘cat’

u /u/ закрытый лабиализованный гласный заднего ряда

ewúrẹ ‘goat’, kúrò ‘leave’, isu ‘yam’

in /ĩ/ закрытый неогубленный назальный гласный переднего ряда

èyìn ‘back’, ìgbín ‘snail’, rìn ‘to walk’

ẹn /ɛ̃/ half open front unrounded nasal vowel полуоткрытый неогубленный назальный гласный переднего ряда

ìyẹn ‘that one’, hẹn ‘yes’

an /ã/ открытый неогубленный назальный гласный смешанного ряда

ìtàn ‘story’, ràn ‘to send’, alákàn ‘crab’

ọn /ɔ̃/ полуоткрытый лабиализованный назальный гласный заднего ряда

ọọnì ‘crocodile’, ìbọn ‘gun’

un /ũ/ close back rounded nasal vowel закрытый лабиализованный назальный гласный заднего ряда

ràkúnmi ‘camel’, ẹkùn ‘tiger’ (Eme C.A., Ebele D.U., 2016, с. 74–75).

В йоруба нет дифтонгов, последовательности гласных звуков обычно произносятся как отдельные слоги. С орфографической точки зрения носовые звуки представлены ртовой гласной, за которой следует согласный «n», то есть in, un, en, on, an (Dingemanse M., 2006, с. 2–3).

4.3. Английский язык на современном этапе

После анализа системы вокализма необходимо дать характеристику английскому языку. АЯ входит в группу западногерманских языков индоевропейской языковой семьи и близко соотносится с такими языками, как фризский, немецкий и голландский (который в Бельгии называют фламандским). Английский язык зародился в Англии и на текущем этапе является главным языком Соединенных Штатов Америки, Соединенного Королевства, Канады, Австралии, Ирландии, Новой Зеландии и различных островных наций в Карибском море и Тихом океане. Это также официальный язык Индии, Филиппин, Сингапура и многих стран Африки, расположенных к югу от Сахары, включая и Южную Африку. Как правило, английский язык выбирают для изучения как иностранный в большинстве других стран мира. Такое положение дел дало АЯ статус глобального лингва франка (далее – ЛФ). Подсчитано, что треть всего населения в мире, около двух миллиардов людей, сейчас используют английский (Crystal D., Potter S.).

Следует также отметить, что английский не является просто лингва франка. Это конгломерат вариантов АЯ, у которых есть культурные и языковые особенности, в том числе и по отношению к субъектам коммуникации. Язык общения участников – это не изолированное явление, легко эксплицируемое через понятия из сферы международной коммуникации. Язык тесно связан с внеязыковыми явлениями и/или факторами прагматики. Эти два аспекта определяют становление английского ЛФ. Отмечается, что важно показать, как происходят изменения языка в рамках международного общения под влиянием сообществ субъ-

ектов, имеющих отличия между собой по разным характеристикам, в том числе, и по уровню владения языком (Мележик К. А., 2018, с. 92).

Для сопоставления с языком йоруба и НВАЯ был взят британский английский, так как Нигерия в прошлом была колонией Великобритании и именно на этом варианте АЯ разговаривали подданные этой страны. В британском английском выделяют 12 монофтонгов: [ɪ], [e], [æ], [ɒ], [ʌ], [ʊ], [ə], [iː], [uː], [ɑː], [ɔː], [ɜː]. Что касается дифтонгов, то их в английском языке их 8: [eɪ], [aɪ], [ɔɪ], [əʊ], [aʊ], [ɪə], [eə], [ʊə] (Бурая Е. А., 2009, с. 74).

4.4. Система вокализма нигерийского варианта английского языка

Система вокализма НВАЯ рассматривается в диссертации Мухаммада Мунзали Джибрила (Muhammad Munzali Jibril) на получение степени Доктора философии под названием «Фонологическое варьирование в НВАЯ» («Phonological Variation in Nigerian English») (Jibril M. M., 1982). На основе анализа речи образованных информантов из трех этнических групп – хауза, игбо и йоруба, которые проживают на территории Федеративной Республики Нигерия, М. М. Джибрил проанализировал фонемный нигерийского варианта английского языка. Третья глава его исследования «Системы вокализма и варианты гласных в НВАЯ» всецело посвящена анализу гласных фонем нигерийского варианта АЯ (Jibril M.M., 1982).

В целом следует отметить, что М. М. Джибрил не даёт единую классификацию гласных, которая подходила бы для использования на всей территории Нигерии всеми носителями нигерийского английского. Как упоминалось выше, для сопоставительного анализа были взяты три этноса, хотя количество этносов составляет 400. Народность хауза проживает на севере страны, а игбо и йоруба – в южных областях. Ввиду особенностей исторического развития так сложилось, что система вокализма НВАЯ хауза отличается от системы вокализма двух других этносов. АЯ, на котором говорят люди игбо и йоруба, используют сходную систему гласных, которую автор называет системой вокализма южного нигерийского варианта английского языка. Таким образом, М. М. Джибрил иллюстрирует на примере трех случаев тот факт, что в Нигерии присутствует фонологическая проблема с выделе-

нием гласных, которые бы использовались представителями всех народностей.

В описании систем вокализма НВАЯ хауза и южного НВАЯ автор также не ограничивается одной определенной классификацией каждой их них. Гласные южного нигерийского варианта английского языка сначала представлены таким образом: /(iː), ɪ, ɛ, eː, a, (aː), ɔ, (ɔː), o, uː, aɪ, ao, ɔɪ, (ɪə), (ɛə), (oə)/ (Jibril M.M., 1982, с. 151). Фонемы в скобках носят маргинальный характер, то есть не всегда употребляются большинством информантов.

Далее М. М. Джибрил проводит сопоставление систем вокализма базового английского у одного представителя этноса игбо и у одного представителя этноса йоруба. В результате получается такой набор гласных: /ɪ, ɛ, eː, a, ɔ, o, uː, aɪ, ɔɪ, ao, (ɪə), (ɛə), (oə)/ (Jibril M.M., 1982, с. 165). Если сопоставить эту систему, которая выведена на основе анализа речи двух информантов, и систему, которая была дана в предыдущем параграфе, то видно, что они почти совпадают. Единственное, что отсутствует в последней, – это маргинальные монофтонги /(iː), (aː), (ɔː)/.

После анализа гласных базового английского автор сопоставляет системы вокализма, используемые в речи образованных людей из народностей игбо и йоруба. Также используются данные по речи двух информантов, один из которых – игбо, другой – йоруба. Предлагается такой набор гласных фонем южного нигерийского варианта английского языка, на котором говорят люди с образованием: /iː, ɪ, ɛ, a [a] [æ], ʌ, ə, əː, ɑː/aː, ɔ, ɔː, o, uː, eɪ, aɪ, ɔɪ, ao, ɪə, ɛə, oə/ (Jibril M.M., 1982, с. 169).

5. Выводы

Сопоставление систем вокализма языка йоруба и БА показывает, что у них общим является одна фонема – /e/. Она также входит в состав дифтонгов /eɪ/ и /eə/. Фонема /a/ присутствует в дифтонгах /aɪ/ и /aʊ/. Гласный /ɔ/ в качестве одного из элементов формирует дифтонг /ɔɪ/. В остальном присутствуют различия. В йоруба имеются назальные гласные, которые отсутствуют в британском варианте АЯ. БА характеризуется наличием дифтонгов и долготой ряда монофтонгов, что отсутствует в йоруба.

Система вокализма базового южного НВАЯ, на котором говорят и представители народности йоруба, содержит такие фонемы: /ɪ, ɛ, eː, a, ɔ, o, uː, aɪ, ao, ɔɪ/. Если эти фонемы сопоставить

с гласными БА в количественном плане, то будут расхождения: в НВАЯ – десять единиц, в британском английском – двадцать. В качественном отношении система вокализма нигерийского варианта английского языка представляет собой пример интерференции двух языков: йоруба и БА. Гласные первого из них представлены такими фонемами: /ɛ, a, ɔ, o/. В фонемном составе языка йоруба присутствуют семь гласных и четыре из них входят в число фонем базового южного НВАЯ. Фонемы /ɪ, u:/ входят в систему вокализма НВАЯ со стороны британского английского. При этом следует отметить, что среди фонем йоруба есть краткая фонема /u/ и можно предположить, что этот факт способствовал приобретению качества долготы этим кратким гласным. Долгота, как качественный показатель, отсутствует в языке йоруба, но присутствует в пяти фонемах БА. В системе вокализма базового южного НВАЯ долгота как характеристика сохраняется и присутствует у двух фонем: /e:, u:/. Первая из них краткая в йоруба и БА. В ходе интерференции /e/ удлиняется. В британском английском языке – восемь дифтонгов. Среди гласных фонем базового южного нигерийского варианта АЯ – три дифтонга /aɪ, ao, ɔɪ/. Дифтонги /aɪ, ɔɪ/ присутствуют среди фонем БА, хотя первые элементы в этих гласных представлены и в языке йоруба. Фонема /ao/ отсутствует среди гласных британского варианта АЯ. Она сформирована из двух фонем, которые представлены в системе вокализма языка йоруба.

Если сопоставить систему вокализма южного НВАЯ, которая используется в речи образованных людей из этноса йоруба, с фонемами БА, то в количественном отношении они одинаковы – двадцать фонем в британском английском и девятнадцать в южном НВАЯ. Кратких гласных в БА и НВАЯ – по семь единиц, долгих – по пять. Число дифтонгов в британском АЯ – восемь, в южном нигерийском варианте АЯ – семь, то есть в качественном отношении наблюдается определенный баланс. Состав фонем также представляет собой интерференцию систем вокализма языка йоруба и БА.

ЛИТЕРАТУРА

Бурая Е. А. (2009) Фонетика современного английского языка. Теоретический курс: учебник для студ. вузов и фак. / Е. А. Бурая, И. Е. Галочкина, Т. И. Шевченко. 3-е изд., стер. М.: Издательский центр «академия». 272 с.

Мележик К. А. (2018) От глобального английского языка к национальному варианту английского лингва франка – проблемы коммуникативно-прагматической вариативности: Дис. … доктора филол наук. Спец. 10.02.04 – Германские языки. Симферополь. 635 с.

Петренко А. Д. (2018 а) Опыт изучения фонетической стратификации речи / А. Д. Петренко // Социофонетика и фоностилистика (опыт, актуальная проблематика, преспективы): монография / А. Д. Петренко, В. М. Бухаров, Д. А. Петренко [и др.]; под ред. д-ра филол. наук, проф. А. Д. Петренко.М.: ИНФРА-М. С. 13–69.

Петренко А. Д. (2018 б) Социофонетика: Крымский резонанс / А. Д. Петренко // Социофонетика и фоностилистика (опыт, актуальная проблематика, преспективы): монография / А. Д. Петренко, В. М. Бухаров, Д.А. Петренко [и др.]; под ред. д-ра филол. наук, проф. А. Д. Петренко. М.: ИНФРА-М. С. 5–12.

Петренко А. Д. (1998) Социофонетическая вариативность современного немецкого языка в Германии / А. Д. Петренко. К.: «Рідна мова». 254 с.

Петренко А. Д., Петренко Д. А. (2018) Билингвизм и вариативность языкового репертуара билингва / А. Д. Петренко, Д. А. Петренко // Иностранные языки в высшей школе. Научный журнал. Выпуск 4 (47). Рязань. С. 42–52.

Петренко Д. А. (2000) Произносительные особенности социолекта политических деятелей Германии: Дис. … канд. филол наук. Спец. 10.02.04 – Германские языки / Д. А. Петренко. Симферополь. 156 с.

ABD – Yoruba Alphabet Part 1 [Electronic Recource]. – Access mode: https://www.youtube.com/watch?v=KQTVj7tioEQ. – (Access date: 05.08.19).

Adeniyi K., Bamigbade O.E. (2019) Customised Ibadan-Yoruba [Electronic Recource] / K. Adeniyi, O.E. Bamigbade // Linguistik Online. Vol. 96. №3. Access mode: https://bop.unibe.ch/linguistik-online/article/view/3563/5413. (Access date: 01.08.19).

Brief History of Yoruba Language. Nairature. Your all round Nigerian Literature resource [Electronic Recource]. Access mode: http://nairature.blogspot.com/2015/10/yoruba-phonetics.html. (Access date: 02.08.19).

Constitution of the Federal Republic of Nigeria 1999 [Electronic Recource]. 118 p. Access mode:

https://www.wipo.int/edocs/lexdocs/laws/en/ng/ng014en.pdf. (Access date: 03.08.19).

Crystal D., Potter S. English language [Electronic Recource] / D. Crystal, S. Potter. Access mode: https://www.britannica.com/topic/english-language. – (Access date: 04.08.19).

Dingemanse M. (2006) The Body in Yorùbá. A Linguistic Study / M. Dingemanse. Leiden: Universiteit Leiden, Juli. 103 p.

Eme C.A., Ebele D.U. (2016) A Contrastive Study of the Phonology of Igbo and Yoruba / C.A. Eme, D.U. Ebele // UJAH: Unizik Journal of Arts and Humanities. Vol 17. №1. P. 65–84.

Jibril M.M. (1982) Phonological Variation in Nigerian English [Electronic Recource]. University of Lancaster, June. 364 p. Access mode: https://www.google.com.ua/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwimw9iMpPvjAhUOdJoKHWtLB74QFjAAegQIARAC&url=http%3A%2F%2Fkubanni.abu.edu.ng%2Fjspui%2Fbitstream%2F123456789%2F4296%2F1%2FPhonological%2520%2520%2520Variation%2520%2520%2520in%2520%2520%2520Nigerian%2520%2520%2520English.pdf&usg=AOvVaw1H_GtTkROSdHnpIE4f48rk. (Access date: 10.08.19).

The Oxford Companion to the English Language (1992) / Editor Tom McArthur. Oxford, New York: Oxford University Press. 1184 p.

Yorùbá Alphabets (with Word Examples and Pictures) | Álífábẹ́ètì Yorùbá [Electronic Recource]. Access mode: https://www.youtube.com/watch?v=OiCqse0M8xE. (Access date: 01.08.19).

Yoruba (Èdè Yorùbá). Omniglot. The Online Encyclopedia of Writing Systems and Languages [Electronic Recource]. Access mode: https://www.omniglot.com/writing/yoruba.htm. (Access date: 08.08.19).

Yoruba. The Language Gulper [Electronic Recource]. Access mode: http://www.languagesgulper.com/eng/Yoruba.html]. (Access date: 09.08.19).

УДК 81’33

EXTRACTING KEYWORDS AND PHRASES FROM TEXTS IN KAZAKH

Diana Rakhimova, Aliya Turganbayeva, Abylai Satybaldiev
Institute of Information and Computational Technologies, Almaty, Kazakhstan
Al-Farabi Kazakh National University, Almaty, Kazakhstan
di.diva@mail.ru, turganbaeva.aliya@bk.ru, nba_abilay.99@mail.ru

In this work, we consider the application of the statistical method of extracting keywords and phrases for the Kazakh language. Direct application of the statistical method tf-idf is not the optimal solution to the question of extracting keywords and phrases in the Kazakh language, since the Kazakh language is an agglutinative type of language. Considering the grammatical features of the Kazakh language, we carry out the pre-processing process, for this purpose we use the stemming algorithm developed by us. In the extraction, we also take into account the syntactic feature of the word or phrases using the morphological analyzer of the Kazakh language. The extraction of keywords and phrases specifically for the Kazakh language is an urgent task in classification, in clustering and in abstracting the text, and of course, in the search for information.

Based on the results of experiments in the work, we show that our approach is the best solution to the issue of extracting keywords and phrases from texts in Kazakh.

Keywords: Kazakh language, information retrieval, keyword extraction, extraction algorithm, tf-idf.

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ И СЛОВОСОЧЕТАНИЙ ИЗ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ

Диана Рахимова, Алия Турганбаева, Абылай Сатыбалдиев
Институт Информационных и Вычислительных Технологии, Алматы, Казахстан
Казахский Национальный Университет им. Аль-Фараби, Алматы, Казахстан
di.diva@mail.ru, turganbaeva.aliya@bk.ru, nba_abilay.99@mail.ru

В этой работе мы рассматриваем применение статистического метода извлечения ключевых слов и словосочетании для казахского языка. Прямое применение статистического метода tf-idf является не оптимальным решением вопроса по извлечению ключевых слов и словосочетаний в казахском языке, так как казахский язык относится

к агглютинативному типу языков. Учитывая грамматические особенности казахского языка, мы выполняем процесс предобработки. Для этого в работе применяется разработанный нами алгоритм стемминга. В извлечении учитывается также и синтаксический признак слово или словосочетания с помощью морфологического анализатора казахского языка. Извлечение ключевых слов и словосочетаний именно для казахского языка является актуальной задачей в классификации, в кластеризации и в реферировании текста, и конечно же, в поиске информации.

По результатам проведенных экспериментов в статье показываем, что предлагаемый нами подход является оптимальным решением вопроса по извлечению ключевых слов и словосочетании из текстов на казахском языке.

Ключевые слова: казахский язык, информационный поиск, извлечение ключевых слов, алгоритм извлечения, tf-idf.

Введение

В настоящее время объемы и динамика информации, которая подлежит обработке в лексикографии и информационном поиске, делают особенно актуальной задачу автоматического извлечения ключевых слов и словосочетаний, которые могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, кластеризация и классификации.

Анализ огромного количества данных может быть упрощен, если у нас будет ключевые слова или словосочетания, которые могут предоставить нам основные характеристики, концепцию и т. д. документа. Соответствующие ключевые слова и словосочетания могут служить кратким изложением документа и помогают нам легко упорядочивать документы и извлекать их на основе их содержания [1]. Необходимо различать два основных подхода к решению проблемы автоматизации выделения ключевых слов и словосочетаний: назначение ключевых слов и словосочетаний (keyphrase assignment) и их извлечение (keyphrase extraction) [2] [3]. Главное отличие заключается в том, что первый подход позволяет выделять только те ключевые слова и словосочетания, которые содержатся в некотором предусмотренном словаре, а второй подход предполагает выбор ключевой информации непосредственно из текста.

Ключевые слова могут быть назначены вручную или автоматически, но первый подход очень трудоемкий и дорогой. Таким образом, существует необходимость в автоматизированном про-

цессе, который извлекает ключевые слова из документов. Есть готовые программные решения этой задачи для распространенных языков (английский, русский, испанский и т.д.), а для казахского языка только единицы и они не в открытом доступе.

Программная реализация и результаты

Алгоритм извлечения ключевых слов из документов на казахском языке будет включать в себя 3 этапа:

1. нахождение кандидатов в ключевые слова;
2. выделение признаков;
3. ранжирование.

На первом этапе решаются 2 задачи:

1. предварительная обработка слов;
2. разделение текста на отдельные слова и словосочетания.

Первая задача является языкозависимой, поэтому здесь должен учитываться морфологическая особенность казахского языка. Для решения этой задачи задействованы система полных окончаний казахского языка, алгоритм стемминга и лемматизации для казахского языка [4]. А для второй использован простой подход – процедура токенизации, с помощью которой весь текст разбивается на отдельные слова.

На втором этапе для каждого найденного кандидата в ключевые слова выделяются признаки, по которым можно будет оценить степень его важности. Выделяемые признаки можно разбить на 3 категории: синтаксические признаки, статистические признаки, структурные признаки. Для выделения синтаксического признака мы использовали морфологический анализ казахского языка. И алгоритм TF-IDF (Term Frequency – Inverse Document Frequency) для определения частотности, то есть для выделения статистических признаков. В программе для биграммного слово указали следующие признаки: существительное+существительное, прилагательное+существительное, существительное+глагол, имя собственное+существительное, числительное+существительное (только некоторые слова).

На третьем этапе ранжируем результаты по статистическому признаку и в соответствующей мере к объему текста.

Реализация алгоритма выполнено на языке программирования Python. Программа запускалось в корпусе разделенный по тематике на казахском языке. Ниже в таблице приведены результаты корпуса по темам «Қазақстан тарихы (История Казахстана)»

(текст содержит 484 предложении), Манчестер сити (текст содержит 370 предложении), «Оскар алғандар (Получившие премию Оскара)» (текст модержит 309 предложении). В результате как показано в таблице приведены 10 ключевых слов и 5 ключевых словосочетании по каждому тексту.

Таблица 1. Полученные результаты.

Темы	Ключевые слова и словосочетания	Показатель tf-idf
Қазақстан тарихы	ғұн	0.013194415852460091
	тайпа	0.006085581972523395
	қытай	0.004175448054575978
	қағанат	0.004175448054575978
	түрік	0.0038034887328271213
	мемлекет	0.0035182270778650877
	күлтегін	0.0031733405214777436
	тонықөк	0.0031733405214777436
	ескерткіш	0.002672286754928626
	жорық	0.002672286754928626
	түрік қағанат	0.0033403584436607825
	ғұн мемлекет	0.002839304677111665
	қола дәуір	0.001336143377464313
	қазақстан жер	0.0011691254552812739
	тас дәуір	0.0010021075330982347
Манчестер сити	сити	0.02602130987211911
	манчестер	0.01520522323852743
	клуб	0.015048468359779727
	лига	0.005799930513665103
	кубок	0.0053296658774219866
	юнайтед	0.005172910998674281
	бапкер	0.0048594012411788704

Продолжение таблицы 1

	футбол	0.00313509757495411
	маусым	0.0029450594945452436
	ойыншы	0.002855815267437812
	манчестер сити	0.009405292724862329
	манчестер юнайтед	0.0047026463624311645
	есеп жең	0.0029783426962064043
	есеп жеңіл	0.0021945683024678767
	роберто манчини	0.0020378134237201712
Оскар алғандар	оскар	0.012804775360332537
	академия	0.005292640482270782
	номинация	0.005121910144133015
	фильм	0.0037560674390308773
	дауыс	0.0030731460864798087
	актер	0.0025609550720665075
	үздік	0.002438136624286715
	сыйлық	0.0024300195937471763
	американдық	0.00239022473392874
	рәсім	0.00239022473392874
	жыл үздік	0.004438988791581946
	оскар сыйлық	0.0029024157483420417
	1928 жыл	0.0027316854102042744
	дауыс бер	0.002219494395790973
	үздік жұмыс	0.0017073033813776715

Заключение

Разработан алгоритм извлечения ключевых слов из текстов на казахском языке. Данная работа найдет свое продолжение в проекте. Результаты задачи извлечение ключевых слов и словосочетании будут использоваться для решения задачи реферирования текстов на казахском языке.

Также разработанный алгоритм будем улучшать современными подходами в данной области.

Благодарность

Данная работа была выполнена и финансирована в рамках проекта AP05132950 «Разработка информационно-аналитической поисковой системы данных на казахском языке» Института информационных и вычислительных технологий, Казахстан, г. Алматы.

ЛИТЕРАТУРА

Шереметьева, С.О., Осминин П.Г., (2015). Методы и модели автоматического извлечения ключевых слов // Вестник Южно-Уральского государственного ун-та. – № 1, Т. 12. – С. 76–81.

Kaur, J., Gupta, V. (2010). Effective Approaches For Extraction Of Keywords // IJCSI International Journal of Computer Science Issues. Vol. 7. Issue 6. http://www.ijcsi.org/papers/7-6-144-148.pdf.

Beliga S. (2014). Keyword extraction a review of methods and approaches. http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approach es.pdf.

Ualsher Tukeyev, Diana Rakhimova, Aliya Turganbayeva, Dina Amirova, Balzhan Abduali, Aidana Karibayeva. (2018). Lexicon-free stemming for Kazakh language information retrieval // IEEE 12th International Conference on Application of Information and Communication Technologies. Almaty. – P. 95–98.

УДК 81’33

DEVELOPMENT OF METHODS OF INTEGRATION OF INFORMATION SYSTEMS USING ONTOLOGY OF THE SUBJECT AREA

Zh. B. Sadirmekova[a], J. A. Tussupov[a], A. M. Kemel[a], M. A. Sambetbayeva[a,b]

[a]Scientific-Research Institute «Artificial intelligence» L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan
[b]Institute of Information and Computational Technologies Almaty, Kazakhstan
janna_1988@mail.ru

This article discusses the development of a universal method of integration of information systems operating in one subject area, using metadata of information systems and ontology of the subject area.

Keywords: domain ontology, interoperability, integration, data storage.

РАЗРАБОТКА МЕТОДОВ ИНТЕГРАЦИИ ИНФОРМАЦИОННЫХ СИСТЕМ, ИСПОЛЬЗУЯ ОНТОЛОГИЮ ПРЕДМЕТНОЙ ОБЛАСТИ

Ж. Б. Садырмекова[a], А. Туссупов[a], А. М. Кемел[a], М. А. Самбетбаева[a,b]

[a]Евразийский национальный университет им. Л. Н. Гумилева. факультет информационных технологий, Нур-Султан, Казахстан,
[b]Институт информационных и вычислительных технологий, Алматы, Казахстан
janna_1988@mail.ru

В статье описывается разработка универсального метода интеграции операционных систем в определенной области, используя метаданные информационных систем и онтологию данной области.

Ключевые слова: онтология предметной области, совместимость, интеграция, хранение данных.

1. Main text

Information systems represent a wide class of software used by various enterprises to automate their work. Since the amount of

information processed is huge, now every organization has its own information system. This system is a complex software product that combines a variety of modern technologies for access, storage, data processing, networking and many others. During the development or merger of firms, their information systems must be scaled up and integrated. As you know, the same IS can be built in different ways, as well as one problem can be solved in different ways. As a result, integrated data stores may not be compatible, even if they perform similar tasks. The situation where an extension of the requirements or integration of system calls need to be redesign or create anew.

Currently, to solve this kind of problems, private means of computer data conversion are used, which are developed for a specific narrow task and do not analyze the converted information.

To date, there is a need to create a method that provides the unification of information systems, as well as a software tool that implements this Association.

2. Integration of information systems and the concept of interoperability

IS interoperability plays a fundamental role in the integration of information systems. Interoperability refers to the ability of an information system to interact with other is. This interaction can be expressed in the form of data sharing, distributed execution of search queries, a consistent database changes (DB), etc. the Need to ensure interope of realnosti occurs when linking the business processes of partner companies, according to the popping of the existing IS with the accepted standard solutions.

Also, the interoperability property is used for integration of several is, inclusion of previously used data warehouses in the created database system, development of complex automated control systems, construction of information storage networks, as well as in many other cases. The problem of IS interoperability is fundamental. It is relevant both for legacy systems that need to be associated with newly created ones (or, at least, to be able to use their databases), and for designed data warehouses, in which it is necessary to provide for the possibility of implementing interaction with other is in the future, when changing the requirements for them [1].

There are two aspects of interoperability: structural and semantic. Structural aspect of interoperability of systems refers to the ability to

structural harmonization of entity systems. The semantic aspect means the possibility of establishing a correspondence between the meanings of the units of information systems.

The existing methods for achieving interoperability are mainly concerned with its syntactic (structural) aspects, i.e. they are aimed at harmonization and pre-formation of data structures through standardization of their formats and the use of extensible metalanguages. There are currently no universal approaches to ensuring IS interoperability at the semantic level. Objectives are private, specific storage, and require manual construction of mappings between their entities, implemented in the mass data conversion.

3. Method of solving the problem ensuring interoperability IS

This study proposes a General solution to the problem of interoperability, by describing the IS metadata within the framework of the developed methodology, and the implementation of mapping entities and links of information systems into each other in terms of a common information field, given by the ontology of the subject area.

Since the knowledge stored in information systems is sufficiently structured, it is possible to build automated models and metamodels of this knowledge.

Conceptual models of information systems are created in accordance with XML and RDF schema standards. XML technology is used to formalize the structure and relationships in IS. A RDF-for the allocation and formalization of semantic units in specific subject areas of use of IS data. Conceptual models of information systems constructed in this way can be used to create a common meta-model that combines representations of entities of two or more data stores. Also define conversion rules of the entities and their relationships one is in the interpretation of the entities and their relationships the other IS.

4. XML and RDF technology

The most important merit of XML technology is that programs from different manufacturers were able to interact in one language. Instead of numerous disparate ways of presenting data, one universal syntax emerged, which formed the basis for the transfer of information between programs operating in different parts of the Internet. An

important quality of this standard is its openness and independence from specific fields of application and sections of knowledge. Its task is to enable users and programs to communicate with each other and with each other, without being limited to any particular subject area [2]. In turn, this universalism led to the creation of standard tools for supporting XML and additional technologies, as well as to the emergence of standard software interfaces for interaction with them. The use of XML technology allows to visualize the communication system, the hierarchy of concepts of the subject area in which the integrated information systems operate.

However, for all its advantages, XML is not able to become a suitable medium for expressing the semantics of marked data. By allowing any information to be encoded and allowing the developer to easily obtain a parser and data manipulation tools, XML satisfies the needs of programmers to have a universal markup tool that has syntactic interoperability (the ability to be a means of interaction between different programs). But it is not able to adequately cope with the task of semantic interoperability.

On the way to the implementation of the task of providing semantic interability, several difficulties can be identified. On the one hand, programs should understand the language of the relevant subject area, on the other – should be able to compare related terms of different subject areas. This requirement is significant, because otherwise programs would be able to work only with certain areas of knowledge, described, for example, specialized XML-languages. The purpose of semantic interoperability is to create a continuous information field.

Here we can give the following example. Suppose that one information system contains data about a company and its employees, another system contains information about people, and a third system contains information about addresses. Obviously, companies, people and addresses belong to separate, relatively independent fields of knowledge. On the other hand, in a continuous information field, the program should easily be able to compare employees and people, the addresses of these people and the subject area of addresses as such.

So, if syntactic interoperability is inextricably linked to parse data, requires semantic analysis of the information, its internal coherence, to establish compliance with the terms and vocabularies in one subject area to the members of another.

XML cannot be a means of communicating different data for a number of reasons. Its main limitation is that XML only describes

grammar. It is impossible to distinguish a semantic unit in a specific subject area, since this language is focused on the structure of the document and does not imply a General interpretation of the data contained in it. XML turns out to be too flexible a means of describing data and allows the same information to be labeled in different ways.

Summing up, it should be noted that for semantically interoperable information systems, in which programs can automatically analyze the content of resources, a new means of expressing the semantics of data, not just their recording, is necessary.

These problems can be solved by using XML together with another data model, such as a semantic network model, to define metadata and data transformation rules when moving from one integrated IS to another. Formally, the semantic network can be defined using the RDF model. Resource description Framework (RDF) was developed to solve problems related to the description of semantics. Fundamental to RDF is the concept of a data model. It is a set of facts and semantic connections between them. The basic building block of a data model is a statement that represents a triple: a resource, a named property, and its value. In RDF terminology, these three parts of the statement are called respectively: subject, predicate, and object. A resource is anything described by RDF, such as a single table or a part of it.

A property should be understood as an aspect, characteristic, attribute, or relation used to describe a resource. Each property has its own specific meaning, valid values, the type of resources to which it can be applied, as well as relations with other properties. RDF notations developed by manufacturers are based on XML.

To better understand the relationship of RDF to XML and other serialization languages, the following analogy can be used. The knowledge that is present in a person's head does not depend in any way on the way it is transmitted to other people. For example, it could be expressed in English, and it can be in Russian. In this abstraction, the RDF data model is equivalent to knowledge, and XML is equivalent to English, which, although only one of the possible ways of representation, has the status of an international means of communication. The two existing XML notations in this case can be compared to different dialects of the same language.

In the RDF model, concept names are selected from a specific concept dictionary and namespace, so their representation is more unified than XML markup. Also, in the proposed model, one concept of the subject area corresponds to a set of concepts-synonyms of the

subject area. This helps to avoid differences in the representation of their relationships without losing the completeness of the representation.

Thus, the joint use of the RDF model with XML technology allows to reflect the semantics of conceptual models of information systems, as well as to avoid the stated limitations of XML.

5. Domain ontology

The mechanism for creating a continuous information field is ontology. Ontology includes a set of terms and rules according to which these terms can be combined to build reliable statements about the state of the system in question at some point in time. In addition, on the basis of these statements, appropriate conclusions can be drawn, allowing to make changes to the system, to improve the efficiency of its functioning [3].

In any system, there are two main categories of objects of perception, such as the objects themselves that make up the system (physical and intellectual) and the relationship between these objects that characterize the state of the system. In terms of ontology, the concept of relationship uniquely describes the relationship between the objects of the system in the real world, and the terms, respectively, describe the real objects themselves. The ontological model represents the most important statements in the subject area. Additionally, this model helps to describe the behavior of objects and the corresponding change in relationships between them; that is the behavior of the system. Thus, ontology is a dictionary of data that includes both terminology and model of system behavior. Since each conceptual model of the domain is a subset of ontology, the problem of integration of information systems is reduced to the problem of combining metamodels of information systems, that is, the construction of maps between these metamodels, in terms of ontology.

In this problem, ontology serves to build correspondences between the concepts of information systems. Ontology also helps to establish links between semantic units within each conceptual domain model of integrated is when defining metadata. Ontology, as well as IS metadata, is described on the basis of XML technology and RDF model. Ontology development was carried out in accordance with the IDEF5 standard. After defining the metadata of information systems and building a common metamodel of data warehouses, it becomes possible to interpret information from one IS by means of another IS.

Thus, interoperability and, consequently, the necessary level of integration of information systems is ensured.

6. Diagram of integration of IS

The developed algorithm of integration of information systems consists of the following basic steps:

1. Analysis of database entities, their attributes and relationships between them. At this stage, the application is building data schemas.

2. Analysis of semantic values of entities and attributes. At this stage, conceptual models of information systems are developed. Domain ontology is used to obtain and analyze semantic meanings.

3. Clarification of semantic correspondences. Ontology is used to identify the missing links between concepts.

4. Building a single metamodel. This metamodel is constructed as a Union of two conceptual models of information systems. At this stage, ontology is used to resolve possible contradictions.

5. Output of resulting mappings between entities and attributes of information systems.

In this diagram (Fig. 1) integration of *IS-A* and *IS-B*. the task of integration is to ensure interaction between IS. To do this, it is necessary to determine the correspondence of is-a entities to is-B entities and the rules of their transformation.

To this end, data schemas are first extracted from information systems. By analyzing data schemes separately, it is possible to establish only structural interoperability, that is, rules for converting field types and entities of information systems into each other. To ensure semantic interoperability, it is necessary to understand the purpose of IS elements. Therefore, it is necessary to use the second component of the metadata-the conceptual model of the subject area (SOFTWARE). It is an add – on to the data schema and defines the system of links between the concepts of the subject area established in this IS. The construction of this model is carried out with the help of domain ontology. The ontology contains a dictionary of concepts and stores the network of relationships between these concepts. That is, each conceptual model is a subset of the SOFTWARE ontology. The use of ontology makes it possible to define conceptual models in the same terms and analyze the connections between their concepts.

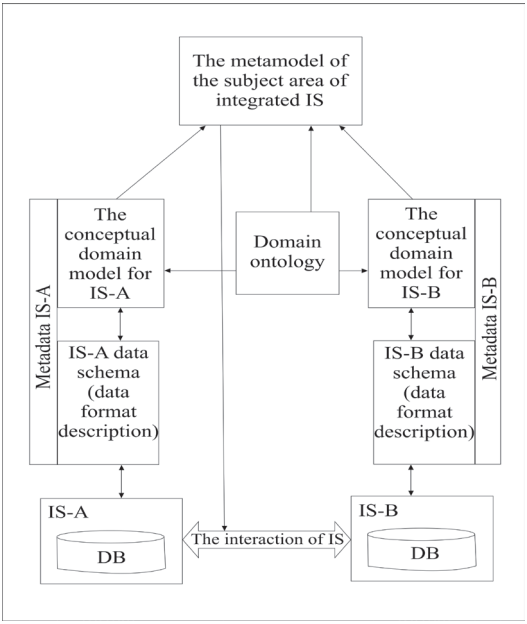

Fig. 1. Scheme of algorithm of integration of two information systems

Further, on the basis of conceptual models of *IS-A* and *IS-B*, the construction of a metamodel of the subject area of integrated is is carried out. This metamodel integrates and adapts itself to both the conceptual model. Ontology is also used at the stage of its construction. A metamodel defines the matching entity *IS-A* entity FROM-TO and rules of their transformation, which allows to establish the interaction between information systems.

7. Conclusion

As a result of this study, a methodology for ensuring interoperability of information systems was developed. After performing this procedure, the IS is included in a single information field, set by ontology, and is able to interact with other IS from the subject area.

The developed methodology can significantly accelerate and simplify the integration of information systems. It also ensures the successful evolution of IS in the course of enterprise development.

The work is supported by the grant of funding of scientific and (or) scientific and technical research for 2018-2020. MES RK (№ AP 05133546), as well as RFBR (№18-07-01457 A).

REFERENCES

[1] Fedotov A. M., Tusupov J. A., Sambetbayeva M. A., Sagnayeva S. K., Bapanov A. A., Nurgulzhanova A. N., Yerimbetova A. S. Using the thesaurus to develop it inquiry systems // Journal of Theoretical and Applied Information Technology. – 2016. – Vol.86, issue 1, – P. 44–61.

[2] Fedotov A. M., Tusupov J. A., Sambetbayeva M. A., Fedotova O. A., Sagnayeva S. K., Bapanov A. A., Tazhibaeva S. Z. Classification model and morphological analysis in multilingual scientific and educational information systems // Journal of Theoretical and Applied Information Technology. – 2016. – Vol.86, issue 1, – P. 96–111.

[3]W3C Recommendation 10 February 2004, RDF/XML Syntax Specification, http://www.w3.org/.

УДК 81’33

ABOUT THE CREATION OF AN ELECTRONIC RESOURCE OF WORKS BY RUSSIAN-SPEAKING WRITERS OF THE REPUBLIC OF BASHKORTOSTAN

Z. A. Sirazitdinov
IALI of the Ufa FIT RAS, Russian Federation, Bashkortostan, Ufa
sazin11@mail.ru

The article considers the issue of creating on the Internet a full-text electronic resource of prose of Russian-speaking authors of the Republic of Bashkortostan. The Russian-language writers are defined as the writers of the Bashkir and other nationalities, living in the republic and writing their works in Russian. The time frames for works to be included in the resource are set, preliminary classification signs of texts in the database are given. There is a positive tendency among developers of informational literary resources towards universality, expressed in the combination of information for the general reader and for the literary scholar. The implementation of corpus linguistics’ potentials in the resources will allow introducing corpus materials of regional authors for scientific research.

The circle of tasks that will be solved by an electronic literary resource is determined.

Keywords: literary criticism, Russian-speaking authors, prose, database, information resource.

О СОЗДАНИИ ЭЛЕКТРОННОГО РЕСУРСА ПРОИЗВЕДЕНИЙ РУССКОЯЗЫЧНЫХ ПИСАТЕЛЕЙ РЕСПУБЛИКИ БАШКОРТОСТАН

З. А. Сиразитдинов
Институт истории, языка и литературы УФИЦ РАН, РФ, Башкортостан, Уфа
sazin11@mail.ru

В статье рассматривается вопрос создания в сети Интернет полнотекстового электронного ресурса прозы русскоязычных авторов Республик Башкортостан. Определяется круг русскоязычных писателей, как пишущие на русском языке литераторы из башкир и других национальностей, проживающих на территории республики, устанавливаются временные рамки для произведений, которые будут включены в ресурс, даются предварительные классификационные признаки текстов в базе данных. Отмечается положительная тенденция среди разработчиков информационных литературных ресурсов к универсальности, выражающаяся в соединении информации для обычного читателя и для исследователя-литературоведа. В связи с этим ставится

на перспективу задача реализация в ресурсах возможностей корпусной лингвистике, что позволит ввести в научный оборот корпусные материалы региональных авторов.

Ключевые слова: литературоведение, русскоязычные авторы, проза, база данных, информационный ресурс.

Современное развитие человечества определяется информационной революцией, в котором большинство членов занято производством, хранением, переработкой и реализацией информации. Международный опыт показывает, что информационные и телекоммуникационные технологии, уже стали локомотивом социально-экономического развития многих стран мира, а обеспечение гарантированного свободного доступа граждан к информации – одной из важнейших задач государств. Учитывая это, Правительством РФ принята Стратегия информационного общества в Российской Федерации от 7 февраля 2008 г. [1], в которой определены цели и направления деятельности в области развития информационного общества в стране. В Стратегии целью и задачами формирования и развития информационного общества в РФ обозначены также и сохранение культуры многонационального народа Российской Федерации, укрепление нравственных и патриотических принципов в общественном сознании, развитие системы культурного и гуманитарного просвещения.

Создание научными, образовательными и культурными учреждениям общедоступных гуманитарных ресурсов и баз данных, в частности в области языкознания и литературоведения, отвечает вышеуказанным целям и задачам Стратегии и ее принципам партнерства государства и гражданского общества и свободы и равенства доступа к информации и знаниям.

Функционирующие в Российском сегменте сети Интернета электронные полнотекстовые информационные ресурсы литературных произведений представлены в виде электронных библиотек художественных произведений, авторских страниц писателей, сайтов литературно-художественных журналов. Среди них популярностью пользуются такие ресурсы как электронная библиотека художественной литературы «Библиотека Максима Мошкова» [2], литературные порталы современной русской поэзии и прозы [3-4], портал сетевой словесности «Современная русская литература в Интернете» [5], авторские ресурсы Бориса Акунина, Виктора Астафьева [6-7], сайты литературно-художественных журналов «Нева» , «Новое литературное обозрение» [8-9].

В Республике Башкортостан функционируют сайты литературной газеты «Истоки» и журнала «Бельские просторы» [10-11], в которых печатаются русскоязычные авторы республики. В архивах этих изданий имеются электронные версии изданий за определенные годы, доступ к ним платный, поэтому говорить о свободном и всеобъемлющем доступе ко всем произведениям авторов, которые печатались в этих СМИ, не приходится. Отметим, часть русскоязычных авторов активно печатаются и в федеральных СМИ.

К сожалению, все существующие литературные Интернет ресурсы не позволяют делать выборки по времени издания, тематике произведения и месту жительства литературных деятелей, которые несомненно являются важными для анализа не только общелитературоведческих вопросов, но и для выявления региональных особенностей авторов их и вкладов в общий культурный процесс страны. Изучение региональной литературы как локально-регионального феномена культуры «позволяет противостоять глобализации и унификации национальных культур и российской культуры, составить целостное представление об общероссийском литературном процессе» [12: с.3].

В связи с этим было бы желательно иметь электронный ресурс региональных авторов, пишущих на русском языке. Данный ресурс на первом этапе должен сделать доступным читателям всех категорий, в том числе и специалистам-литературоведам, произведения русскоязычных авторов Республики Башкортостан.

В последнее время среди разработчиков информационных литературных ресурсов наблюдается стремление к универсальности, выражающееся в соединении информации для обычного читателя и для исследователя-литературоведа, что является «одним из основных направлений развития информационных ресурсов в области художественной литературы и литературоведения» [13]. Учитывая это, в перспективе данный ресурс представляется содержащим критические и научные статьи по творчеству писателей и средства лингвистического поиска. Реализация в ресурсах возможностей корпусной лингвистике было бы не только новым явлением, но и позволило бы ввести в научный оборот корпусные материалы региональных авторов. Сегодня, к сожалению, тексты русскоязычных авторов национальных республик не находят всеобъемлющего отражения в Национальном корпусе русского языка.

Русские писатели Башкортостана и пишущие на русском языке литераторы из башкир и других национальностей, проживающих

на территории республики, вносят значительный вклад в общую национальную культуру республики, знакомя русскоязычного читателя с башкирской действительностью. В эпоху информационной открытости через произведения этих авторов не только многонациональный народ Российский Федерации, но и все мировое сообщество открывает для себя самобытную национальную культуру башкир.

Часть русскоязычных писателей республики являются уроженцами Башкортостана: Ю. Андриянов, Р. Ахмедов, М. Гафуров, Р. Паль, А. Филиппов, М. Чванов, Д. Швецов – всю жизнь и творческую деятельность связали с республикой, другие же, как А. Генатулин, М. Львов, Я. Мустафин, Б. Романов, В. Сорокин, Р. Хуснутдинова, Л. Шикина, впоследствии переехали в Москву. А такие литераторы, как Г. Шафиков, В. Денисов, А. Докучаева, Б. Павлов, В. Перчаткин, И. Слободчиков, И. Сотников, В. Трубицын, родившиеся за пределами Башкортостана, переехали в республику и творческую деятельность связали со второй малой родиной.

Численность русскоязычных писателей в республике неуклонно растет, что объясняется наличием печатных литературных СМИ и возможностью свободной публикации произведений в бесплатных Интернет ресурсах. Отражением этого является существование в республике нескольких общественных объединений русскоязычных писателей: объединение при Союзе писателей РБ, Уфимское литературное объединение (УфЛИ), Уфимское литературное созидательное содружество (УЛИСС).

Создание полного электронного ресурса русскоязычных писателей Башкортостана позволит представить полную картину культурно-духовных связей башкир и русских в Республике Башкортостан. Отметим, что определенные попытки такого представления на основе выборочного анализа доступных материалов для периода 19 – начала 20 веков позволили выявить предпосылки в развитии башкирско-русских контактных литературных явлений рассматриваемого периода и раскрыть закономерности образования литературного синтеза и роли контактных связей в эволюции башкирской и русской словесности [14: с. 5].

Привлечение больших полнотекстовых литературных данных для сравнительного анализа русскоязычных и башкироязычных писателей позволит полнее изучить контактные литературные отношения, поскольку исследования в этом русле помогают решать актуальные проблемы не только литературно-историческо-

го, но и ценностнокультурного характера [14, с. 20]. Создаваемый электронный ресурс может способствовать пониманию многих закономерностей и особенностей возникновения и развития башкирско-русских литературных контактных связей и в наше время.

Электронный ресурс должен содержать полнотекстовые произведения с классификацией по жанрам. Проблема классификации художественной прозы является спорной и как отмечают литературоведы «установление систем и классификаций жанров всегда будет сохранять опасность случайности и субъективизма» [15: с. 215]. Однако мы считаем, что прозаические тексты должны включать жанровые классификационные характеристики. В литературоведении жанры прозаического текста достаточно полно определены, ядерную часть которого составляют роман, повесть, рассказ, новелла, очерк, притча, эссе, зарисовка, фельетон, литературная сказка. Для эпических жанров характерно тематическое деление, они могут быть фантастическими, детективными, философскими, историческими, приключенческими [16: с. 42]. Учитывая это, нами в жанровую классификацию включается и тематика прозаических произведений.

Каждый текст художественного произведения наряду с указанием автора, должен иметь дату первой публикации и источник публикации в виде названия печатного издания (журнал, литературная газета, альманах), издательства, электронного ресурса в котором опубликован.

Хронологические рамки электронного ресурса, по-видимому, должны определяться как XIX–XXI вв. Начало хронологической рамки как XIX в. выбрано с целью охватить творчество таких писателей как С. Т. Аксакова, П. М. Кудряшева, В. С. Юматова, В. С. Лосиевского, М. Л. Михайлова, М. В. Авдеева и других, выходцев с исторического Башкортостана, в чьих произведениях нашли отражение национальные черты башкирского народа, его быт, природа края и взаимоотношения с русским народом.

Такой электронный ресурс в сети Интернет в свободном доступе будет играть роль важную роль в культурной жизни Республики Башкортостан, позволит популяризации произведений, отражающих интернациональную тему в современной отечественной литературе, явится одной из актуальных мер в создании и укреплении единой российской нации, являющейся союзом разных народов РФ, крепко связанных историческими и культурными узами.

ЛИТЕРАТУРА

1. Стратегия развития информационного общества в Российской Федерации от 7 февраля 2008 г. № Пр-212 [электронный ресурс]. URL: https://rg.ru/2008/02/16/informacia-strategia-dok.html. (дата обращения: 10.10.2019).

2. Библиотека Максима Мошкова [электронный ресурс]. URL:http:// http://lib.ru/ (дата обращения: 10.10.2019).

3. Портал Стихи.ру [электронный ресурс]. URL:http://www.stihi.ru (дата обращения: 10.10.2019).

4. Российский литературный портал Проза.ру [электронный ресурс]. URL:http://www.proza.ru (дата обращения: 10.10.2019).

5. Портал сетевой словесности «Современная русская литература в Интернете» [электронный ресурс] URL:http://www.netslova.ru) (дата обращения: 10.10.2019).

6. Борис Акунин. Сочинения. Полное интерактивное собрание [электронный ресурс]. URL: http://www.akunin.ru (дата обращения: 10.10.2019).

7. Сайт фонда им. Виктора Петровича Астафьева [электронный ресурс]. URL: http://www.astafiev.ru (дата обращения: 10.10.2019).

8. Литературный журнал «Нева» [электронный ресурс]. URL: http:// http://www.nevajournal.ru/ (дата обращения: 10.10.2019).

9. Литературный журнал «Новое литературное обозрение» [электронный ресурс]. URL: https://www.nlobooks.ru (дата обращения: 10.10.2019).

10. Литературная газета «Истоки» [электронный ресурс]. URL: https://www.istokirb.ru (дата обращения: 10.10.2019).

11. Литературный журнал «Бельские просторы» [электронный ресурс]. URL: https://bp.rbsmi.ru (дата обращения: 10.10.2019).

12. Прокофьева И. О. Уфимская художественная проза конца XIX – начала XX века: состояние и функционирование. Автореф. Дис. Канд. Филол. Наук. Екатеринбург, 2016, 25 с.

13. Буранбаев А. М. Информационные ресурсы в области художественной литературы и литературоведения в сети Интернет: опыт анализа // Современная техника и технологии. 2016. № 2 [Электронный ресурс]. URL: http://technology.snauka.ru/2016/02/9493 (дата обращения: 10.10.2019)].

14. Абидова Э.Х. Башкирско-русские контактные литературные связи XIX – начала XX веков. Автореф. Дисс. Канд. Филол. Наук. Уфа, 2011. 23 с.].

15. Хализев, В. Е. Теория литературы / В. Е. Хализев. – М.: Высшая школа, 2002 – 438 с.

16. Шатрова Е. Д., Ласица Л. А. К проблеме определения жанров произведений фанфикшн// Вестник Оренбургского государственного университета 2017 № 1 (201), С. 41–48.

УДК 81’33

ON RELATIVE CLAUSE IN TURKISH

Ercan Solak
Işık University, İstanbul, Turkey
ercan.solak@isikun.edu.tr

In the tradition of the generative theories of grammar, Relative Clause (RC) in Turkish poses interesting challenges. Traditionally, RC in Turkish has been treated as transformation and movement based structure. This approach has led to a dual analysis where the relativized constituents are roughly divided into objectivized or subjectivized heads. In addition to not providing a theoretical basis for such a distinction, the transformation/ movement treatment cannot be easily extended to oblique objects. As a result, one ends up with a set of disparate relativization rules. Moreover, the semantic content is often mixed with syntactic structure to explain the choice of a particular RC construction. In this paper, I propose a unified analysis of RC in Turkish without a recourse to transformation or movement. I propose to treat RCs as complex Noun Phrases with their heads derived from stemmed sentencial clauses. Thus, my approach eliminates the need to assume an *a priori* non-relativized clause that lies beneath the surface form. Also, the distinction between the object and subject relativization disappears. I illustrate the consistency of the analysis through several examples with their syntax trees.

Keywords: Relative Clause; Participles; Syntax trees; Turkish.

ОБ ОТНОСИТЕЛЬНОМ ПРИДАТОЧНОМ ПРЕДЛОЖЕНИИ В ТУРЕЦКОМ ЯЗЫКЕ

Эрджан Солак
Университет Ышык, Стамбул, Турция
ercan.solak@isikun.edu.tr

В традиции теории порождающей грамматики относительное придаточное предложение (RC) в турецком языке ставит интересные задачи. Традиционно RC в турецком языке рассматривается как структура, основанная на трансформации и движении. Этот подход привел к двойственному анализу, где релятивизированные составляющие делятся на объективизированные и субъективизированные. В дополнение к отсутствию теоретической основы для такого разделения, отношение к трансформации/движению не может распространяться на косвенные предметы. В результате получается набор разрозненных правил релятивизации. Более того, семантический контент часто смешивается с синтаксической структурой, чтобы объяснить выбор определенной конструкции RC. В данной статье предлагается унифицированный

анализ RC в турецком языке без обращения к трансформации или движению. Данный анализ предполагает относится к RC как к сложным именным словосочетаниям. Таким образом, данный подход исключает необходимость принятия нерелятивизированного предложения, лежащего ниже поверхностной формы, априори. Также исчезает различие между объективной и субъективной релятивизацией. Последовательность анализа демонстрируется на нескольких примерах с их синтаксическими деревьями.

Ключевые слова: относительное придаточное предложение; причастие; синтаксические деревья; турецкий язык.

1. Introduction

Turkish relative clause (RC) has generally been treated in a gapping and movement context (Kornfilt, Turkish 1997). In this view, relative clauses are constructed by transforming a finite clause and changing its semantic content to one of an adjective phrase. This view results in a range of complex and distinct rules to explain its construction for different cases of relativization targets. Moreover, these rules stand in contrast to the rest of the Turkish grammar, which has a rather regular morphosyntactic structure.

Two of the earliest attempts to analyze the syntax of relativization in Turkish were (Underhill 1972) and (Hankamer ve Knecht 1976) which provided an exhaustive lists of ways in which RCs are constructed. These early accounts do not provide a unified view of the different relativization and participle selection processes.

(Güngördü ve Engdahl 1998) offers a HPSG account of the relativization in Turkish using lexically modified MOD values, valency lists, and non-local feature handling. An analysis of RCs within the framework of Minimalist Program is proposed in (Çağrı 2005). More recently, (Özçelik 2016) gives an antisymmetric analysis of relative clauses.

In this paper, we explain the relativization in Turkish by viewing participle morphemes as derivations acting on stemmed verb phrases where the finite verb is stripped of its tense and person morphemes. A claim similar to our present proposal was forwarded in (Kornfilt, Agreement – subject case correlations in Turkish and beyond 2008) for the case of relativized object. We claim that the apparent asymmetry of relativizations with respect to subject and object corresponds to two different strategies of deriving noun phrases.

The rest of the paper is organized as follows. In the next section, we give a short description of subject and object relativization in Turkish. In Section 3, we highlight the problems of the traditional view in the

case of relativization out of constituent. In section 4, we present our proposal for the analysis of Turkish relative clauses. We end the paper with concluding remarks.

2. Relative Clauses in Turkish

Turkish is a head final language with a prevalent SOV word order where scrambling is also licensed. Turkish also has a rich morphology where the syntactic functions are usually marked with case suffixes.

The most prevalent view of relative clause in Turkish is to invoke a relativization process which moves and gaps the constituents of a sentence that contain no relative structures and to arrive at a relative clause. The relativized constituent is referred to as the target of relativization. The example sentences (1) and (2) illustrate the transformation.

(1) Adam-lar su iç-iyor-lar. Man-
PL water drink-IMPF-3PL Men
are drinking water.

In order to relativize the target subject ``adam-lar» (men) in (1), we move it to immediately after the verb and the verb is modified with one of the participle morphemes -(y)An, -DIK, -(y)AcAK. In the example (2), the subject is the target of relativization and so the sentence is relativized using -(y)An as in (2).

(2) su iç-en adam-lar.
water drink-PART man-PL
the men who are drinking water.

In this relativization process, apart from the constituent movement, a single participle morpheme -(y)An replaces both the tense and person morphemes of the verb. In the traditional view of the Turkish relative clause, the -(y)An morpheme derives an adjective from a verb stem.

If the target of relativization is a direct object of the verb, -DIK participle suffix is used. In this case, a GEN-POSS structure is used as well as the movement. The sentence (3) is the object-relativized transformation of the sentence in (1), where, this time, the target of relativization is the indefinite object ``su» (water).

(3) adam-lar-ın iç-tiğ-i su man-PL-GEN
drink-PART-P3SG water the water that
the men are drinking

Note that, in object relativization, a genitive morpheme is appended to the subject of the original phrase, «adam-lar» (men).

The most frequent occurrences of relativization involves the two cases where the target of relativization is the subject or object of the verb.

3. Relativizing within a constituent

In the transformation view, construction of relative clauses are further parameterized by how the target of relativization functions within its surrounding constituent. A particularly common case is when the possessor of a subject possessive NP is the target. This process is illustrated in sentences (4) and (5).

(4) Kız-ın kalem-i düş-tü
Girl-GEN pencil-P3SG fall-PAST-A3SG
Girl's pencil fell.

In the original sentence (4), when moving to after the verb «düş-tü», the possessor «kız-ın» looses its genitive morpheme while the possessed constituent «kalem» retains its possessive morpheme –I.

(5) kalem-i düş-en kız pencil-
P3SG fall-PART girl the
girl whose pencil fell

Moreover, the person agreement within the NP is lost when its possessor is relativized out of the NP. For example, in sentences (6) and (7), the possessed «babası» (his/her father) retains its third person marking while its relativized possessors become «sen» (you) and «ben» (me), respectively.

(6) baba-sı gel-en sen father-
P3SG come-PART you you,
whose father came
(cf. sen-in baba-n gel-di)
your father came

(7) baba-sı gel-en ben father-
P3SG come-PART me I,
whose father came
(cf. ben-im baba-m gel-
di) my father came

Another interesting case is when the target of relativization is a case-marked oblique object of the verb. In such cases, as in the possessive NP, the case marking of the target drops when it moves to after the verb. Moreover, possessive markers are inserted after the participle. In the following example, the dative case marker -DA drops and possessive marker -(I)m is inserted after the participle. Sentence (8) illustrates this case.

(8) uyu-duğ-um yatak
sleep-PART-P1SG bed
the bed that I sleep in
(cf. yatak-ta uy-uyor-um)
bed-DAT sleep-IMPF-1SG

4. A new analysis of relative clause

In order to unify the disparate relative clause constructions of Turkish exemplified in the previous section, we make the following two claims.

1. There are no participles in Turkish. All participles should be treated as suffixes deriving noun phrases out of stemmed verb phrases.
2. There are no relative clauses in Turkish. All relative clauses can be treated as noun phrases.

In the rest of this section, we justify these claims.

4.1 –(y)An mopheme

The participle suffix –(y)An, when appended to single verb stems, derives an adjective with the semantics of actor, similar to the -er suffix in English. Examples of this derivation are, «gel-en» (comer), «otur-an» (sitter), «sor-an» (asker).

In Turkish, the boundary between adjectives and nouns are not clear cut. Any adjective can syntactically stand for a noun and conversely many nouns can serve as adjectives before other nouns. The latter is more common in idiomatic constructions. For example, the qualified noun «adam» (man) in the subject of the sentence (9) is dropped in (10) and the adjective «zengin» (rich) serves as the subject.

(9) zengin adam konuş-ur
rich man speak-AOR-
3SG the rich man speaks

(10) zengin konuş-ur
rich speak-AOR-3SG
the rich one speaks

For nouns serving as adjectives, examples are «parmak çocuk» (finger kid), «ressam adam» (painter man), «ada ülke» (island country).

In this paper, we propose to treat the adjectival participle –(y)An as a derivational suffix modifying a VP with its verb stemmed. It produces a NP out of a stemmed VP. The NP thus produced can stand on its own or serve as an adjective before a common noun.

As an example, consider the phrase in (11).

(11) ev-de kal-an çocuk
house-DAT stay-PART child
the child who stays (stayed) home

Figure 1 illustrates the new structure in the parse tree of the sentence (11). In Figure 1, we use several new tags to help us with the new analysis of the –(y)An participle. They are

V-S : verb stem,
VP-S : verb phrase with its verb stemmed,
-ER : adjectival suffix (corresponding to -(y)An)) that attaches to VP stems.

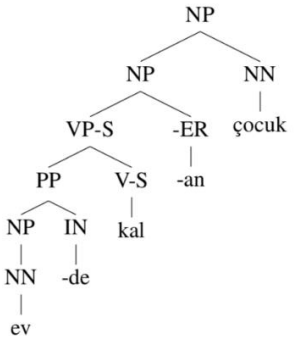

Fig. 1. The parse of «ev-de kal-an çocuk» with the new tags VP-S, V-S and -ER

In order to consistently parse the sentences like (5), (6) and (7), we break the long distance dependency between the -P3SG morpheme and the noun that the RC qualifies. Instead, we treat -P3SG as completely contained within its surrounding clause. Thus, -P3SG here generically

marks a possessed noun whose possessor left indeterminate. Such a view, immediately explains the case of headless RC's, (Kornfilt, Turkish 1997).

In order to emphasize the lack of person agreement in the possessive marker, we use -WP$ as its tag. This also accords with its counterpart in English which tags the possessive relative pronoun «whose.»

Using the analysis that we propose with the new set of tags, the parse of the sentence (7) is given in Figure 2.

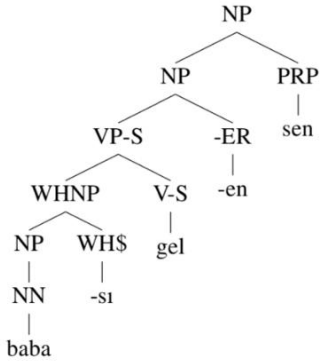

Fig. 2. The parse tree of «babası gelen sen»

4.2 -DIk and –(y)AcAK morphemes

There are plenty of examples to suggest that words derived from verb stems using -DIK morpheme behave like nouns in contexts other than relative clauses. An example sentence is (12). In the analysis of the sentence (12), we intentionally used -DIK morpheme instead of the -PART as done in traditional analysis.

(12) biz-e ye-dik-ler-in-i iç-tik-ler-in-i söyle
us-DAT eat-DIK-PL-P2SG-ACC drink-DIK-PLU-P2SG-ACC
tell tell us what you ate and drank

In this paper, we look for a unifying analysis that treats -DIK as a noun generating morpheme even in RC contexts. Indeed, such an analysis is supported by the presence of possessive morpheme after -DIK and the attachment of genitive morpheme to the subject of the RC as in (13).

(13) sen-in gel-diğ-in-i duy-du-m
you-GEN come-DIK-P2SG-ACC
I heard you came.

Thus, the words derived by the -DIK morpheme obey the noun inflection paradigm.

For (13), an alternate yet a more literal translation would be «I heard your coming.» Thus, -DIK roughly corresponds to -ing gerundive form for verbs in English. For the present analysis, we will use -INF tag to denote the -DIK morpheme. This puts -DIK morpheme in the same lexical category as the infinitive morpheme -mA as in «gel-me-si» (his coming). The difference between -DIk and -mA is one of a semantics. While the morpheme-DIk denotes a past or present action, the morpheme -mA denotes an action without a specific time. Similarly, -AcAK morpheme specifies the infinitive form of an action in the future.

The parse tree for the sentence (13) is given in Figure 3.

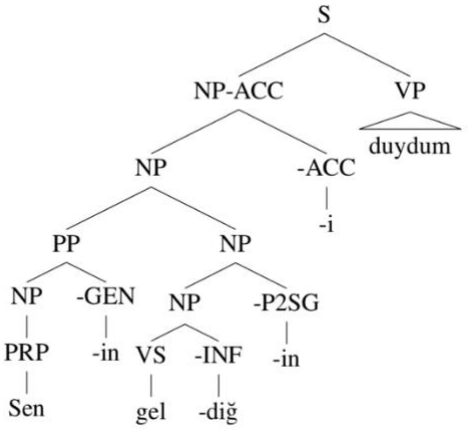

Fig. 3. The parse tree of «Senin geldiğini duydum»

A particular difficulty in the transformational view of relative clauses is the insertion of auxiliary «ol-» (be) in relativizing nominal sentences. For example, in relativizing the sentence (14), we are forced to insert the verb «ol-» as in (15).

(14) Ali sen-in öğretmen-in.
Ali you-GEN teacher-P2SG.
Ali is your teacher.

(15) Sen-in öğretmen-in ol-an Ali you-GEN teacher-P2SG be-PART Ali Ali who is your teacher.

Such an insertion is impossible to arrive at by any movement and gapping of the original sentence (14) even when we imagine a dropped copula at the end as «Ali senin öğretmenin-*dir*» (Ali *is* your teacher). However, in our new analysis of the relative clause, we decouple (14) and (15) and treat them as having different syntaxes albeit with similar semantics. The parse trees of the sentences (14) and (15) are given in Figure 4.

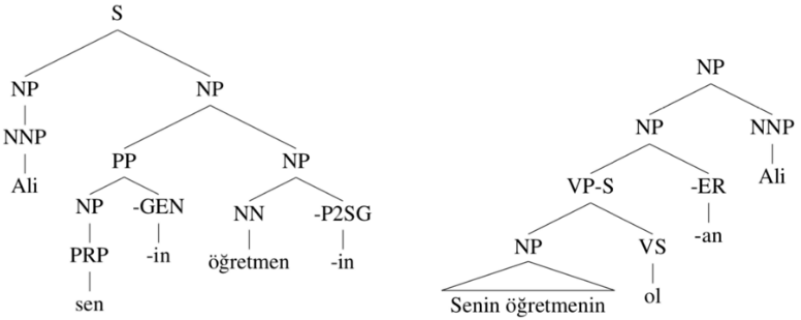

Fig. 4. The parse trees of the sentences «Ali senin öğretmenin» and «senin öğretmenin olan Ali»

5. Conclusion

Relative clauses in Turkish are tricky structures. Up to now, their syntactic structures have not been analyzed under a consistent framework. In this paper, we proposed to resolve the inconsistency by claiming that there are actually no relative clauses in Turkish. Such structures are not sentential clauses but rather noun phrases. As for the morphemes -(y)An, DIK and –(y)AcAK, traditionally viewed as participles used in relativization, we proposed a new view that treats them as productive suffixes deriving NP's out of verb phrase stems. The resulting view results in consistent parses for (traditional) relative clauses. Moreover, our new analysis easily explains the insertion of genitive morphemes and auxiliary verbs in relativization.

REFERENCES

Çöltekin, Çağrı. 2014. «A set of open source tools for Turkish natural language processing.» Ninth International Conference on Language Resources and Evaluation. Reykjavik: European Language Resources Association. 1079–1086.

Çağrı, M., İ. 2005. Minimality and Turkish relative clauses. University of Maryland.

Özçelik, Ö. 2016. «An antisymmetric analysis of Turkish relative clauses: implications from prosody.» Turkic Languages, 3 87–99.

Özenç, Berke. 2019. DiaMor. 15 August. https://github.com/berkeozenc/Diamor.

Özenc, Berke, Razieh Ehsani, and Ercan Solak. 2018. «Moraz: an open-source morphological analyzer for Azerbaijani Turkish.» 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels: Association for Computational Linguistics. 25–29.

Güngördü, Z, and E. Engdahl. 1998. «A relational approach to relativization in Turkish.» Joint Conference on Formal Grammar, HPSG and Categorial Grammar.

Hankamer, J., and Laura Knecht. 1976. «The role of the subject/non-subject distinction in determining the choice of relative clause participle in Turkish.» NELS 4. 123–135.

Kessikbayeva, Gulshat Cicekli, Ilyas. 2016. «A RuleBased Morphological Analyzer and a Morphological Disambiguator for Kazakh Language.» Linguistics and Literature Studies 96–104.

Kornfilt, J. 2008. Agreement – subject case correlations in Turkish and beyond. Leipzig Spring School on Linguistic Diversity: Topics in Turkic Syntax. –. 1997. Turkish. Abingdon : Routledge.

Koskenniemi, Kimmo. 1983. Two Level Mophology: A General Computational Model for Word-Form

Recognition and Production. Helsinki: University of Heksinki.

Linden, Krister, Erik Axelson, Sam Hardwick, Miikka Silfverberg, and Tommi Pirinen. 2011. «HFST– framework for compiling and applying morphologies.» Second International Workshop on Systems and Frameworks for Computational Morphology. Zurich: Springer. 77–85.

Matlatipov, Gayrat, and Zygmunt Vetulani. 2009. «Representation of Uzbek Morphology in Prolog.» In Aspects of Natural Language Processing, by Marciniak M. and Mykowiecka A., 83–110. Berlin: Springer.

Oflazer, Kemal. 1994. «Two-level description of Turkish morphology.» Literary and Linguistic Computing 137–148.

Tantuğ, A. Cüneyd, Eşref Adalı, and Kemal Oflazer. 2006. «Computer Analysis of the Turkmen Language Morphology.» Advances in Natural Language Processing. Finland: Springer. 186–193.

Underhill, R. 1972. «Turkish participles.» Linguistic Inquiry, 3(1) 87–99.

Washington, Jonathan, Mirlan Ipasov, and Francis Tyers. 2012. «A finite-state morphological transducer for Kyrgyz.» Eighth International Conference on Language Resources and Evaluation. İstanbul: European Language Resources Association. 934–940.

УДК 81’33

ON THE DEVELOPMENT OF THE SEMANTIC-SYNTACTIC ANALYZER OF THE TATAR SENTENCE: THE RULES OF CONTEXT-FREE GRAMMAR

D. Sh. Suleymanov, A. R. Gatiatullin
Academy of Sciences of the Republic of Tatarstan, Russian Federation, Tatarstan, Kazan
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

The article describes a new stage in the development of a semantic-syntactic analyzer of simple sentences of the Tatar language. The analyzer uses a combined approach and in the process of functioning, both formal grammars and relational-situational models are connected to represent semantic information. Thanks to the addition of the linguistic base with semantic information, the number of ambiguities arising in the case of conventional parsing is reduced.

Keywords: semantic-syntactic analyzer, Tatar language, trees of immediate components.

О РАЗРАБОТКЕ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА ТАТАРСКОГО ПРЕДЛОЖЕНИЯ: ПРАВИЛА КОНТЕКСТНО-СВОБОДНОЙ ГРАММАТИКИ

Д. Ш. Сулейманов, А. Р. Гатиатуллин
Академия наук Республики Татарстан, РФ, Татарстан, Казань
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

В статье описывается новый этап разработки семантико-синтаксического анализатора простых предложений татарского языка. В анализаторе используется комбинированный подход и в процессе функционировании для представления семантической информации подключаются как формальные грамматики, так и реляционно-ситуационные модели. Благодаря дополнению лингвистической базы семантической информацией, уменьшается количество неоднозначностей, возникающие в случае обычного синтаксического анализа.

Ключевые слова: семантико-синтаксический анализатор, татарский язык, деревья непосредственных составляющих.

Введение

В настоящее время в области обработки естественного языка наблюдается бум технологий, основанных на нейронных сетях

и машинном обучении. Эти технологии пытаются использовать в любой области, где производится обработка больших объемов данных. Главное преимущество нейронных сетей – нет необходимости в детальной формализации знаний, формализация заменяется обучением на примерах. Однако, при этом наблюдается трудность вербализации результатов работы нейронной сети и объяснений, почему она приняла то или иное решение, а также невозможно гарантировать повторяемость и однозначность получения результатов.

Еще одной особенностью систем машинного обучения является потребность в больших объемах данных с готовой морфологической, семантической и синтаксической разметкой со снятой многозначностью. Кроме того, большой объем лингвистических баз данных, полученных в процессе разработки лингвопроцессоров, основанных на формальных методах, также могут быть использованы в других задачах компьютерной обработки естественного языка.

Этот набор перечисленных требований и то, что для многих малоресурсных языков отсутствуют электронные корпусы с синтаксической и семантической разметкой, являются обоснованием того, что наряду с технологиями машинного обучения продолжаются работы по созданию программного обеспечения, основанного на правилах и формальных методах. Публикации последних нескольких лет [1-4] подтверждают, что продолжается создание семантико-синтаксических анализаторов для русского языка. Для тюркских языков мы не нашли разработок семантико-синтаксического анализа. Имеется лишь ряд работ по созданию чисто синтаксических анализаторов для тюркских языков. В основном это разработки для турецкого, уйгурского и казахского языков [5-6].

Семантико-синтаксический анализатор татарского предложения, представленный в этой статье, разрабатывается для предложений татарского языка, на которые накладывается ряд ограничений. На данном этапе анализатор разрабатывается для простых предложений татарского языка без причастных и деепричастных оборотов.

Если рассмотреть стандартные синтаксические анализаторы, то они в результате своей работы получают на выходе синтаксические деревья двух видов: деревья зависимостей или деревья непосредственных составляющих. На рис. 1 представлен пример дерева непосредственных составляющих для предложения на русском языке.

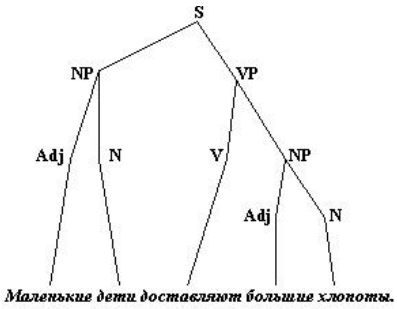

Рис. 1. Пример дерева непосредственных составляющих для русского языка

У каждого из типов синтаксических деревьев есть свои преимущества и свои недостатки. Особенностью деревьев непосредственных составляющих является способность отображать порядок следования элементов в предложении.

Отличием семантико-синтаксического анализатора, представленного в данном проекте, от классических синтаксических анализаторов является то, что в нем присутствует семантическая информация и эта семантическая информация получается не на следующем этапе анализа после синтаксического, а параллельно с ней. Для получения такого результата используются реляционно-ситуационные модели и все простые предложения татарского языка рассматриваются в виде некоторых ситуаций. Таким образом, составляющими в данном дереве становятся не просто именные или глагольные группы, а группы, обозначающие в текущей ситуации определенные семантические роли.

Для обозначения видов ситуаций и семантических ролей в нашей модели анализа введена своя система обозначений. Эта система обозначений используется как для представления результатов анализа, так и при определении правил контекстно-свободной грамматики, которые определяют алгоритм работы анализатора. В качестве первоначальной версии системы обозначений ситуаций (ситуационных фреймов) использована система обозначений, представленная в ресурсе FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/). Данная система обозначений взята в связи с тем, что это один из наиболее проработанных и известных ресурсов с описанием семантических ситуационных

фреймов. В дальнейшем в процессе работы над анализатором происходит корректирование и уточнение этой системы в соответствии со структурой и семантикой татарского языка, с особенностями работы нашего анализатора и необходимостью использования этих результатов для разметки электронного корпуса татарского языка.

Рассмотрим систему обозначений, использованную в нашем семантико-синтаксическом анализаторе.

Для отличия названий ситуаций всем именам добавлен префикс S.

Например:

S_Filling – ситуация, описывающая процесс заполнения некоторой сущности.

S_Motion – ситуация, описывающая процесс движения.

Основным элементом каждой ситуации является предикат, который в данной модели семантико-синтаксического анализа представляется именем ситуации и префиксом P.

Например:

P_Filling – предикат, выражающий процесс заполнения некоторой сущности,

P_Motion – предикат движения.

Семантические роли в представленной нами семантико-синтаксической модели подразделяются на три типа. Роли, представленные в предложении подлежащим и обозначающие исполнителей ситуации, обозначаются с помощью префикса A.

Например:

A_Actor – исполнитель ситуации,

A_Cause – причина возникновения ситуации.

Оставшиеся роли подразделяются на те роли, которые входят в ядро предиката и те, которые не входят в ядро. Ядро ситуации образуют роли, которые формируют специфику данной ситуации, а роли, не входящие в ядро – это роли, которые присущи всем ситуациям. Роли, не входящие в ядро, могут обозначать, например, место, время или способ свершения ситуации.

Роли, входящие в ядро:

C_Goal – конечная точка движения объекта,

C_Source – начальная точка движения объекта.

Роли, не входящие в ядро ситуационного фрейма:

NC_Time – роль, выражающая время выполнения действия или события,

NC_Place – роль, выражающая место выполнения действия или события.

Программа семантико-синтаксического анализа выполнена в виде нескольких процедур, выполняющихся последовательно. Последовательность данных процедур представлена на рис. 2.

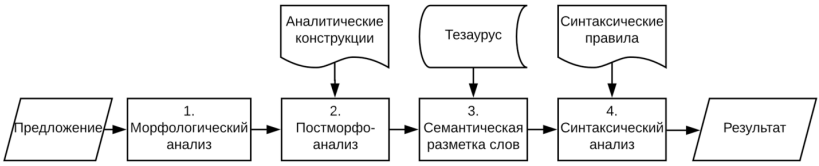

Рис. 2. Последовательность работы анализатора

Рассмотрим эти этапы.

На вход анализатора поступает предложение на татарском языке, с которым производятся последовательные преобразования и в результате на выходе должно получиться синтаксическое дерево с дополнительной семантической информацией. Первый этап – это этап морфологического анализа [7] на выходе которого получаются морфологические структуры отдельных словоформ татарского языка.

На вход поступило предложение: *Машиналар авылдан шәһәргә юл буйлап баралар 'Машины едут из деревни в город вдоль дороги'*. На выходе морфологического анализа получен следующий текст:

Машиналар
машина+N+PL(ЛАр)+Nom;
авылдан
авыл+N+Sg+ABL(ДАн);
шәһәргә
шәһәр+N+Sg+DIR(ГА);
юл
юл+N+Sg+Nom;
буйлап
буй+N+DISTR(лАп);буй+PROP+DISTR(лАп);буйла+V+ADVV_ACC(ЫП);буйлап+POST;
баралар
бар+V+PRES(Й)+3PL(ЛАр);

Как видно из этого примера, результат морфологического анализа содержит варианты неоднозначностей. В данном случае словоформа *буйлап* может быть проанализирована как имя существительное, глагол и послелог.

Постморфоанализ осуществляет выявление в тексте аналитических форм и аналитических конструкций. Этап семантической разметки слов осуществляет семантическую классификацию словоформ, выделенных аналитических форм или аналитических конструкций. Эта классификация осуществляется с использованием тезауруса и помогает уточнять роли, которые способны замещать эти словоформы в ситуации в зависимости от их значения. Например, в зависимости от того к какому классу в тезаурусе относится сущность, она может играть роль Агента или активной причины A_Cause.

Завершающим этапом, представленным на данном этапе, является этап построения непосредственно синтаксических деревьев. Выполнение этого этапа производится с использованием библиотеки NLTK [8]. Для осуществления синтаксического анализа используются правила контекстно-свободных грамматик. В нашей программе правила имеют комбинированное представление, содержащие как синтаксическую, так и семантическую информацию. Набор всех правил образует лингвистическую базу данных.

Рассмотрим эти правила.

Первое правило показывает, что любое простое предложение является описанием одного из типов ситуации:

S -> S_Arriving | S_Burying | S_Cause_benefit_or_detriment | S_Cause_bodily_experience | S_Cause_change | S_Cause_harm | S_Cause_impact | S_Cause_motion | S_Cause_temperature_change | S_Cause_to_be_dry | S_Cause_to_be_included | S_Cause_to_be_sharp | S_Cause_to_be_wet | S_Cause_to_end | S_Cause_to_fragment | S_Cause_to_make_noise | S_Cause_to_move_in_place | S_Cause_to_resume | S_Cause_to_rot | S_Cause_to_wake | S_Control | S_Corroding_caused | S_Cotheme | S_Creating | S_Cure | S_Damaging | S_Departing | S_Destroying | S_Dispersal | S_Downing | S_Emptying | S_Endangering | S_Erasing | S_Filling | S_Fleeing | S_Fluidic_motion | S_Grinding | S_Imposing_obligation | S_Infecting | S_Intentional_traversing | S_Intentionally_affect | S_Interrupt_process | S_Killing | S_Mass_motion | S_Motion | S_Motion_directional | S_Motion_noise

| S_Objective_influence | S_Placing | S_Quitting_a_place | S_Removing | S_Render_nonfunctional | S_Rejuvenation | S_Reshaping | S_Self_motion | S_Setting_back_burn | S_Setting_fire | S_Travel | SN_Container_focused_placing | SN_Container_focused_removing | SN_Transitive_action

Данное правило показывает, что предложение S может выражать одну из ситуаций перечисленного списка. В данном примере приведен фрагмент перечисления ситуаций. Названия ситуаций разделены символом дизъюнкции '|'.

Например ситуация S_Motion – это ситуация движения физического объекта, а S_Cause_motion – ситуация, являющаяся причиной движения физического объекта. Действующие лица самих ситуаций представляются соответствующими правилами, которые рассмотрим далее в статье.

При создании правил формирования структуры предложения в реляционно-ситуационной модели нами использовано несколько основных гипотез:

1. Предикат в предложении стоит в конце предложения.

2. Зависимое самостоятельное слово стоит в тексте слева от главного.

3. Зависимое служебное слово (послелоги и послеложные слова, служебные глаголы) стоят справа от главного.

4. Подлежащее, в отличие от английского языка, не всегда стоит вначале предложения.

Правила для описания структуры самих ситуаций представляют возможную комбинаторику предиката и его семантических ролей.

Приведем пример одного из правил для описания ситуации движения:

S_Motion -> A_Theme P_Motion EOS | A_Theme NC_Speed P_Motion EOS | A_Theme C_Direction P_Motion EOS | A_Theme C_Direction NC_Speed P_Motion EOS | A_Theme C_Goal P_Motion EOS | A_Theme C_Goal C_Source P_Motion EOS | A_Theme C_Direction C_Source P_Motion EOS | A_Theme C_Source C_Goal P_Motion EOS | NC_Time A_Theme C_Source C_Goal P_Motion EOS | A_Theme NC_Time C_Source C_Goal P_Motion EOS | A_Theme C_Source C_Goal NC_Time P_Motion EOS | A_Theme C_Goal C_Source NC_Time Path P_Motion EOS | A_Theme C_Source C_Goal NC_Time Path P_Motion EOS | A_Theme NC_Time C_Source C_Goal Path P_Motion EOS | A_Theme NC_Time C_Goal C_Source Path P_Motion EOS | A_Theme NC_Time C_Goal C_Source Path

Distance P_Motion EOS | A_Theme NC_Time C_Source C_Goal Path Distance P_Motion EOS

В данном примере аббревиатура EOS обозначает конец предложения. Далее следующие правила определяют свойства ролей, участвующих в описании ситуации.

Правило **P_Motion** -> V[lemma=’бар’] показывает, что предикат P_Motion выражается глаголом с основой *бар ‘идти’*.

В ситуации движения **S_Motion** описывается движение некоторой сущности **A_Theme.**

Остальные роли в ситуации делятся на 2 типа: роли, которые входят в ядро, и роли, которые не входят в ядро.

Ролями, которые образуют ядро ситуации движения, кроме самого движущегося объекта, являются еще следующие роли:

C_Direction -> N[last_affix=DIR] POST[last_affix=POST, lemma=’таба’] | Adj N[last_affix=DIR] POST[last_affix=POST, lemma=’таба’] – роль, которая показывает направление движения. Это правило показывает, что данная роль выражается именем существительным N в форме директива и послелогом *таба ‘в сторону’*.

Например:

Урманга таба ‘в сторону леса’

C_Goal -> N[last_affix=DIR] | Adj N[last_affix=DIR] – роль, которая показывает конечную цель движения и выражается именем существительным в форме директива (DIR).

Например:

авылга ‘в деревню’

C_Source -> N[last_affix=ABL] | Adj N[last_affix=ABL] – роль, которая показывает начальную точку движения и выражается с помощью имени существительного в форме Аблатива (ABL).

Например:

шәһәрдән ‘из города’

Ролями, которые не входят в ядро ситуации движения являются роли времени NC_Time и скорости NC_Speed. В правилах ниже представлено, чем могут быть выражены эти роли в тексте:

NC_Speed -> Adv[lemma=’тиз’] | Adv[lemma=’кызу’] | Adv[lemma=’җитез-җитез’] | Adv[lemma=’җәлт-җәлт’] | Adv[-lemma=’әкрен’] | Adv[lemma=’акрын’] | Adv[lemma=’акрын’] | Adv[lemma=’бик’] Adv[lemma=’тиз’] | Adv[lemma=’бик’] Adv[lemma=’кызу’] | Adv[lemma=’җитез-җитез’] | Adv[lem-

ma=’җәлт-җәлт’] | Adv[lemma=’бик’] Adv[lemma=’әкрен’] | Adv[-lemma=’бик’] Adv[lemma=’акрын’]

NC_Time -> Adv[lemma=’бүген’] | Adv[lemma=’быел’] | Adv[-lemma=’былтыр’] | Adv[lemma=’иртәгә’] | Adv[lemma=’көндез’] | Adv[lemma=’төнлә’] | Adv[lemma=’иртән’] | Adv[lemma=’кичен’] | Adv[lemma=’язын’] | Adv[lemma=’көзен’]

Как представлено в этих правилах, роли, которые выражают скорость и время свершения действия, могут быть выражены с помощью наречия (Adv). Причем, структурных особенностей, свойственных всех элементам, представленным в этой роли, эти наречия не имеют, поэтому для того, чтобы была возможность выделить эти роли, необходимо указывать их лемму.

Таким образом, лингвистические базы данных с формальным и структурно-функциональным описанием позволяют получить различные виды лингвистических баз данных, которые описывают особенности татарского языка. Тестирование этого анализатора на реальных текстах также позволяет проверить достоверность получаемых данных.

Результатом работы анализатора является дерево непосредственных составляющих, которое может быть представлено разных форматах. В графическом формате, формате XML или внутреннем представлении NLTK. На рис. 3 представлен результат анализа во внутреннем формате NLTK.

В данном примере на вход анализатора поступает предложение на татарском языке:

Машиналар былтыр шәһәрдән авылга урман аркылы бардылар. ‘В прошлом году машины ехали из города в деревню через лес’.

```
(S[]
  -->(S_Motion[]
        -->(A_Theme[]
          -->(N[case=’Nom’, last_affix=’PL’, lemma=’машина’] Машиналар))
        -->(NC_Time[] (Adv[last_affix=’Adv’, lemma=’былтыр’] былтыр))
       -->(Source[] (N[last_affix=’ABL’, lemma=’шәһәр’] шәһәрдән))
      -->(Goal[] (N[last_affix=’DIR’, lemma=’авыл’] авылга))
     -->(Path[]
        -->(N[case=’Nom’, last_affix=’Sg’, lemma=’урман’] урман)
        -->(POST[last_affix=’POST’, lemma=’аркылы’] аркылы))
        -->(P_Motion[] (V[last_affix=’3PL’, lemma=’бар’] бардылар))
  -->(EOS[] )))
```

Рис. 3. Результат семантико-синтаксического анализа

Выданный анализатором результат представляет следующую актантную структуру.

Ситуация: S_Motion
Исполнитель (A_Theme): *Машиналар*
Цель (Goal): *авылга*
Исходная точка (Source): *шәһәрдән*
Путь (Path): *урман аркылы*
Время совершения действия: *былтыр*
Предикат (P_Motion) *бардылар*

Заключение

В статье описан ряд результатов, полученных при разработке семантико-синтаксического анализатора для анализа простого татарского предложения. Полученные результаты планируется использовать для разметки электронного корпуса, информационного поиска и других задач. Полученный результат легко переносим на другие тюркские языки.

Данная статья выполнена при поддержке Российского фонда фундаментальных исследований РФФИ 18-47-160014 «Разработка интегральной компьютерной модели и программного инструментария для семантико-синтаксического анализа татарских текстов».

ЛИТЕРАТУРА

1. Боярский, К. К. Семантико-синтаксический парсер SemSin / К. К. Боярский, Е. А. Каневский // Научно-технический вестник информационных технологий, механики и оптики. – 2015. – Т. 15. – № 5. – С. 869–876.

2. Осипов Г. С., Шелманов А. О. Метод повышения качества синтаксического анализа на основе взаимодействия синтаксических и семантических правил // Труды шестой международной конференции «Системный анализ и информационные технологии» (САЙТ). – Т. 1. – 2015. – С. 229–240.

3. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies / K. V. Anisimovich, K. Ju. Druzhkin, F. R. Minlos et al. // Papers from the Annual International Conference «Dialogue» (2012). – Vol. 2. – 2012. – P. 91–103.

4. Kuznetsov, V. A. Ontological-semantic text analysis and the question answering system using data from ontology / V. A. Kuznetsov, V. A. Mochalov, A. V. Mochalova // ICACT Transactions on Advanced

Communications Technology (TACT). – 2015. – Vol. 4, Issue 4. – P. 651–658.

5. Eryiğit G., J. Nivre, and K. Oflazer, Dependency Parsing of Turkish, Computational Linguistics, vol. 34, Sep. 2008, pp. 357–389.

6. Başer, Z. (2019). A universal parser or language specific parsing strategies: A study on relative clause attachment preference in Turkish. RumeliDE Dil ve Edebiyat Araştırmaları Dergisi , () , 1–21 . DOI: 10.29000/rumelide.648403

7. Гатауллин, Р. Р. Гибридный морфологический анализатор татарского языка на основе правил и статистики / Р. Р. Гатауллин // Научно-технический вестник Поволжья. № 9 2018 г. – Научно-технический вестник Поволжья, 2018. – С. 89–92.

8. Perkins, Jacob. Python Text Processing with NLTK 2.0 Cookbook. – Packt Publishing, 2010.

УДК 81’33

STRUCTURAL-PARAMETRIC COMPUTER MODEL OF THE TURKIC MORPHEME AS THE BASIS OF THE MULTIFUNCTIONAL MULTILINGUAL INTERNET SERVICE

D. Sh. Suleymanov, A. R. Gatiatullin
Academy of Sciences of the Republic of Tatarstan,
Russian Federation, Tatarstan, Kazan
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

The development of natural language processing programs requires a large volume of linguistic databases. The article describes the implementing of a multifunctional, multilingual linguistic Internet service based on the structural-parametric functional model of the Turkic morpheme, in terms of representing Turkic attributive linguistic elements (adjectives and adverbs). The main objectives of the service are the creation of a linguistic resource base for software products that perform computer processing of Turkic languages, such as machine translation systems, information retrieval systems, electronic corpus annotation systems, data extraction, etc.

Keywords: multifunctional Internet service; Turkic languages; language units; technological tools.

СТРУКТУРНО-ПАРАМЕТРИЧЕСКАЯ КОМПЬЮТЕРНАЯ МОДЕЛЬ ТЮРКСКОЙ МОРФЕМЫ КАК ОСНОВА МНОГОФУНКЦИОНАЛЬНОГО МНОГОЯЗЫЧНОГО ИНТЕРНЕТ-СЕРВИСА

Д. Ш. Сулейманов, А. Р. Гатиатуллин
Академия наук Республики Татарстан, РФ, Татарстан, Казань
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

Для создания программного обеспечения, работающего с естественными языками, необходимо наличие лингвистических баз данных большого объема. В статье описывается развитие многофункционального, многоязычного лингвистического интернет-сервиса на базе структурно-параметрической функциональной модели тюркской морфемы, в части представления тюркских атрибутивных лингвистических элементов (прилагательных и наречий). Основной задачей сервиса являются формирование лингвистической ресурсной базы для технологий, осуществляющих компьютерную обработку тюркских языков, таких как системы машинного перевода, информационно-поисковые системы, системы разметки электронных корпусов, извлечения данных и др.

Ключевые слова: многофункциональный интернет-сервис; тюркские языки; языковые единицы; технологический инструментарий.

Введение

В настоящее время проблема создания лингвистических баз данных для малоресурсных языков стоит особо остро, что связано, в первую очередь, с появлением эффективных методов обработки естественного языка на основе нейронных технологий и, соответственно, с необходимостью больших объемов языковых данных для машинного обучения. Практически все языки тюркского семейства, кроме турецкого языка, можно отнести к малоресурсным языкам. Очевидно, отсутствие баз знаний с большим объемом текстов существенно ограничивает возможности использования современного программного инструментария и технологий для обработки большинства тюркских языков. Соответственно, для внедрения тюркских языков в инфокоммуникационные технологии требуется решение задач создания множества лингвистических баз данных, в то время как для многих европейских языков такие ресурсы уже созданы. Кроме того, в силу структурных особенностей тюркских языков, многие модели и технологии, разработанные для европейских языков, неэффективны для прямого использования для языков тюркского семейства. Таким образом, актуальной является задача создания компьютерных моделей, лингвистических ресурсов и технологий обработки тюркских языков с учетом их структурно-функциональных особенностей. Все названные задачи являются междисциплинарными, и очевидно, что они эффективно могут быть решены только совместными усилиями специалистов в областях информационных технологий, математики и лингвистики.

Согласно мониторингу публикаций по внедрению тюркских языков в инфокоммуникационные технологии, наиболее активно ведутся работы для турецкого, казахского, татарского языков. Созданием моделей, технологий, и программного обеспечения для татарского языка занимается Институт прикладной семиотики Академии наук Республики Татарстан. Сотрудники института осуществляют разработку целого ряда программных продуктов для татарского языка, таких как, система машинного перевода, синтезатор и распознаватель речи, электронный корпус, русско-татарский тезаурус, электронный диалектологический атлас и другие. Эти программные продукты создаются на основе технологий разного типа. С одной стороны, это технологии машинного обучения, которые требуют наличия больших электронных корпусов, а с другой стороны, технологии, требующие наличия

лингвистических ресурсов типа тезаурусов, баз данных фреймов и правил синтаксиса и морфологии, семантических моделей. Создание подобных электронных лингвистические ресурсов трудоемкий и времязатратный процесс, особенно если они предназначены для многоязычных систем, где необходимость определения межъязыковых эквивалентов вызывает дополнительные трудности и увеличивает время разработки.

Несмотря на то, что активные разработки для целого ряда тюркских языков ведутся уже с 1990-х годов, долгое время между разработчиками отсутствовала реальная интеграция исследований, происходило дублирование лингвистических моделей и ресурсов, а также программных модулей их обработки. Причем, эти происходило в то время, когда для европейских языков создавались интегрированные и унифицированные многоязычные лингвистические базы данных, такие как, Euro WordNet, MultiWordNet, BalkaNet, BabelNet.

Среди языков, входящих в эти многоязычные лингвистические ресурсы, можно выделить отдельные группы со структурно и лексически близкими языками, так называемыми близкородственными языками, для которых возможно использование общих технологий и программных средств. Создание многоязычных ресурсов и библиотек программных продуктов для работы с близкородственными языками позволяет преодолевать дублирование и объединять усилия в совместных разработках, а также экономить ресурсы, как финансовые, так и кадровые. Важным требованием к разработкам по созданию многоязычных лингвистических ресурсов и библиотек сервисов для работы с этими ресурсами сегодня, тем более в перспективе, является размещение их в виде общедоступных открытых интернет-ресурсов и сервисов.

Одним из таких многоязычных лингвистических интернет-сервисов для использования в открытом доступе является Многофункциональный многоязычный лингвистический интернет-сервис на базе модели тюркской морфемы.

Важным свойством архитектуры многофункционального интернет-сервиса является ее открытость, и многоступенчатая система доступа к ресурсам. Одни пользователи могут только просматривать информацию, другие – заполнять и редактировать информацию для выбранного языка, третьи – вносить изменение в общие параметры для всех языков, существующих в базе данных. Уровень доступа каждого пользователя определяется администратором системы.

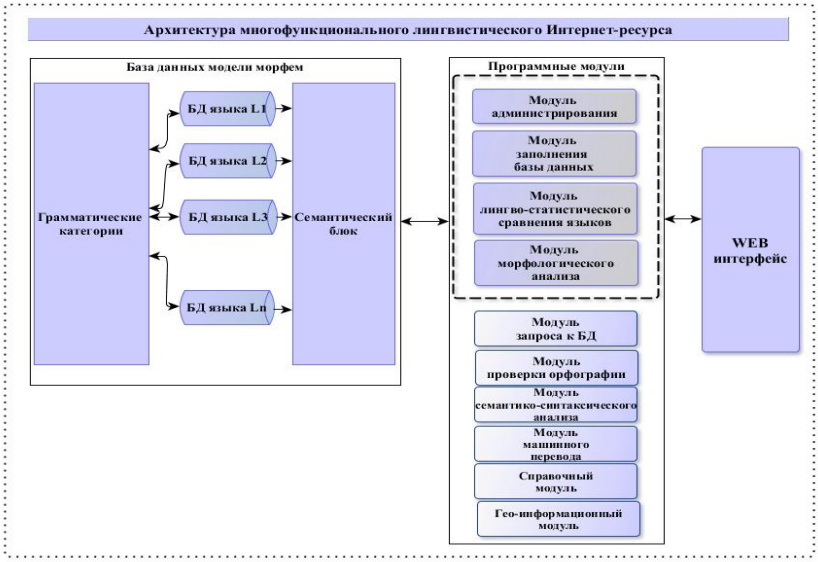

Рис. 1. Архитектура многофункционального лингвистического интернет-сервиса

Данный многофункциональный интернет-сервис на базе тюркской морфемы представляет собой веб-сайт для работы с этим интернет-сервисом, и на этом сайте представлен каталог с описанием программных модулей для компьютерной обработки тюркских языков (Рис. 1.).

В основе сервиса лежит структурно-параметрическая компьютерная модель тюркской морфемы. Эта модель представляет собой прагматически-ориентированное структурно-функциональное описание элементов морфологии. Структура модели позволяет осуществить полную «инвентаризацию» тюркских морфем с описанием характеристик и ситуаций их проявления на всех языковых уровнях (фонологическом, морфологическом, морфонологическом, синтаксическом).

Архитектура структурно-параметрической функциональной модели тюркской морфемы представляет собой иерархическую модель, состоящую из комплекса структурно-параметрических функциональных подмоделей, количество которых зависит от количества языков и диалектов, описанных в модели, а также

концептуально-формальных подмоделей с описанием технологий обработки текстов на тюркских языках (Рис. 2).

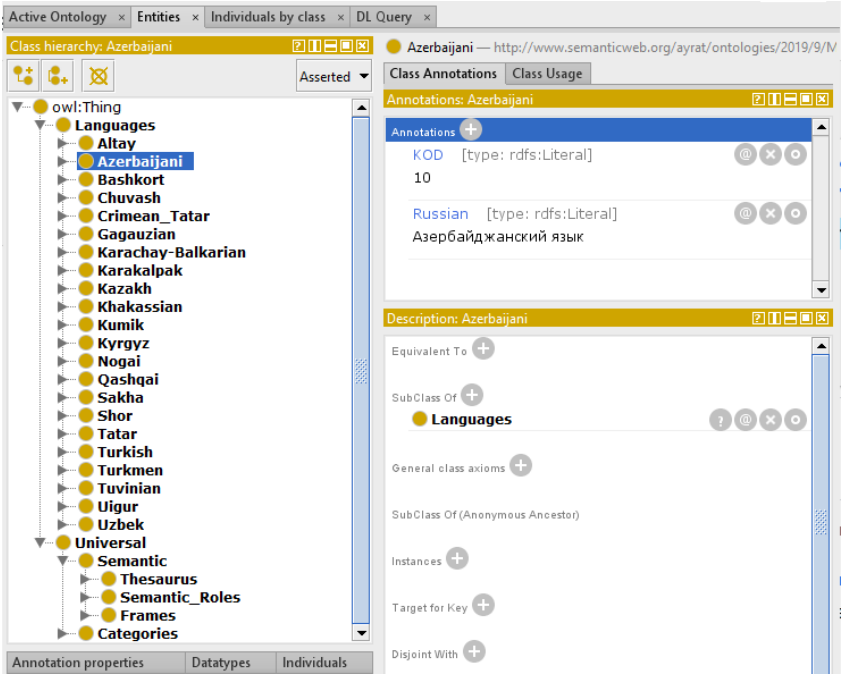

Рис. 2. Структура верхнего уровня модели

Модель состоит из двух основных блоков (Рис. 2):

1. Languages – описание языковых единиц каждого из тюркских языков;

2. Universal – языконезависимая часть модели с описаниями грамматических категорий и семантический блок.

Связующим элементом моделей описания языковых единиц отдельных языков являются семантические модели (Semantic). В нашей модели для представления семантики используются онтологические модели двух типов. Первый тип – это тезаурусы (Thezaurus), а второй тип – реляционно-ситуационные фреймы (Frames). Тезаурусы используются для описания значений корневых морфем, выражающих именные сущности. Структура части тезауруса, которая используется для представления именных концептов, аналогична структуре тезауруса WordNet.

Реализация семантического аспекта структурно-параметрической функциональной модели тюркских морфем в виде базы данных для описания значений **атрибутивных (прилагательные, наречия)** корневых тюркских морфем имеет ряд особенностей. Это связано со структурно-семантическими особенностями атрибутивных частей речи в тюркских языках. Для реализации моделей и лингвистической базы данных был произведен анализ свойств атрибутивных частей речи. Как показал этот анализ публикаций, изучению и описанию атрибутивных частей речи (прилагательные и наречия) уделяется намного меньше внимания по сравнению с такими лексическими категориями, как существительное и глагол.

В отличие от русского языка, в словарном представлении у корневых морфем в тюркских языках категориальный статус прилагательного или наречия может отсутствовать и возникать лишь в применении его в некотором контексте. В итоге категориальный статус атрибутивных языковых единиц в тюркских языках может зависеть от того, какую функцию в предложении они выполняют. Для категоризации этих единиц в современных грамматиках, как правило, учитываются все три набора свойств: морфологический, синтаксический и семантический. Это совпадает с тем, что в структурно-параметрической модели тюркской морфемы лингвистические элементы (морфемы, словоформы, словосочетания, предложения) также описываются на разных языковых уровнях. Произведя определенную структуризацию и формализацию информации, представленную в грамматиках тюркских языков, получили возможность отобразить ее в базе данных модели.

Рассмотрим эти свойства языковых единиц в тюркских языках.

Прилагательные

Согласно работам по грамматике тюркских языков прилагательные:

– обозначают признаки предметов,

– не имеют формальных показателей,

– всегда предшествуют определяемому слову,

– не согласуются с существительным: *зәңгәр чәчәк* – голубой цветок, *зәңгәр чәчәкләргә* – голубым цветам.

Если использовать для представления этих свойств типологические классификации, то типологи выделяют 2 вида употребления лексических единиц:

– немаркированные (без специального оформления),

– маркированные (с оформлением) употребления.

В тюркских языках возможны, как немаркированные, так и маркированные варианты прилагательного. К маркированным прилагательным относятся прилагательные, получаемые с помощью аффиксальных морфем: -лЫ, -сЫз, -ГЫ: язгы (тат.), көктемгі (каз.) ‘весенний’, тозлы (тат.), тұзды (каз.) ‘соленый’.

К немаркированным прилагательным относятся прилагательные, выраженные только корневой морфемой.

Например:

матур (тат.), әдемі (каз.), güzel (тур.) ‘красивый’,

кызыл (тат.), қызыл (каз.), қырмызы (каз.), kırmızı (тур.) ‘красный’.

Для выделения таких корневых морфем в модели тюркских морфем (Рис.3) ввели 2 класса лексем:

Simple Lexemes – немаркированные,

Compound Lexemes – маркированные.

Simple Lexemes – лексемы, состоящие из одной корневой морфемы. Это в модели прописано в свойствах лексем, как: equivalent some Tatar_Lexical_Roots (Рис.3).

При этом маркированные лексемы получаются разным способом: аффиксальным или образованием парных лексем. Аффиксальным способом они образуются от корневых морфем разного типа. Эти способы представлены в модели (рис.3.):

– имени существительного N+Aff,

– из имени прилагательного A+Aff и т.п.

Например, образуемые из корневой морфемы типы существительного (???) имеют следующую структуру:

1. N+ГЫ: язгы ‘весенний’, бүгенге ‘сегодняшний’
2. N+лЫ: татлы «сладкий»’
3. N+сЫз: татсыз «не сладкий»’
4. N+СЫл: аксыл ‘беловатый’
5. N+чЫл: вакчыл ‘мелочный’

Функции, которые выполняют прилагательные и наречия в предложении – это функции модификации. Прилагательные модифицируют имена существительные, а наречия модифицируют глаголы или прилагательные.

При этом в роли модификаторов выступает и немаркированное употребление существительных *таш ‘камень’* и *агач ‘дерево’* в функции модификации.

Например:

а. таш йорт ‘каменный дом’

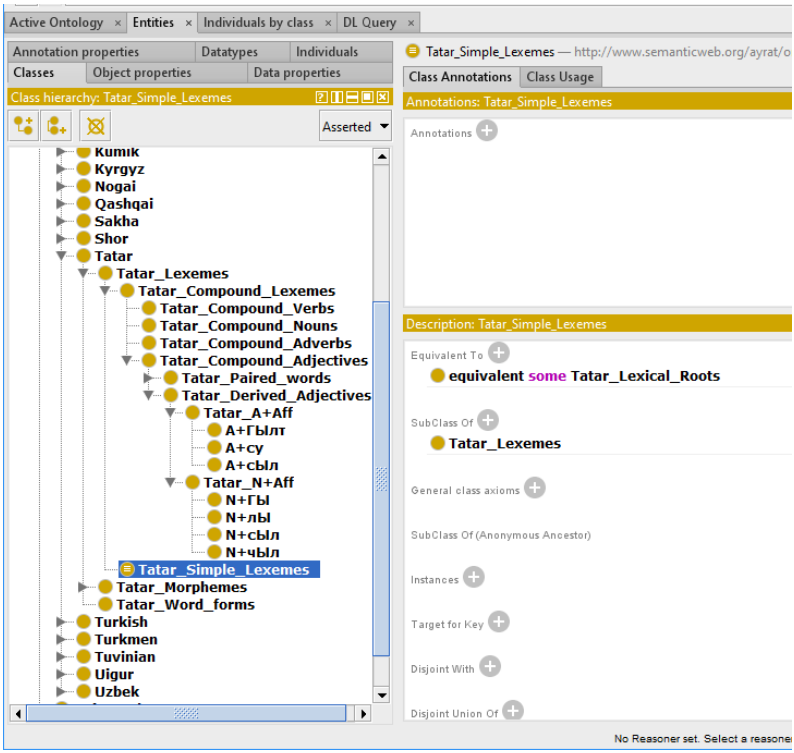

Рис. 3. Типы атрибутивных элементов в модели тюркской морфемы

b. агач күпер ‘деревянный мост’

Это свойство неоднозначности классификации в модели выражается тем, что эти языковые элементы принадлежат сразу двум классам.

Например:

татарское существительное *таш ‘камень’*

таш йорт ‘каменный дом’

таш ташладым ‘бросил камень’.

Для представления семантики в модели тюркской морфемы используются семантические универсалии разных типов. Для имен существительных – это тезаурус. В результате использования тезауруса синонимичные языковые единицы разных тюркских языков ссылаются на один и тот же концепт в этом тезаурусе. А для представления семантики глаголов дополнительно к тезау-

русу используется и база ситуационных фреймов с обозначением семантических ролей.

Для представления структуры семантических универсалий и определения структуры тезауруса для описания атрибутивных элементов были использованы классификации, предложенные филологами. Так для имен прилагательных Диксоном было предложено 7 классов прилагательных с близкими (внутри класса) семантическими и грамматическими свойствами [Dixon, 1977]:

– размер (большой, маленький, длинный, короткий, широкий, узкий, ...),

– физические свойства (жесткий, мягкий, тяжелый, легкий, ...),

– цвет (черный, белый, красный, ...),

– состояние человека (ревнивый, счастливый, добрый, умный, веселый, жестокий, ...)

– время (новый, молодой, старый, ...)

– оценка (хороший, плохой, правильный, совершенный, ...)

– скорость (быстрый, медленный, ...)

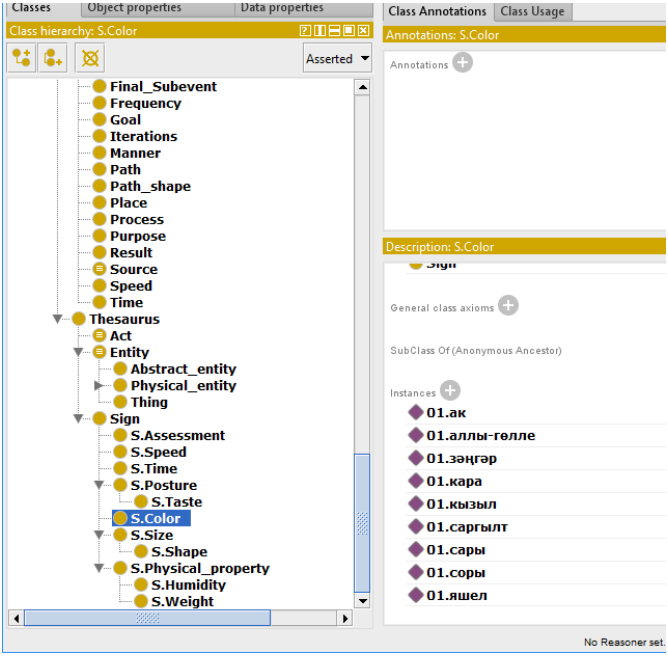

Рис. 4. Классы тезауруса для описания свойств атрибутивных элементов

Эти классы были добавлены в модель морфем для определения веток верхнего уровня тезауруса. Также для дальнейшего уточнения структуры тезауруса планируется провести сравнительный анализ концептов таких тезаурусов как РуТез и WordNet. Однако в РуТезе классификация сделана не отдельно для концептов каждого типа, а в одном концепте объединены текстовые входы, представленные разными частями речи.

Морфологические и синтаксические свойства морфем.

Морфологические свойства прилагательных и наречий в типологии определены грамматическими категориями, которые можно разделить на:

собственные (интерпретируемые),

согласовательные (неинтерпретируемые).

Согласовательные категории повторяют некоторое подмножество именных категорий в их неинтерпретируемом варианте.

Так, например, в русском языке полные формы прилагательного выражают согласовательные род, одушевленность, число и падеж, а краткие формы – род и число. В отличие от них в тюркских языках прилагательные не выражают никаких согласовательных характеристик.

В качестве одного из синтаксических свойств атрибутивных конструкций определен порядок их следования в тексте относительно определяемого слова. Филологами в разных работах приводится разный порядок следования модификаторов для разных языков:

$Adj_{quantification} > Adj_{quality} > Adj_{size} > Adj_{shape} > Adj_{color} > Adj_{nationality}$ [Cinque, 1994]

Ordinal > Cardinal > Subject Comment > Evidential > Size > Length > Height > Speed > Depth > Width > Temperature > Wetness > Age > Shape > Color > Nationality/Origin > Material [Scott, 2002]

$[_{quantif}$ Ordinal > Cardinal] >
$[_{speak\text{-}orient}$ Subject Comment > Evidential] >
$[_{scalar\ phys.\ prop.}$ Size > Length > Height > Speed > Depth > Width]>
$[_{measure}$ Weight > Temperature > Wetness > Age] >
$[_{non\text{-}scalar\ phys.\ prop.}$ Shape > Color > Nationality/Origin > Material]
[Laenzlinger, 2005]

В этих работах утверждается, что жесткий порядок следования прилагательных относительно определяемого слова в разноструктурных языках связан с тем, что порядок следования функцио-

нальных проекций в именной группе универсален. Исследуя взаимный порядок следования фразовых модификаторов в именной группе и порядок следования аффиксов в именной словоформе, можно восстановить универсальную структуру проекций, окружающих лексическую вершину.

Для описания таких свойств атрибутивных корневых морфем в модели тюркской морфемы ввели параметр:

Be_to_the_right_in_NP ‘быть справа в именной группе’

Эта функция определяет, какой из двух атрибутов должен быть ближе к главному слову в именной группе: яхшы бала – яхшы укый. В модели морфем это свойство определяется таким образом: определяется класс корневых морфем, который одновременно является подклассом двух классов – модификаторов действия и модификаторов признака.

Наречия

Как показал анализ литературы по грамматикам тюркских языков, наречия в тюркских языках изучены без сравнения с другими родственными языками. Есть отдельные работы сравнительного характера, но сравнения в них носят эпизодический, не системный характер. Для установления общих и специфических особенностей образования наречий в тюркских языках нужны сравнительные исследования системного характера. Инструментом, который может помочь осуществить такой сравнительный анализ, является структурно-параметрическая модель тюркской морфемы, представленный в этой работе, поскольку свойства текстовых единиц в этой модели уже структурированы и частично формализованы.

В тюркских языках наречиями становятся, как падежные формы имен существительных и прилагательных, местоимений, числительных, так и словарные формы имен прилагательных, а также имен существительных и глаголов. Однако адвербиализация словарных форм последних двух частей речи весьма ограничена.

Группа таких наречий, как: узакъда (каз.) ‘далеко’, уңда (тат.) ‘справа’, алға (каз.) ‘вперед’, монда (тат.) ‘здесь’ и др., будучи формально связана со словоизменительными категориями различных частей речи, приобрела наречное значение благодаря постоянному употреблению в синтаксической функции обстоятельств. Как видно из примеров, эти слова внешне представляют собой

падежные формы имен существительных, реже – прилагательных. Главным отличительным признаком является то, что их значение раскрывается только в контексте, где обнаруживается, что эти слова не имеют конкретного вещественного значения, а выражают различные обстоятельственные отношения и изолируются от системы словоизменения. В силу специальных причин (обычно, семантико-синтаксических) одна из форм с падежным аффиксом принимает новое значение, лишь отдаленно связанное с корневой морфемой основного слова. Таким образом, в результате постоянного употребления в синтаксической функции обстоятельств ряд слов приобретает признаки новой грамматической категории – наречия.

Как указывается в книгах по грамматике, в современных тюркских языках наречия являются наименее устойчивыми грамматическими категориями, что выражается в следующем:

1. постоянный переход части наречий в служебные слова (послелоги, союзы, частицы): соң ‘поздно’ – соң ‘после’;

2. наречия по некоторым своим грамматическим признакам сближаются с прилагательными и в значительной степени совпадают с ними морфологически и по функциям, выполняемым в предложении:

матур кыз ‘красивая девушка’ – матур сөйли ‘красиво говорит’

3. наречия находятся в постоянной взаимосвязи с другими частями речи (прилагательными, существительными, глаголами и др.), что затрудняет установление лексико-грамматических границ между ними. У некоторых наречий имеются как лексические, так грамматические синонимы, выходящие за пределы данной категории.

Филологи отмечают сложность выделения наречий в отдельную часть речи в тюркских языках. Это связано с большой неоднородностью групп слов, которые принято относить к наречиям, отсутствием общих положительных парадигматических признаков, релевантных для этой части речи и общих для всех видов наречия, а также отсутствием единого семантического критерия и невозможностью подвести все типы наречия под одну синтаксическую категорию.

Для представления значения наречия в модели тюркской морфемы также разрабатывается тезаурус наречий. Для представления тезауруса необходима семантическая классификация этих наречий. В грамматике дается следующая классификация наречий.

Эта классификация взята в качестве базовой в процессе построения тезауруса.

Семантически наречия делятся на 2 класса:

– определительные наречия,

– обстоятельственные наречия.

Заключение

Многофункциональный лингвистический интернет-сервис на основе структурно-функциональной модели тюркской морфемы, частично описанный в этой статье, находится на стадии заполнения лингвистических баз данных для татарского, казахского, турецкого, крымскотатарского, узбекского языков. Авторы надеются, что данный сервис будет активно использоваться исследователями тюркских языков и разработчиками соответствующих лингвопроцессоров, и будет способствовать сближению тюркских языков в области создания и применения общих новых терминов и понятий, особенно в области информатики и компьютерных технологий.

ЛИТЕРАТУРА

1. Сулейманов Д. Ш., Гатиатуллин А. Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань: Фэн, 2003. – 220 с.

2. Cinque, 1994 – Cinque G. On the evidence for partial N-movement in the Romance DP. Paths towards universal grammar. G. Cinque, J. Koster, J.-Y. Pollock, L. Rizzi, R. Zanuttini (eds.). Georgetown, 1994. Pp. 85–110.

3. Dixon, 1977 – Dixon R.M.W. Where have all the adjectives gone? Studies in Language. 1977. Vol. 1. Pp. 19–80.

4. Laenzlinger C. Some Notes on DP-internal movement. Generative Grammar in Geneva. 2005. Vol. 4. Pp. 227–260.

5. Scott G.-J. Stacked adjectival modification and the structure of nominal phrases. Functional Structure in DP and IP. The Cartography of Syntactic Structures. Vol. 1. G. Cinque (ed.). Oxford, 2002. Pp. 91–120.

6. Дмитриев Н. К. Грамматика башкирского языка. -М.- Л.: Изд-во АН СССР, 1948. – 276 с.

7. Хангилдин В. Н. Татар теле грамматикасы: морфология һәм синтаксис. – Казан: Тат. кит. нәшр., 1959. – 642 б.

УДК 81’33

ON THE IMPLEMENTATION OF THE ONTOLOGICAL MODEL OF THE SYNTACTIC LEVEL OF THE TATAR LANGUAGE GRAMMAR

D. Sh. Suleymanov, A. R. Gatiatullin
Academy of Sciences of the Republic of Tatarstan,
Russian Federation, Tatarstan, Kazan
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

The paper describes the work on a project to create an ontological model of the grammar of Turkic languages. At the current stage of the project, a description of the syntactic level of the Tatar language is carried out. The project comprises tasks of preparing linguistic knowledge bases which are necessary to integrate various types of linguistic data and to create a single information space for diverse projects in the field of Turkic computational linguistics. The result of the work should be a unified, dynamically developing model of knowledge on the subject area – the grammar of Turkic languages.

Keywords: ontological model, grammar model, Tatar language, language syntactic level.

О РЕАЛИЗАЦИИ ОНТОЛОГИЧЕСКОЙ МОДЕЛИ СИНТАКСИЧЕСКОГО УРОВНЯ ГРАММАТИКИ ТАТАРСКОГО ЯЗЫКА

Д. Ш. Сулейманов, А. Р. Гатиатуллин
Академия наук Республики Татарстан, РФ, Татарстан, Казань
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

В статье описывается работа над проектом по созданию онтологической модели грамматики тюркских языков. На данном этапе проекта решается описание синтаксического уровня татарского языка. В проекте решались задачи подготовки лингвистических баз знаний, которые необходимы для решения проблемы интеграции различных лингвистических данных и создания единого информационного пространства для различных разработок в области тюркской компьютерной лингвистики. Результатом работы должна стать динамически единая развивающаяся модель знаний о предметной области – грамматике тюркских языков.

Ключевые слова: онтологическая модель; модель грамматики; татарский язык; синтаксический уровень языка.

Введение

Разработка лингвистических онтологий ведется уже в течение последних 15–20 лет и не является новой задачей [1,2]. Новизна и актуальность результатов данной работы заключается в построении онтологических моделей для тюркских языков, которые отражают специфические структурно-функциональные особенности именно тюркских языков.

Задачей, решенной на данном этапе проекта, было построение онтологической модели синтаксиса. Это означает, что необходимо было произвести классификацию и описание языковых единиц синтаксического уровня, а также отношений между этими языковыми единицами. Для решения поставленных задач использованы методы системного анализа и онтологического инжиниринга. Выделены ключевые классы онтологии грамматики, составляющие общий словарь терминов для представления знаний о предметной области. Построена и обоснована таксономия классов онтологии грамматики, представляющая иерархию терминов по отношению вложения. Установлено множество межклассовых отношений (объектных свойств), определяющих смысловую структуру рассматриваемой предметной области.

Особенностью данного проекта является то, что для обеспечения единого информационного и терминологического пространства одна из моделей была выбрана в качестве исходной, все модели для других языков разрабатывались в соотношении к ней. В роли этой модели использована онтологическая модель казахской грамматики. Это позволяет использовать единую систему обозначений концептов, классов и отношений.

Данная работа производилась с использованием грамматик описываемых языков. Главное отличие получаемого лингвистического ресурса от классических грамматик – это структуризация и формализация представления этих грамматик, что должно позволить производить компьютерную обработку, получаемой лингвистической базы данных.

1. Реализация табличного описания грамматик

1.1. Грамматические единицы

В разрабатываемой онтологической модели синтаксического уровня языка представлены грамматические сущности следующих видов: словоформы, словосочетания, предложения.

Для этих сущностей произведена подробная классификация по синтаксическим, морфологическим и семантическим признакам: по частям речи, по присоединяемым аффиксальным цепочкам и ролям, которые они могут играть в грамматических единицах более высокого уровня.

Например, фрагмент классификации:

2. Словосочетание

1.1. Устойчивое словосочетание (Фразеологизм)

1.2. Свободное словосочетание

1.2.1. Именное словосочетание

1.2.2. Глагольное словосочетание

1.2.3. Прилагательное словосочетание

В процессе заполнения онтологической модели, подтверждается гипотеза о том, что грамматические единицы, верхнего уровня для всех тюркских языков совпадают. Различия начинаются на уровне реализации (экземпляров), нижнем уровне онтологического дерева, где рассматриваются конкретные модели управления и грамматические средства, которыми эти конструкции выражаются.

Для всех единиц производится заполнение информации на национальных языках. Пример такого заполнения фрагмента модели приведен в таблице 1.

Таблица 1

1.	**Словосочетание**	Казах	Татар	Примеры
1.1.	**Устойчивое словосочетание (Фразеологизм)**	Тұрақты сөз тіркесі	Фразеологик сүзтезмә	җил куу, тырай тибү, үги ана яфрагы, эт шомырты, искә төшерү, хәтергә алу
1.2.	**Свободное словосочетание**	Еркін сөз тіркесі	Ирекле сүзтезмә	әтигә багышлау, Ватанны ярату, агачтан ясау, бер атнада тәмамлау, бер атна

Продолжение таблицы 1

1.2.1.	**Именное словосочетание**	Есімді тіркес	Исем сүзтезмә	укучының китабы, ефәк күлмәк, җиде ел, укыган кеше, муеннан кар
1.2.2.	**Глагольное словосочетание**	Етістікті тіркес	Фигыль сүзтезмә	аңа сөйләү, берәүне күрү, укыганны ярату, бишәүләп язу, матур тегү

В процессе работы над данным проектом произведена подробная классификация не только самих грамматических сущностей, так и отношений между ними. В качестве основных отношений онтологии в лингвистических онтологиях используется следующий набор надёжных отношений: выше-ниже (класс-подкласс), часть-целое, отношение онтологической зависимости, обозначаемое как несимметричная ассоциация: асц1-асц2.

Информация, представленная в данных моделях, опирается и тесно переплетается с материалами, представленными на морфологическом уровне онтологической модели (предыдущий этап проекта). Так с одной стороны, словоформы сами состоят из нескольких морфем или одной морфемы, а с другой стороны морфемы используются для выражения большой группы синтаксических отношений. Это могут быть, как аффиксальные, так и послеложные морфемы.

Например:

урманга бара ‘идет в лес’

урманга кадәр бара ‘идет до леса’

1.2. Синтаксические отношения

Грамматические отношения между членами предложения и членами словосочетания представлены в качестве экземпляров и классов концептов. В нашей модели рассматриваются и описываются на содержательном и формальном уровне такие отношения, как: Согласование, Управление, Примыкание.

В таблице 2 приведен пример представления информации, от-

ражающей структуру отношений глагольного управления, когда в словосочетании главное слово является глаголом.

Таблица 2

№	Зависимое слово	Форма вязи	Главное слово	Примеры
1	Глагольное управление			
1.1.	Глагольное управление с окончанием направительного падежа			
1.1.1.	Имя существительное	Направительный падеж	Глагол	*Урманга бара*
1.1.2.	Имя прилагательное	Направительный падеж	Глагол	*Кызылга буяды*
1.1.3.	Имя числительное	Направительный падеж	Глагол	*Бишкә тапкырлады*
1.1.4.	Местоимение	Направительный падеж	Глагол	*Аңа әйтте*
1.1.5.	Причастие	Направительный падеж	Глагол	*Булганга була*
1.2.	Глагольное управление с окончанием исходного падежа			
1.1.1.	Имя существительное	Направительный падеж	Глагол	*Юлдан бара*
1.1.2.	Имя прилагательное	Направительный падеж	Глагол	*Матурдан киенгән*

Информация в таблице имеет древовидную структуру, которая определяется номерами в левом столбце.

Например:

1. – 1-й элемент, а 1.1. – потомок элемента с номером 1.

Отношения согласования

В таблицах по описанию правил **согласования** произведена классификация типов правил согласования подлежащего и сказуемого по предикативности (персональности), лицу, числу (Таблица 3.)

Таблица 3

Зависимое слово	**Главное слово**		**Форма**	**Примеры**
I	*сыйфат фигыль*	*зат кушымчалары*	*-мын, -мен, -м*	*мин баручымын*
II (гади)	*сыйфат фигыль*	*зат кушымчалары*	*-сың, -сең*	*син баручысың*
II (ихтирамлы)	*сыйфат фигыль*	*зат кушымчалары*	*-сыз, -сез*	*сез баручысыз*
III	*сыйфат фигыль*	*зат кушымчалары*	*0*	*алар баручы*
I	*сыйфат фигыль*	*зат кушымчалары*	*-быз, -без*	*без баручыбыз*
II (гади)	*сыйфат фигыль*	*зат кушымчалары (гади)*	*-сыз,-сез*	*сез баручысыз*
II (ихтирамлы)	*сыйфат фигыль*	*зат кушымчалары (ихтирамлы)*	*-сыз,-сез*	*сез баручысыз*
III	*сыйфат фигыль*	*зат кушымчалары*	*0*	*алар баручы*

В таблице представлена информация, где отражается один из видов отличия татарской модели от казахского. Этим отличием является отсутствие в татарском языке грамматических средств для выражения вежливой формы предикативности 2-го лица.

Например:

Казахский	**Татарский**
келген-син	килгән-сең ‘ты пришел’
келген-сез	килгән-сез ‘Вы пришли’
келген-синдер	килгән-сез ‘вы пришли’
келген-сиздер	килгән-сез ‘Вы пришли’

2. Заполнение базы данных

Информация с описанием каждого описываемого грамматического элемента в базе представляется на национальных языках,

в нашей части проекта соответственно на татарском языке. Это такая информация, как название элемента, текстовое описание и примеры. Наличие текстового описания и большого количества примеров позволяет использовать нашу базу данных для создания многоязычной обучающей системы по грамматикам тюркских языков.

Сравнительный анализ татарского и казахского языков в процессе построения онтологической модели грамматики показывает, что на верхнем уровне онтологические модели совпадают. Разница начинается на уровне детализации.

Так, например, в татарском и казахском языках различаются количество форм причастия. В татарском языке отсутствует аффиксальная форма -АтЫн, она в татарском языке представляется несколькими словоформами:

баратын кісі (каз.) = бара торган кеше (тат.) = баручы кеше (тат.) = идущий человек (рус.)

И наоборот, в татарском языке есть аффикс причастия будущего времени -АчАк, который отсутствует в казахском языке.

Соответственно, формы синтаксических связей, которые выражаются с помощью этих аффиксов, в другом тюркском языке выражаются другими грамматическими средствами.

Реализация на языке онтологии

Программная часть онтологической модели грамматики тюркских языков технически реализована на языке OWL в программной системе Protege. Protege является программным продуктом, разработанным в Стэндфордском университете (США, Калифорния) и распространяемым по лицензии Mozilla Public License (MPL). Система является свободным программным обеспечением с открытым исходным кодом. Система Protege позволяет упростить процесс разработки онтологии, она свободно поставляется, постоянно обновляется и может интегрироваться в другие проекты.

Основным элементом, реализованным в системе Protégé являются классы, которые представляют описание единиц синтаксического уровня татарского языка. Также описаны экземпляры классов, свойства классов и экземпляров на формальном языке и отношения между этими классами (Рис. 1).

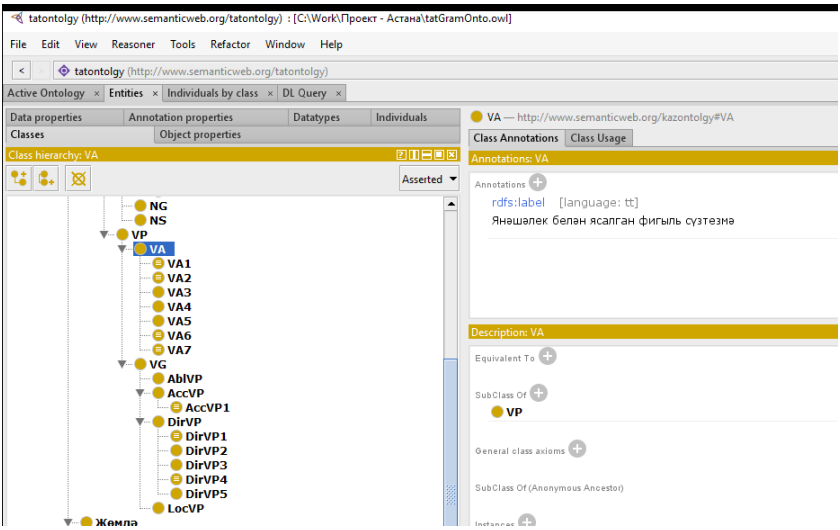

Рис. 1. Окно синтаксических классов

В системе Protégé для представления элементов каждого типа используется отдельное окно:

– classes – окно для представления классов;

– Individuals – окно для представления экземпляров класса.

– Annotation – окно для представления текстового описания элементов. В этом окне могут быть представлены, как чисто комментарии с пояснениями, так и поля с конкретными полями. Например, в данном проекте представлены такие элементы, как Question и Definitions. В поле Questions представляется список вопросов, на которые отвечает данная единица, а в поле Definitions даются определения лингвистических единиц.

На рис.1 **Классы** синтаксических элементов представлены в левом окне описания классов. Они представлены в виде иерархического дерева и имена классов представлены разными способами. Первый способ – это формальные названия грамматических категорий: NP, VP, AblVP, AccVP, LocVP, второй = это названия классов на национальном языке, в данном случае на татарском языке:

‘Җөмлә’ (‘Сөйлем’(каз.)/‘Предложение’(рус.)) и члены предложения:

Ия (‘Бастауыш’(каз.)/‘Подлежащее’(рус.),

‘Хәбәр’ (‘Баяндауыш’(каз.)/‘ ‘Сказуемое’),
‘Тәмамлык’(‘Толықтауыш’ (каз.)/‘Дополнение’(рус.)),
‘Аергыч’ (‘Анықтауыш’(каз.)/‘Определение’(рус.)),
‘Хәл’ (‘Пысықтауыш’ (каз.)/‘Обстоятельство’(рус.)).

Окно **Annotation** на рис.1 представлено в правой части окна программы и содержит текстовое пояснение о классе на татарском языке.

Например (Рис.1.):

для класса VA – ‘Янәшәлек белән ясалган фигыль сүзтезмә’.

В нижней части справа в окне **Description** представлены свойства этих классов.

При разработке онтологической базы этих классов, их названий, свойств, и отношений между этими классами необходимо было для одних и тех же категорий во всех описываемых языках, использовать одни и те же обозначения (теги). Только такой подход позволит соединить все онтологические модели в единую мультиязычную общетюркскую онтологическую модель грамматики. Это условие потребовало проведения большого объема компаративистской и типологической работы, выявляя сходства и различия в тюркских языках. При этом были использованы описательные грамматики по отдельным языкам, поиск примеров в электронных корпусах и лингвистическая интуиция участников проекта. Для поиска примеров татарского языка был использован электронный корпус татарского языка «Туган тел» (tugantel.tatar).

В окне **Individuals** реализованы экземпляры классов (Рис. 2.). Экземпляры представляют собой примеры заполнения классов для конкретного языка. Поскольку онтологическая модель данного этапа является продолжением предыдущего этапа с описанием единиц морфологического уровня, то в базе данных хранится информация с описанием всех классов и экземпляров морфологического уровня (морфемы, алломорфы, грамматические категории).

На данном этапе к ним добавлены описания классов и экземпляров синтаксического уровня языка (словоформы, словосочетания, предложения). Все конструкции (классы, экземпляры) синтаксического уровня образованы с помощью единиц морфологического уровня языка, поэтому в описании свойств этих единиц (структура, функции) использованы единицы морфологического уровня. Это образует целостность и непротиворечивость единой онтологической модели, состоящей из классов и экземпляров разного уровня.

На рис. 2 представлен пример с экземпляром предложения на татарском языке:

Балалар мәктәпкә барганнар. ‘Дети ходили в школу.’

В окне Descriptions (Рис. 2) показано, что экземпляр ‘*Балалар мәктәпкә барганнар*’ является экземпляром класса *Җөмлә ‘Предложение’,* а в окне Property assertions представлено, что этот экземпляр содержит две группы:

– Именная группа hasNP – *балалар ‘дети’*

– Глагольная группа hasVP – *мәктәпкә барганнар ‘ходили в школу’.*

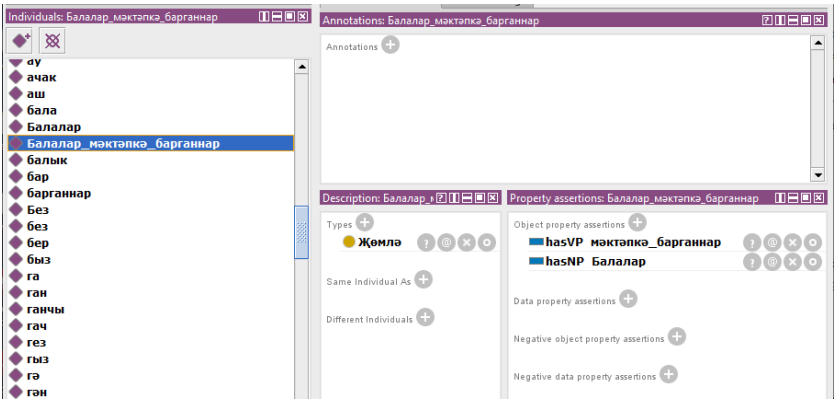

Рис. 2. Пример экземпляра класса ‘Предложение’

Элементы представленные в окнах системы Protégé являются гиперссылками и нажимая на них можно открывать страницы с описанием этих элементов. В результате пройдя по гипертекстовым ссылкам и открыв описание именной и глагольных групп можно увидеть, что экземпляр *мәктәпкә барганнар ‘ходили в школу’* принадлежит к классу P: сүзтезмә ‘словосочетание’.

Структура экземпляра ‘мәктәпкә барганнар’ имеет следующий вид:

– главная часть hasHead – *барганнар ‘ходили’;*

– зависимая часть hasDependent – *мәктәпкә ‘в школу’.*

Такая полная структурированность и взаимосвязанность описываемых моделей предоставляет возможность с одной стороны стратифицированно (по языковым уровням), а с другой целостно исследовать сходство-различие тюркских языков.

Дальнейшим развитием данной модели может стать присваивание языковым единицам, классам или свойствам, определенных весовых значений. Данные весовые значения могут, например представлять степень важности данного свойства по отношению к другим параметрам. Эти весовые параметры могут быть получены на основе статистического анализа на корпусных данных. Наличие таких параметров позволит использовать статистический аппарат для компаративных исследований по оценке близости-дальности тюркских языков и других типологических исследований. Это показывает необходимость продолжения проекта по онтологическому описанию для других языковых уровней, а также других языков тюркской группы.

На рис. 3 приведен пример формального описания свойств класса, где отношение примыкания типа VA1 (примыкание глагола с наречием) должно обладать следующими свойствами:

$$VA1 \equiv \exists hasDependent(Adv) \cap \exists hasHead(V)$$

Здесь:

- VA1 – глагольное примыкание (наречие + глагол);
- hasDependent – имеется зависимое слово;
- Adv – наречие;
- hasHead – имеется главное слово;
- V – глагол.

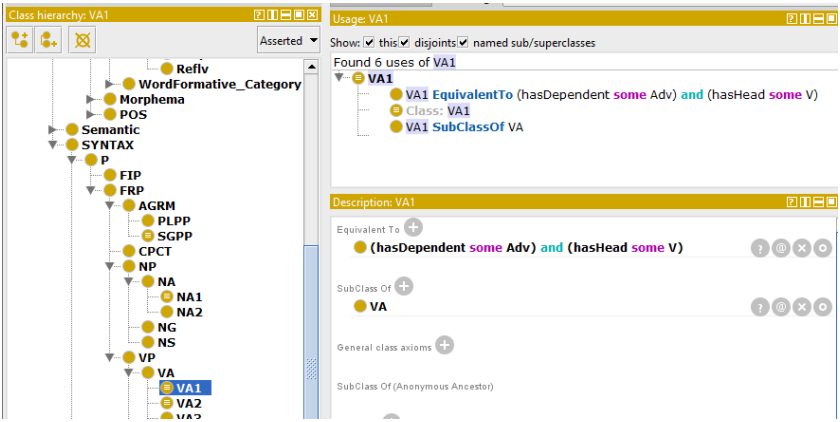

Рис. 3. Пример формального описания свойств классов

Различия казахской и татарской моделей онтологии

Одна из задач, которая решалась в процессе заполнения онтологической модели грамматики – это сравнение моделей разных тюркских языков. Одним из основных отличий, которое было выявлено в процессе заполнения модели является разница в моделях глагольного управления. Оно означает, что словоформы, играющие одну и ту же семантическую роль в разных тюркских языках, описываемых в проекте, имеют разные грамматические формы.

Рассмотрим в качестве примера предложение на казахском языке: *Балалар тамакка тойды. «Дети насытились».* На татарском языке такое предложение некорректно. Правильным вариантом с тем же значением на татарском языке будет *Балаларның тамаклары туйды. «Дети насытились».*

Этот пример демонстрирует, что в татарском и казахском языке различаются модели глагольного управления. Та синтаксическая классификация, которая реализована в данной онтологической модели не позволяет полноценно отражать подобную разницу в языках. Для получения такой возможности, кроме синтаксической и морфологической модели в общую онтологическую модель необходимо добавить семантико-синтаксическую подмодель в виде модели управления глаголов, и модели ситуационных фреймов.

Заключение

Объединение онтологических моделей разных тюркских языков в единую онтологическую модель позволит получить несколько полезных результатов:

1. Общий многоязычный терминологический словарь грамматических терминов, который можно использовать для решения задач машинного перевода между тюркскими языками, многоязычного поиска лингвистических терминов, создания многоязычных курсов автоматизированного обучения по тюркской лингвистике;

2. Полученная лингвистическая база данных позволит производить сравнительный статистический анализ грамматик тюркских языков, и определять какие из тюркских языков наиболее структурно близки друг к другу.

Благодарности

Статья подготовлена в рамках проекта APS05132249 «Разработка электронных версий тюркских языков для создания многоязычных поисковых и основанных на знаниях систем» по контракту № 132 от 12 марта 2018 года.

ЛИТЕРАТУРА

1. Боровикова О. И., Загорулько Ю. А., Загорулько Г. Б., Кононенко И. С., Соколова Е. Г. Разработка портала знаний по компьютерной лингвистике // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием КИИ-2008. М.: ЛЕНАНД, 2008. Т. 3. С. 380–388.

2. Портал знаний по компьютерной лингвистике. http://uniserv.iis.nsk.su/cl/.

УДК 81.322

MORPHOLOGICAL GLOSSING IN THE INTERPRETATION OF TRANSLATION TRANSFORMATIONS

G. G. Torotoyev, S. G. Torotoyeva
Ammosov North-Eastern Federal University, RF, Sakha, Yakutsk
torgav@mail.ru, tsandaara@mail.ru

In the interpretation of translation transformations, the authors propose a tagging system based on the Leipzig glossing rules. Morpheme-by-morpheme glosses can be effectively used by linguists in fundamental and applied linguistic research, as well as for educational and methodological purposes, in particular, in training courses on translation theory and workshops on Russian-Yakut translation.

Keywords: morpheme-by-morpheme glosses, modeling method, system of identifiers (tags), translation transformations.

МОРФОЛОГИЧЕСКОЕ ГЛОССИРОВАНИЕ В ИНТЕРПРЕТАЦИИ ПЕРЕВОДЧЕСКИХ ТРАНСФОРМАЦИЙ

Г. Г. Торотоев, С. Г. Торотоева
Северо-Восточный федеральный университет
им. М. К. Аммосова, РФ, Саха, г. Якутск
torgav@mail.ru, tsandaara@mail.ru

В интерпретации переводческих трансформаций авторами предлагаются система тэгов, базирующаяся на Лейпцигских правилах глоссирования. Поморфемная нотация может быть эффективно использована языковедами в фундаментальных и прикладных лингвистических исследованиях, а также в учебно-методических целях, в частности, в учебных курсах по теории перевода и практикумах по русско-якутскому переводу.

Ключевые слова: поморфемное глоссирование, метод моделирования, система идентификаторов (тэгов), переводческие трансформации.

В последнее время в связи с интенсивным развитием компьютерных технологий назрела необходимость в разработке системы грамматической разметки для автоматического анализа текстов, хранящихся в электронных корпусах тюркских языков. В целях повышения эффективности работы языковедов при проведении сравнительно-сопоставительных исследований и для получения

объективных языковых данных необходима унификация системы идентификаторов (тэгов). Для того, чтобы компьютер мог автоматически проанализировать тексты любой сложности, представленные в электронном корпусе якутского языка, необходимо описать унифицированными тэгами все грамматические категории языка саха. При решении данной проблемы станет возможным создание новых компьютерных программ, таких как онлайн-переводчик, автоматический анализатор текстов, синтезатор речи и др.

Творческим коллективом (Торотоев Г. Г., Леонтьев Н. А., Новицына А. Н., Бочкарев В. В., Торотоева С. Г.) проделана большая работа: создана База данных по словоизменительным аффиксам языка саха [3], [4], разработана компьютерная программа «Морфологический анализатор якутского языка», на стадии разработки находятся программы «Морфонологический анализатор языка саха», «Саха-тыва-казахский переводчик».

Условные обозначения частей речи в языке саха

Английские термины	Русские термины	Якутские термины	Унифицированные тэги	Тэги собственной разработки
Nouns	Имена существительные	Аат тыл	N	A
Possessive Case of Nouns*	Имена притяжательные	Тардыылаах аат тыл	POSS	A//
Pronouns	Местоимения	Солбуйар аат	PRO	CA
Numerals	Имена числительные	Ахсаан аат	NUM	AA
Adjectives	Имена прилагательные	Даҕааһын аат	ADJ	Д
Verbs	Глагол	Туохтуур	V	T
Participles*	Причастия	Аат туохтуур	PCP	AT
-	Деепричастия	Сыһыат туохтуур	CONV	CT
Adverbs	Наречия	Сыһыат	ADV	C
Modal words*	Модальные слова*	Сыһыан тыл	MOD	СыһТ
Interjections	Междометия	Саҥа аллайыы	INTJ	САл
Conjunctions	Союзы	Ситим тыл	CONJ	Сит
Particles*	Частицы	Эбиискэ	PART	Э

Продолжение таблицы

Prepositions	Предлоги	-	PREP	-
Postpositions*	-	Дьөһүөл	POST	Дь
Articles	-	-	ART	-

В данной таблице представлены лексико-грамматические разряды английского, русского и якутского языков в сопоставительном аспекте. Из контента таблицы следует, что унифицированные тэги кодифицированы на базе английских терминов. Звездочкой помечены те категории, грамматические значения которых в вышеперечисленных языках, не полностью совпадают. Система условных сокращений, основанная на якутских терминах, была разработана нами для внутреннего пользования, иными словами, студент использует данные тэги тогда, когда он производит лингвистический анализ на языке саха. Таблицы подобного рода способствуют студентам самостоятельно находить языковые универсалии и параллели в разносистемных языках. [2].

Многие компаративисты сходятся во мнении, что «поморфемная нотация лучше всего подходит для описания агглютинативных языков, соответствующих элементно-комбинаторной модели морфологии». [1]. Как известно, язык саха относится к языкам агглютинативного типа. И это означает, что у языка саха в области компьютерной лингвистики большие перспективы, следовательно, и в машинном переводе. В целях репрезентативности основные морфологические категории и признаки имени существительного и глагола языка саха представим в виде таблиц.

1. Морфологическая структура слова в языке саха

Ø	1	2
Корневая морфема (именная/глагольная)	Словообразовательная морфема	Словоизменительная морфема

2. Иерархическая последовательность морфем в имени существительном языка саха

Ø	1	2	3
Именная основа	Множественность	Притяжательность	Падежность

Для наглядности возьмем синонимы *убай* и *бии* (*старший брат*) и продемонстрируем их последовательную морфологическую трансформацию до формы притяжательного аккузатива > *убайдарбын/биилэрбин* (*моих старших братьев*):

Ø	1	2	3
Именная основа	Множественность	Притяжательность	Падежность
Убай	*-лар*	*-(ы)м*	*-ын*
Бии (архаизм)	*-лэр*	*-(и)м*	*-ин*

3. Основные категории глагола в языке саха

Глагольная основа **Ø**	Категория отрицания	Категория залога	Модальность	Аспектуальность	Категория времени	Персональность	Множественность	Диминутив

Все вышеперечисленные показатели основных категорий глагола имеют строго закрепленные места в глагольной словоформе, что являются большим преимуществом в компьютерном моделировании словообразовательных процессов.

Прежде чем мы приступим к интерпретации переводческих трансформаций с помощью морфологических идентификаторов проясним следующее. Как правило, текст с поморфемной нотацией состоит из трех параллельных строк. В первой строке располагается оригинал текста, где морфологические показатели формально отделены друг от друга. Во второй строке дается поморфемный перевод текста на язык-посредник с использованием соответствующих тэгов, в третьей строке – литературный перевод на тот же язык.

Рассмотрим глоссирование параллелистической конструкции олонхо К. Г. Оросина «Дьулуруйар Ньургун Боотур»:

(1) Орто дойду дьол-у-н тох-тор-д-(у)лар,
средняя страна счастье-POSS.3SG-ACC пролить-CAUS-PAST-3PL
‘Пролили счастье срединного мира’

(2) Төр-үүр оҕо уйа-ты-н түҥ-нэр-д-(и)лэр,
родиться-PCP.PRES ребенок/дитя гнездо-POSS.3SG-ACC опрокинуть-CAUS-PAST-3PL
‘Опрокинули колыбели рождающихся детей’

(3) Иит-эр	сүөһү	күрүө-тү-н	ин-нэр-д-(и)лэр...
разводить-PCP.PRES	скот	изгородь-POSS.3SG-ACC	повалить-CAUS-PAST-3PL

‘Разрушили изгороди разводимого скота’

Далее, опираясь на метод моделирования, можно обнаружить те или иные формальные соответствия или расхождения в параллелистических конструкциях оригинала и подстрочника.

Орто дойду дьолун тохтордулар,	ADJ+N+POSS+V
Төрүүр оҕо уйатын түҥнэрдилэр,	ADJ+N+POSS+V
Иитэр сүөһү күрүөтүн иҥнэрдилэр…	ADJ+N+POSS+V

Пролили счастье срединного мира,	V+N+ADJ+N
Опрокинули колыбели рождающихся детей,	V+N+ADJ+N
Разрушили изгороди разводимого скота...	V+N+ADJ+N

В ходе анализа нами выявлены следующие типы переводческих трансформаций:

– Лексические преобразования (замена лексических единиц: страна – мир, повалить – разрушить; паронимы: средний – срединный; предикаты могут быть выражены потенциальными синонимами);

– Морфологические преобразования (единственное число – множественное число: ребенок/дитя – дети; адъективация и др.);

– Синтаксические преобразования (перестановка предикатов; изменение порядка следования синтаксических элементов в тексте перевода по сравнению с текстом подлинника).

Вышеперечисленные типы переводческих трансформаций обусловлены тем, что русский и якутский языки являются разноструктурными (неродственными). Использование метода моделирования в познании процесса транскодирования естественного языка позволяет теоретически осмыслить и адекватно интерпретировать различные типы переводческих трансформаций.

В заключение хочется отметить, что подстрочник выдерживает основные стихообразующие каноны якутского героического эпоса, адекватно передаёт содержательно-концептуальную информацию, формульность и образность олонхо, в целом, его художественное своеобразие.

ЛИТЕРАТУРА

1. Плунгян В. А. Общая морфология: Введение в проблематику: Учебное пособие. – М.: Эдиториал УРСС, 2000. – С. 331.

2. Торотоев Г. Г. Метод моделирования в исследовании стихообразующего каркаса олонхо // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Изд-во «Фэн» Академии наук РТ, 2014. – С. 243–247.

3. Торотоев Г. Г., Ноговицына А. Н. Проблема аннотирования грамматических категорий в языке саха (на примере наклонений якутского глагола) // Пятая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2017». Труды конференции. В 2-х томах. Т.1. – Казань: Издательство Академии наук Республики Татарстан, 2017. – С. 336–355.

4. Торотоев Г. Г., Торотоева С. Г. Морфологическое аннотирование словоизменительных категорий имени существительного в языке саха // Сравнительно-сопоставительное изучение тюркских и монгольских языков: материалы Международной научно-практической конференции. г. Якутск, 18–19 октября 2018 г. / [под ред. А. К. Прокопьевой, А. Е. Шамаевой]. – Якутск: Издательский дом СВФУ, 2018. – С. 253–263.

Расшифровка использованных тэгов

3SG – единственное число, 3 лицо
3PL – множественное число, 3 лицо
ACC – винительный падеж
PRES – настоящее время
PAST – недавнопрошедшее время
CAUS – побудительный залог

УДК 81.322.4

MORPHOLOGICAL SEGMENTATION FOR THE KAZAKH LANGUAGE IN NEURAL MACHINE TRANSLATION

U. Tukeyev, A.Karibayeva, B. Abduali
Al-Farabi Kazakh National University,
Almaty, Kazakhstan
ualsher.tukeyev@gmail.com, a.s.karibayeva@gmail.com

The paper describes solving to the problem of morphological segmentation in neural machine translation for the Kazakh language. This problem of morphological segmentation in neural machine translation arises in connection with the increase the volume of the dictionary of the machine translation system while increasing the amount of source data for the training. This problem is especially relevant for neural machine translation of agglutinative languages. The solution to this problem is carried out through the use of morphological segmentation of the words of the source training data. In this area, the most famous and actively used method is BPE (byte pair encoding). However, this method is not effective enough for agglutinative languages. We investigate the use of morphological segmentation for the Kazakh language as a representative of the agglutinative group of languages. A model, algorithm and implementation of the morphological segmentation of the texts of the Kazakh language based on the complete system of endings of the Kazakh language are proposed. Experimental results on the use of the proposed approach for neural machine translation for the Kazakh-English pair of languages are given.

Keywords: morphological segmentation; Kazakh language; neural machine translation.

МОРФОЛОГИЧЕСКАЯ СЕГМЕНТАЦИЯ ДЛЯ КАЗАХСКОГО ЯЗЫКА В НЕЙРОННОМ МАШИННОМ ПЕРЕВОДЕ

У. Тукеев, А. Карибаева, Б. Абдуали
Казахский Национальный Университет им.аль-Фараби,
Алматы, Казахстан
ualsher.tukeyev@gmail.com, a.s.karibayeva@gmail.com

В работе описывается решение проблемы морфологической сегментации в нейронном машинном переводе казахского языка. Данная проблема морфологической сегментации в нейронном машинном переводе возникает в связи с ростом объема словаря системы машинного перевода при наращивании объема исходных данных для более качественного обучения. Особенно эта проблема актуальна для нейрон-

ного машинного перевода агглютинативных языков. Решение данной проблемы выполняется через применение морфологической сегментации слов исходных данных обучения. В данной области наиболее известным и активно используемым является метод BPE (bite pair encoding). Однако, данный метод недостаточно эффективен для агглютинативных языков. Мы исследуем использование морфологической сегментации для казахского языка, как представителя агглютинативной группы языков. Предлагается модель, алгоритм и реализация морфологической сегментации текстов казахского языка на основе полной системы окончаний казахского языка. Приводятся экспериментальные результаты по использованию предложенного подхода для нейронного машинного перевода для казахско-английской пары языков.

Ключевые слова: морфологическая сегментация; казахский язык; нейронный машинный перевод.

Введение

Нейронный машинный перевод в настоящее время является превалирующей технологией машинного перевода, показывающей впечатляющие результаты. Однако, качество нейронного машинного перевода существенно определяется объемом исходных данных для обучения системы нейронного машинного перевода.

В настоящее время имеются языки с достаточно объемными лингвистическими ресурсами для реализации нейронного машинного перевода, такие как английский, немецкий, французский, китайский и т.д. Однако, имеются много языков с малыми лингвистическими ресурсами для машинного обучения по технологии нейронного машинного перевода. Это создает проблемы по качеству машинного перевода для таких малоресурсных языков. Вместе с тем, для реализации нейронного машинного перевода имеются ограничения в виде ресурсов вычислительных систем. Это ограничение начинает проявляться с ростом объема исходных данных и, соответственно, с ростом объема словаря системы машинного перевода. Данная проблема решается путем уменьшения объема словаря через морфологическую сегментацию слов исходного текста для обучения и в литературе описан ряд таких методов. Особенно эта проблема актуальна для нейронного машинного перевода агглютинативных языков так, как существующие методы недостаточно эффективно решают ее.

Традиционные методы, которые направлены на решение проблемы морфологической сегментации, опираются в основном на статистические характеристики элементов языков.

В последнее время, попарное байтовое кодирование (BPE) используется де-факто стандартом для сегментации в нейронном машинном перевода. Эффективность использования BPE в нейронном машинном перевода для аналитических языков приведены во многих исследованиях, но применение в агглютинативных языках не улучшают качество перевода.

Сэнрич и другие (Sennrich, 2016) в соавторстве разработали метод, который сегментируют корпус по наиболее частым последовательностям символов. Они адаптировали алгоритм BPE к задаче сегментации для создания открытого словаря, который показал эффективные результаты по сравнению с использованием словаря большого объема. BPE разбивает слова на различные варианты основ и суффиксов, но он плохо работает для языков с богатой морфологией, как казахский язык.

Использование BPE было рассмотрено в других исследованиях с языковой группой, аналогичной казахскому языку. Атаман и другие (2017) предсказывает сегменты подслов с использованием алгоритма обучения морфологии без учителя, который основан на модели морфологии. Они использовали сегментацию BPE в двух экспериментах. В первом BPE используется на стороне исходного языка, во втором – на стороне целевого языка. В турецком переводе с BPE они получили ненадежный перевод, и сегменты были неоднозначными.

Ву и Чжао (2018) предлагают обобщенную сегментацию BPE и экспериментировали на немецко-английском и китайско-английском языковых парах. Они объединили обобщенную сегментацию BPE и расширенные меры подстрок для лучшего представления подслов в нейронном машинном переводе. Они сравнили различные типы показателей качества подстроки и пришли к выводу, что стратегия сегментации чувствительна к языковым парам. Вывод этого алгоритма не имеет смыслового значения для языков с богатой морфологией. Например, он может разбить слово где-то посередине, что вызывает семантическое смещение.

В данной работе предлагается метод морфологической сегментации казахского языка на основе полной системы окончаний, который рассматривается как альтернатива существующим методам морфологической сегментации для агглютинативных языков. В экспериментальной части данной работы представлено сравнение результатов сегментации при помощи BPE и сегментации на основе полной системы окончаний.

1. Полная система окончаний казахского языка

Система окончаний казахского языка делится на две группы: именные окончания (существительные, прилагательные, числительные) и глагольные окончания (глаголы, причастия, герундий, наклонение и залог). В казахском языке слово образуется при помощи 4 видов окончаний/аффиксов. Эти виды: С-падежные, Т-притягательные, К-множественные, J-личные. Окончания казахского языка могут представлены как всевозможные комбинации этих базовых видов аффиксов. Всевозможные комбинации базовых видов аффиксов состоят из комбинаций одного типа, комбинаций двух типов, комбинаций трех типов и комбинаций четырех типов. Общее количество комбинаций определяется по формуле: $A_n^k = n!/(n-k)!$.

Тогда количество комбинаций (размещений) будет определяться следующим образом:

$$A41 = 4! / (4-1)! = 4;\ A42 = 4! / (4-2)! = 12;\ A43 = 4! / (4-3)! = 24;$$
$$A44 = 4! / (4- 4)! = 24.$$

Всего 64 возможных размещений для именной основы. Рассмотрим, какие комбинации семантически допустимы. Окончания комбинаций однотипных (К, Т, С, J) семантически действительны. Окончания комбинаций двух типов следующие: КТ, ТС, СJ, JK, КС, ТJ, СТ, JТ, КJ, ТК, СК, JС. Семантический анализ комбинаций двух типов окончаний показывает, что допустимы только шесть комбинаций (КТ, ТС, СJ, КС, ТJ, КJ), а остальные комбинации неприемлемы. Например, ТК недопустим: множественные окончания не используются после притяжательных окончаний, СК недопустимы: множественные аффиксы не допускаются после окончаний падежа, JС неприемлемы: окончания падежа не допускаются после личных окончаний, СТ недопустимы: притяжательные окончания не могут быть используется после падежных окончаний, JТ неприемлемо: притяжательные окончания не могут использоваться после личных окончаний. JK неприемлемо: множественные окончания не могут использоваться после личных окончаний, поскольку эта комбинация покрыта множественными личными окончаниями. Для комбинаций трех и четырех типов окончаний определение приемлемых комбинаций окончаний выполняется в соответствии с правилом: если внутри данной комбинации имеется недопустимая комбинация двух типов, то эта ком-

бинация неприемлема. Тогда правильные комбинации окончаний трех типов будут равны 4 (KTC, KTJ, TCJ, KCJ), а правильные комбинации окончаний четырех типов равны 1. С учетом семантической допустимости размещений число всевозможных размещений для именной основы уменьшено до 15.

Аналогичным образом выполнено определение всевозможных размещений для окончаний с глагольной основой, что составило 55 семантически допустимых типов окончаний. В общем, общее количество типов окончаний для именных оснований плюс общее количество типов окончаний слов с глагольной основой равно 70. В соответствии с этими типами окончаний для всех основных частей речи казахского языка построены конечные наборы окончаний. Так, для частей речи с именными основаниями количество окончаний составляет 1213 (учитываются все варианты множественного числа), а количество окончаний частей речи с устными основаниями составляет: глаголы – 432, причастия – 1582, наречие – 48, настроения – 240, голоса – 80. Всего – 3565 (Tukeyev, 2015).

В нижеприведенной таблице показаны примеры изменении именного слова по окончаниям:

Таблица 1. Пример именных конструкции

Тип	Пример слова	Тип	Пример слова
K	студент+тер	T-C	студент+ім+е
K-C	студент+тер+де	K-T-C	студент+тер+іңіз+ге
T	студент+ім	K-T-J-C	студент+тер+і+міз+ден

Все правила последовательности комбинации окончаний были использованы в морфологической сегментации в нейронном машинном переводе казахского языка, которое рассмотрено в следующем разделе.

2. Морфологическая сегментация казахского языка

Алгоритм морфологической сегментации слов казахского языка включает два этапа:

1) деление основ и окончаний слов;

2) сегментирование окончаний слова на сегменты суффиксов.

Стадия деления основа и окончания слова может быть выполнена с использованием лемматизатора, также основанного на использовании полной системы окончаний казахского языка. В системе окончаний казахского языка все окончания делятся на классы по длине окончаний. Во второй стадии, отделенная часть окончаний сегментируется как показано в Таблице 2, в правой части.

В таблице 2 продемонстрирован шаблон сегментации, символ «@@» – используется от деления сегментов.

Таблица 2. Тип окончания с его сегментацией (фрагмент)

Тип окончания	Шаблон сегментации
терімізбенбіз	тер@@ і@@ міз@@ бен@@ біз
терімізбенмін	тер@@ і@@ міз@@ бен@@ мін
терімізбенсіз	тер@@ і@@ міз@@ бен@@ сіз
терімізбенсің	тер@@ і@@ міз@@ бен@@ сің
терменбіз	тер@@ мен@@ біз
терменмін	тер@@ мен@@ мін
терменсіз	тер@@ мен@@ сіз
терменсің	тер@@ мен@@ сің
терменмінбіз	тер@@ мен@@ біз
терменмін	тер@@ мен@@ мін
терменсіз	тер@@ мен@@ сіз
терменсің	тер@@ мен@@ сің
уің	у@@ ім
уің	у@@ ің
ге	ге
ға	ға
ы	ы

При морфологической сегментации казахского языка для предотвращения многозначности основы слова с видом окончания, был создан словарь слов исключений. Например, если слово «астық» не добавлено в исключение, то при сегментации «қ» отде-

лится от основы и это будет неправильно сегментирован. Словарь исключении был создан для предотвращения многозначности основы с типом окончаний. Этот словарь исключений состоит из 863 слов-основ, который еще требует расширения во избежания неправильной сегментацией. Эксперименты с 863 слов-основами выдали ошибки сегментации в диапазоне 12–15%.

3. Описание экспериментов и полученные результаты

Экспериментальная часть проводилась для казахско-английской языковой пары. Система обучалась на 143262 параллельных предложений казахско-английского корпуса, которые собраны из разных источников, таких как официальные сайты Республики Казахстана, а также с разных литературных произведении и книг. Все данные были предварительно обработаны, включая токенизацию и нормализацию. Данные были разделены на две части: обучение и тестирование. В таблице 3 представлено описание параллельного казахско-английского корпуса КазНУ коллекции. В таблице 4 представлены экспериментальные данные по обучению и тестированию без сегментации и с сегментацией предложенным методом с использованием параллельного казахско-английского корпуса КазНУ коллекции.

Таблица 3. Параллельные казахско-английские корпуса для обучения (КазНУ коллекция)

Имя корпуса	Количество предложений
OPUS	4 480
New World Bible	38 358
Lab IIS	5 925
Akorda	24 148
TED	6 120
Zakon	20 961
Other sites	43 270
Всего	143 262

Таблица 4. Результаты по обучению и тестированию без сегментации и с сегментацией методом полной системы окончаний казахского языка с использованием параллельного казахско-английского корпуса КазНУ коллекции

Название языковой пары	Без сегментации, BLEU	С сегментацией, BLEU
Kazakh-English	test – 20.1	test – 24.1
	dev – 17.7	dev – 22.1

В таблице 5 представлены сравнительные экспериментальные данные по использованию метода морфологической сегментации BPE и предложенного метода, основанного на полной системы окончаний казахского языка с использованием параллельных корпусов WMT 2019.

Таблица 5. Сравнительные экспериментальные данные по использованию метода морфологической сегментации BPE и предложенного метода

Языковая пара	BPE сегментация, BLEU	KAZNU сегментация, BLEU
Kazakh-English	test – 9.7	test – 14.2
	dev – 10.4	dev – 14.1

Из таблицы 4 видно, что предложенный метод морфологической сегментации позволяет улучшить метрику качества на 4.0 пункта , а из таблицы 5 результатов можно увидеть, что предложенный метод морфологической сегментации по сравнению с методом морфологической сегментации BPE улучшает метрику качества на 4.5 пунктов в казахско-английском машинном перевода.

Заключение

В этой статье был предложен новый метод сегментации для казахского языка на основе полной системы окончаний, который был использован для обучения казахско-английской системы нейронного машинного перевода. В дальнейшим планируется увеличить объем слов исключений словаря сегментации казахского языка, применить предложенный метод морфологической сегментации для других языков тюркской группы.

ЛИТЕРАТУРА

Sennrich, Rico, Barry Haddow and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, 1715–1725.

Ataman, Duygu, Matteo Negri, Marco Turchi and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. The Prague Bulletin of Mathematical Linguistics 108, no. 1, 331–342.

Wu, Yingting and Hai Zhao. 2018. Finding Better Subword Segmentation for Neural Machine Translation. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, 53–64.

Ualsher Tukeyev. AUTOMATON MODELS OF THE MORPHOLOGY ANALYSIS AND THE COMPLETENESS OF THE ENDINGS OF THE KAZAKH LANGUAGE. Proceedings of Turklang-2015. Tatarstan, Kazan. – 91–100 pp.

УДК 81.373

LEXICAL STATISTICS OF UZBEK FOLKLORE TEXT

D. B. Urinbaeva
Samarkand State University, Samarkand, Uzbekistan
dilbarxon@inbox.ru

The article deals with the lexical and statistical structure of Uzbek folklore texts on the basis of quantitative comparative research: the average frequency of word forms, text cover ability with different parts of the frequency lexicon, and the ratio of rare (random) lexical units.

Keywords: the average frequency of words; fill factor; epos; folklore; frequency; quantitative; rate synthetism; statistics.

ЛЕКСИЧЕСКАЯ СТАТИСТИКА УЗБЕКСКОГО ФОЛЬКЛОРНОГО ТЕКСТА

Д. Б. Уринбаева
Самаркандский государственный университет, Узбекистан, Самарканд
dilbarxon@inbox.ru

В статье рассматривается лексическая структура текстов узбекского фольклора на основе количественного сравнительного анализа: средняя частота словоформ, частота лексикона в разных частях текста и соотношение редких (случайных) лексических единиц.

Ключевые слова: средняя частотность слов; коэффициент заполнения; эпос; фольклор; частота; количественный, уровень синтетизма; статистика.

Наша основная задача, оценить степень квантитативно-лексической близости дастана, сказки, песни, пословицы, загадки и определить место рассматриваемых жанров в узбекском литературном языке. Решая эту задачу, мы подвергнем квантитативному анализу первые 300 словоформ, потом общий объём текстов. Сопоставление рангов и стоящих за ними частот, служит средством обнаружения сходств и различий в речевом функционировании лексических единиц в пяти рассматриваемых жанрах. Для этого мы применим метод попарного сравнения рангов «дастан» и «сказки», «дастан» и «пословицы», «дастан» и «загадки», «дастан» и «народные песни», «сказки» и «пословицы», «песни» и «пословицы», «пословицы» и «загадки», «загадки» «песни». Результаты этого сравнения оцениваются с помощью коэффицента

Спирмена [1, 11; 2, 99; 5,104]. Метод ранговой корреляции применяется нами для выявления содержательной близости между жанрами фольклорного языка.

Наблюдение над рангами лексических единиц и их употребительностью в текстах рассматриваемых жанров показывает, что наибольшие схождения дают служебные слова, глаголы и местоимения и далее, сравнения словника частотный словарь фольклорных жанров с наиболее полным корпусом лексики современного литературного языка (Ўзбек тилининг имло луғати. Москва, 1981) показывает, что в результате проведения нашего лингвостатистического эксперимента обнаружилось около 2750 слов, не засвидетельствованных в словаре. Условно эти слова можно объединить в следующие группы:

1. Название одежды: а) «головной убор»: ***qalpoq*** (Д.i=55; П.i=11; З.i=11)[1]***, lovdon ro'mol*** (Д.i=3)***, ponza ro'mol*** (Д.i=5)***, telpak*** (Д.i=17; П.i=4; З.i=3; С.i=2; Пос.i=2)***, po'ta*** (Д.i=9; П.i=3; З.i=1; С.i=1)***, ovsar*** (Д.i=1)***, jelak*** (Д.i=4;З.i=1)***, dastor*** (Д.i=2;З.i=4; П.i=1)***, jig'a*** (Д.i=6) каби.

б) «верхняя одежда»: *to'n* (Д.i=10; С.i=1; Пос.i=23; З.i=5; П.i=8), ***chopon*** (Д.i=6; С.i=7; Пос.i=2; З.i=1; П.i=3)***, kebanak*** **(**Д.i=4)***, jegda*** (Д.i=1), ***sholvor*** (Д.i=1) каби.

в) «обувь»: *choriq* (Д.i=2; Пос.i=8; П.i=1), ***etik*** (Д.i=10; С.i=1; Пос.i=23; З.i=5; П.i=8), ***kovush*** (Д.i=3; З.i=2; П.i=7), ***maxsi*** (Д.i=2; С.i=2; З.i=1; П.i=2), ***paypoq*** (Д.i=1; З.i=1) каби.

2. Военная терминология:

а) «воинские части»: *qo'shin* (Д.i=29; С.i=9; Пос.i=1), ***qo'sh*** (Д.i=10; С.i=11; Пос.i=11; З.i=25; П.i=9), ***lashkar*** (Д.i=49; С.i=30; Пос.i=2), ***dasta*** (Д.i=22; С.i=3; Пос.i=3; З.i=3; П.i=10), ***navkar*** (Д.i=10; С.i=3; Пос.i=3), ***suvori*** (Д.i=1)***.***

б) «название видов оружья»: *yoy* (Д.i=83; С.i=15; Пос.i=1; З.i=3)***, kamon*** (Д.i=4; С.i=5; Пос.i=1; П.i=4)***, qilich*** (Д.i=59; С.i=110; Пос.i=23; П.i=9; З.i=11)***, isfaxon qilich*** (Д.i=2)***, dastgir qilich*** (Д.i=2)***,***

---

[1] **Д** –*Alpomish.* Fozil Yo'ldosh o'g'li. – Toshkent: «Sharq» nashriyoti-matbaa konserni bosh tahririyati, 1998; *Ravshan.* – Toshkent: Fan, 1954; **П** – *Boychechak*. – Toshkent: G'.G'ulom nomidagi Adabiyot va san'at nashriyoti, 1984; **З** – *Topishmoqlar.* – Toshkent: G'.G'ulom nomidagi Adabiyot va san'at nashriyoti, 1981; **Пос.** – *O'zbek xalq maqollari*. – Toshkent: «Sharq» nashriyot-matbaa aksiyadorlik kompaniyasi bosh tahririyati, 2005; **С** – *O'zbek xalq ertaklari.* I tom. Toshkent: «O'qituvchi» nashriyot-matbaa ijodiy uyi, 2007.
i – частотность слова.

gurzi (Д.i=2; С.i=3)**,** ***parang miltiq*** (Д.i=2; П.i=1; З.i=2)**,** ***qora miltiq*** (Д.i=2)**,** ***jazoyil*** (Д.i=4)**,** ***xurma to'p*** (Д.i=1)**,** ***xitoycha*** (Д.i=1)**.**

в) «название средства защиты» МГ: ***qalqon*** (Д.i=11; С.i=3; Пос.i=2; П.i=2)**,** ***kirovka*** (Д.i=6)**,** ***dubulg'a*** (Д.i=1)**.**

г) «военные передвижение»: ***g'azot*** **(**Д.i=5)**,** ***shabgir tortmoq*** **(**Д.i=6)**,** ***savash*** **(Д**.i=25)**.**

3. Гиппологическая терминология: ***ayil*** (Пос.i=3; Д.i=12; П.i=1; З.i=1)**,** ***pushtan*** (Д.i=14)**,** ***ayil-pushtan*** (Д.i=3)**,** ***toziyona*** (Д.i=1)**,** ***uzangi*** (Пос.i=2; Д.i=23; П.i=3; З.i=2)**,** ***bellik*** (Д.i=3)**,** ***terlik*** (Д.i=2)**,** ***jilov*** (Пос.i=2; Д.i=29; П.i=2; З.i=2; С.i=2)**,** ***zul*** (Д.i=1)**,** ***umuldirik*** (Д.i=16)**,** ***jahaldirik*** (Д.i=2)**,** ***abzal*** (Пос.i=1; Д.i=16) каби.

4. Название животных:

а) «скот»: ***novvos*** (Д.i=3; Пос.i=1)**,** ***ho'kiz*** (Д.i=4; С.i=78; Пос.i=28; П.i=14; З.i=19)**,** ***sigir*** (С.i=4; Пос.i=32; П.i=14; З.i=22)**,** ***tana*** (П.i=1) каби.

б) «конь»: ***baytal*** (Пос.i=2; П.i=2)**,** ***biya*** (Д.i=11; С.i=2; Пос.i=28; П.i=4; З.i=4)**,** ***ayg'ir*** (Д.i=1; С.i=1; Пос.i=4; З.i=2)**,** ***yo'rg'a*** (Д.i=16; С.i=3; Пос.i=10; П.i=10; З.i=14)**,** ***saman*** (Д.i=11; С.i=1; П.i=7; З.i=3)**,** ***do'non*** (Д.i=11; Пос.i=2; З.i=1)**,** ***toy*** (Д.i=17; С.i=2; Пос.i=26; П.i=18; З.i=7)**,** ***to'bichoq*** (Д.i=7)**,** ***bedov*** (Д.i=80; Пос.i=2; П.i=4)**,** ***asov*** (Д.i=5)**,** ***yobi*** (Д.i=10; Пос.i=1)**,** ***to'riq*** (Д.i=7; Пос.i=1; П.i=1; З.i=2) кабилар.

в) «собаки»: ***ko'ppak*** (Д.i=3; Пос.i=2)**,** ***tozi*** (Д.i=95; С.i=15; Пос.i=10; П.i=5; З.i=8)**,** ***kuchuk*** (Д.i=17; С.i=9; Пос.i=7; П.i=23; З.i=4)**,** ***it*** (Д.i=12; С.i=37; Пос.i=110; П.i=23; З.i=20) каби.

г) «свиньи»: ***qobon*** (Д.i=1)**,** ***megajin*** (Д.i=1)**.**

д) «дикие животные»: ***bo'ri*** (Д.i=6; С.i=219; Пос.i=46; П.i=20; З.i=12)**,** ***sher*** (Д.i=39; С.i=64; Пос.i=32; П.i=5; З.i=2)**,** ***yo'lbars*** (Д.i=31; С.i=33; Пос.i=12; П.i=2; З.i=4)**,** ***qarsoq*** (Д.i=1; Пос.i=1)**,** ***tulki*** (Д.i=3; С.i=271; Пос.i=34; П.i=13; З.i=11)**.**

е) «траваядные»: ***kulon*** (Д.i=4; П.i=1)**,** ***bulon*** (Д.i=3)**,** ***kiyik*** (Д.i=4; С.i=49; Пос.i=6; П.i=4; З.i=4)**,** ***ohu*** (Д.i=12; Пос.i=1)**.**

5. Название птиц:

а) «певчие птицы»: ***bulbul*** (Д.i=45; С.i=23; Пос.i=16; П.i=42; З.i=1)**,** ***tovus*** (Д.i=3; С.i=9; Пос.i=2; П.i=1; З.i=1)**,** ***to'ti*** (Д.i=4; С.i=14; П.i=3; З.i=1)**.**

б) «степные птицы»: ***mayna*** (Д.i=59; С.i=2; Пос.i=1; П.i=1; З.i=1)**,** ***laylak*** (Д.i=36; С.i=53; Пос.i=2; П.i=25; З.i=11)**,** ***g'urrak*** (Д.i=1; П.i=1)**,** ***g'azalay*** (Д.i=3)**,** ***puchchakkalon*** (Д.i=1; П.i=1)**,** ***olashaqshaq*** (Д.i=1; З.i=1)**,** ***musicha*** (Д.i=4; Пос.i=1; П.i=4)**,** ***g'oz*** (Д.i=64)**.**

в) «дикие птицы»: ***g'ajir*** (Д.i=8; Пос.i=1; З.i=1)**,** ***boyqush*** (Д.i=1; Пос.i=1)**,** ***burgut*** (Д.i=7; С.i=17; Пос.i=4; П.i=4; З.i=2)**,** ***qarchig'ay*** (Д.i=26; С.i=2; Пос.i=1; П.i=1) кабилар.

6. Название продуктов питания: ***bulamig'*** (Пос.i=2)**,** ***atala*** (С.i=3; П.i=8)**,** ***pista*** (Д.i=7; Пос.i=2; П.i=27; З.i=1)**,** ***pista-rusta*** (Д.i=1)**,** ***qatlama*** (Д.i=3; С.i=25; З.i=4) кабилар.

7. Природные явление: ***chaqmoq*** (Д.i=1; С.i=2; Пос.i=4; П.i=2; З.i=2)**,** ***qor-yomg'ir*** (С.i=2)**,** ***yomg'ir-yomg'ir*** (Д.i=1)**,** ***momaqaldiroq*** (Д.i=1; Пос.i=1; З.i=3)**.**

8. Название степени и званий: ***o'gach*** (Д.i=1), ***o'dag'a*** (Д.i=3), ***salom og'a*** (Д.i=1), ***xudaychi*** (Д.i=2), ***sayis*** (Д.i=1), ***kengashboshi*** (Д.i=1).

9. Название музыкальных инструментов: ***karnay*** (Д.i=9; С.i=18; Пос.i=2; П.i=3; З.i=4)**,** ***karnay-surnay*** (Д.i=4; С.i=4)**,** ***do'mbira*** (Д.i=1)**,** ***doira*** (С.i=9; З.i=3).

Эти группы слов можно объединить в следующей таблице:

№	Семантические группы	дастан	сказки	послови-цы	песни	загадки
1.	Название растений	22	33	43	58	55
2.	Название животных	44	52	56	58	56
3.	Название птиц	21	21	23	25	24
4.	Название топоним.объек.	31	10	6	24	11
5.	Название посуды	20	21	34	39	36
6.	Название прод.питания	28	35	43	50	53
7.	Название муз.инструм.	4	8	9	10	13
8.	Религиозные термины	119	10	12	9	6
9.	Название обычаев	30	0	0	0	0
10.	Архитектурные названия	20	1	1	1	2
11.	Название степеней и званий	37	2	5	2	3
12.	Название виды оружия и.т.д	58	10	4	4	4
13.	Гиппологическая терминология	72	7	8	6	10
14.	Диалектизмы	220	47	38	57	65

Отметим, что остальные слова являются почти без исключения стилистически нейтральными в современном языке и совершенно не содержат устаревших словообразовательных элементов.

Изложенные выше наблюдения дают основания для следующего предположения. Наше восприятие словарного состава текстов фольклора как близко к современному литературному языку, очевидно, связано не вообще с составом словника произведений, но прежде всего с составом той небольшой группы наиболее частых слов, которые составляют преобладающую долю словоупотреблений в тексте. Напомним, что 200 наиболее частых слов словника фольклорных жанров, среди которых всего 20 слов могут быть отнесены к устаревшим или значительно изменившим систему значений, составляют около 62 % словоупотреблений всего текста жанров.

Подробные соображения, как нам кажется, должны учитываться при всяком сравнении словарного состава текстов, относящихся к разным периодам истории литературного языка.

ЛИТЕРАТУРА:

1. Айимбетов М. К. Проблемы и методы квантитативно-типологического измерения близости тюркских языков. Автореф.дисс..док. филол.наук. – Ташкент, 1997.

2. Квантитативная лингвистика и автоматический анализ текстов. – ТАРТУ, 1985.

3. Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А. Математическая лингвистика: учебное пособие для пединститутов. – М.: Высшая школа, 1977. С.368.

4. Статистика речи и автоматический анализ текста. – Ленинград, 1980.

5. Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики. – Таллин: Валгус, 1987.

Материалы для обработки

1. Dastan. *Alpomish.* (1998) Fozil Yo‘ldosh o‘g‘li. Toshkent: «Sharq» nashriyoti-matbaa konserni bosh tahririyati: *Ravshan.* (1954) Toshkent: Fan.

2. Сказки. *O’zbek xalq ertaklari.* (2007) I tom. Toshkent: «O’qituvchi» nashriyot-matbaa ijodiy uyi.

3. Пословицы. *O’zbek xalq maqollari*. (2005) Toshkent: «Sharq» nashriyot-matbaa aksiyadorlik kompaniyasi bosh tahririyati.

4. Загадки. *Topishmoqlar.* (1981) Toshkent: G‘.G‘ulom nomidagi Adabiyot va san’at nashriyoti.

5. Песни. *Бойчечак*. (1984) Тошкент: Ғ.Ғулом номидаги Адабиёт ва санъат нашриёти.

УДК 81’33

PRINCIPLES OF CREATING THE INTERFACE OF THE UZBEK LANGUAGE’S AUTHORSHIP CORPUS (ON THE EXAMPLE OF THE AUTHORSHIP CORPUS OF ABDULLAH KAHHAR)

Shahlo Khamroeva[a], Bakhtiyor Mengliev[b]
[a]Bukhara State University, Bukhara, Uzbekistan.
[b]Tashkent State University of the Uzbek Language and literature, Tashkent, Uzbekistan
hamroyeva81@mail.ru, b.mengliev@inbox.com

The article analyzes the principles of creating the authorship corpus of the Uzbek language, the interface and the search engine of the authorship corpus, the annotation and markup system, the corpus units, the design of the interface. The interface of the authorship corpus provides a an opportunity to understand the genre classification of the author’s creative heritage. Designing an interface requires a lot of responsibility. The interfaces of the authorship corpus show that they have a different design and structure, and a perfect interface requires special aesthetics from the creators. An important factor of designing the interface of the authorship corpus is the reflaction of the author’s creative heritage in it. The interface of A. Kahhor’s corps can be divided into the following windows: «Autobiography of Abdulla Kahhar», «Creation of Abdulla Kahhar», «Memoirs about Abdulla Kahhar», «Search (field for entering words)», «Concordance», «Advanced Search». During the creation of the corpus the peculiarities of the Uzbek language and the work of A. Kahhar were considered.

Keywords: authorship corpus, interface, search engine, annotation and markup system, corpus material, corpus units, interface design.

ПРИНЦИПЫ СОЗДАНИЯ ИНТЕРФЕЙСА АВТОРСКИХ КОРПУСОВ УЗБЕКСКОГО ЯЗЫКА (НА ПРИМЕРЕ АВТОРСКОГО КОРПУСА АБДУЛЛЫ КАХХАРА)

Шахло Хамроева[a], Бахтиёр Менглиев[b]
[a]Бухарский Государственный Университет, Бухара, Узбекистан
[b]Ташкентский Государственный Университет узбекского языка и литературы, Ташкент, Узбекистан
hamroyeva81@mail.ru, b.mengliev@inbox.com

В статье представлен анализ принципов создания авторских корпусов узбекского языка, интерфейса и поисковой системы авторского корпуса, системы аннотирования и разметки, проектирования и дизайна интерфейса. Интерфейс авторского корпуса предоставляет возможность понять жанровую классификацию творческого наследия автора, что является существенным отличием корпуса от других электронных продуктов. Как известно, разработка интерфейса требует большой ответственности, поскольку интерфейс является важной, неотъемлемой частью корпуса и его сайта. Интерфейсы авторских корпусов показывают, что они имеют различный дизайн и структуру, что для идеального интерфейса требуется особая эстетика от создателей корпуса. Для создания авторского корпуса важным фактором является отражение в интерфейсе авторского творческого наследия. Интерфейс корпуса А. Каххара можно разделить на следующие окна: «Автобиография Абдуллы Каххара», «Творчество Абдуллы Каххара», «Мемуары и воспоминания об Абдулле Каххаре», «Поиск (поле для ввода слова)», «Конкорданс», «Расширенный поиск». При создании корпуса учтены особенности узбекского языка и многогранность творчества Абдуллы Каххара.

Ключевые слова: авторский корпус; интерфейс; поисковая система; система аннотирования и разметки; единицы корпуса.

1. Введение

Требования к интерфейсу авторского корпуса

Интерфейс корпуса является очень важным компонентом, который знакомит пользователя со всем корпусом (Добровольский О. Д., 2005). Интерфейс авторского корпуса предоставляет хорошую возможность понять жанровую классификацию творческого наследия автора. По сути, это и есть отличие корпуса от других электронных продуктов. Разработка интерфейса требует большой ответственности, поскольку интерфейс является важной, неотъемлемой частью корпуса и его сайта. Наблюдение за актуальными интерфейсами авторских корпусов показывают, что они имеют различный дизайн и структуру, что для идеального интерфейса требуется особая эстетика от авторов корпуса (Потемкин С. Б.). Национальная индивидуальность в корпусе обеспечивается проявлением этнического колорита в общем плане интерфейса, орнаменты, отражающие национальный колорит на заднем плане, а также классические или современные дизайнерские особенности.

2. Проектирование интерфейса авторского корпуса Абдуллы Каххара

Интерфейс корпуса А.Каххора можно разделить на следующие окна: «Автобиография Абдуллы Каххара», «Творчество Абдуллы Каххара», «Мемуары и воспоминания об Абдулле Каххаре», «Поиск (поле для ввода слова)», «Конкорданс», «Расширенный поиск».

3. Отражение авторского творческого наследия в интерфейсе

Корпус Абдуллы Каххара должен охватывать различные тексты (корпусные единицы), связанные с его жизнью и творчеством, поскольку тексты корпуса не могут состоять только из литературных примеров. Поэтому необходимо выбирать все работы: прежде всего, научные, публицистические статьи и воспоминания о писателе, о его жизни и творчестве. Главное, чтобы тексты (корпусные единицы) полностью показывали жизнь и творчество Абдуллы Каххара. В окно «Жизненный путь А.Каххара» будет включена информация о жизни А. Каххара, которая и позже может будет обогащена. Произведения писателя будут помещены в окно «Творчество Абдуллы Каххара» на основе жанровых классификаций: «Роман», «Сказка», «Пьеса», «Сказка», «Фельетон», «Портрет», «Статьи и беседы», «Записки из тетради». Это окно является самой важной частью – основной базой корпуса. Корпус будет выбирать наиболее полное издание (Каххар А., 1987) его произведений, а также его произведений в сборниках, которые были опубликованы за годы независимости.

В окно «Воспоминания об Абдулле Каххаре» будут включены все воспоминания и статьи о жизни и творчестве писателя. Материалы, которые находятся в интерфейсе, позволят использовать корпус в качестве электронной библиотеки. Окно «Поиск» является очень важной частью поисковой системы корпуса. В этой системе окно «Поиск из всех произведений», «Конкорданс», «Поиск речи героев» и «Полный конкорданс произведений Абдуллы Каххара» позволит использовать все возможности корпуса. Обеспечение поиска на основе поискового окна связано с программным обеспечением и аннотацией единиц корпуса.

Окно «Расширенный поиск» поможет выполнить поиск произведений, жанров и речи героев. Работа этих окон связана с экс-

тралингвистической аннотацией единиц корпуса – текста. То есть для каждого из этих окон должны быть активны данные, которые записывались, как жанр произведения, год, место и т. д. для каждого текста корпуса.

Окно «Конкорданс» – это компонент, который позволяет вам создать частотный словарь произведений Абдуллы Каххара или использовать корпуса в качестве частотного словаря. Известно, что в практике лексикографии были составлены несколько словарей (Каримов С., Қаршиев А., Исроилова Г., 2007) на основе произведений А. Каххара. При наличии доступного программного обеспечения, можно создать новое поколение словарей (частотные, идеографические словари) по творчеству А. Каххара.

Поиск конкорданса должен определить частоту любого слова, найденного в произведениях А. Каххара и перейти к контексту этого слова. В окне поиска вы можете искать материал корпуса, то есть литературное наследие писателя. Если введется условие поиска в это поле, можно будет выбрать один из следующих параметров.

Откроется окно «Список контекста», в котором вы можете найти результат запроса. В верхней части окна отображается количество результатов.

Например: *по этому запросу найдено 29 контекстов. Ключевые слова:* _____

1. Результат появляется примерно в следующем виде:

№	Название произведение	Из речи какого героя	Номер страницы в произведение	Результат поиска – контекст

Если вы ищете слово в окне «Поиск конкорданс», вы найдете частоту слова в корпусе, то есть в произведениях А.Каххара. Результат будет выглядеть примерно так:

«Қаҳҳор асарлари конкорданси»
- ***даҳшат*** *сўзи бўйича ушбу материаллар топилди*
- ***даҳшат*** *сўзи 21 та асарда 146 гап ичида 158 марта ишлатилган.*

Қуйида ***даҳшат*** *сўзи учраган матн ва сўз миқдорини кўринг.*

Поиск по параметру «Конкорданс» дает информацию о количестве словоформ в произведениях автора. При нажатии слово переходит к контексту, в котором оно встречается. Наблюдения за существующими корпусами показали, что поиск может быть выполнен по любой части слова или фразы[1].

«Поиск речи героя» обычно осуществляется по тексту драматического произведения. В зависимости от уровня синтаксической разметки можно осуществлять поиск по прозаическим текстам. Если вы введете имя нужного героя в это поле поиска и введете команду поиска, вы можете получить следующие результаты[2]:

речь		**герой**		**произведение** (нажмите на имя героя, чтобы увидеть его речь)
33		**Lodovico**		Othello

При нажатии на окно с частотой слов, содержащее работу, будут показаны все контексты: название пьесы, заголовок и количество просмотров и контекст. Могут быть выбраны также два параметра: только контекст речи героя или диалог речи героя.

2	IV,1,2651	***Othello.*** *With all my heart, sir.* **Lodovico.** The duke and senators of Venice greet you.

Отдельной областью интерфейса является страница, показывающая возможности автора корпуса, которая содержит несколько важных окон. Главное окно этой платформы – «Конкорданс» (Столяров А. И., 2017). Запустив это окно, вы перейдете на страницу «Полный конкорданс произведений Абдуллы Каххара». Эта страница содержит окно «алфавит» и «цифры»: это алфавитный указатель, то есть частотный словарь – конкорданс произведений А. Каххара. Число перед каждой буквой указывает на количество слов, начинающихся с этой буквы. Для упрощения и автоматизации поиска Корконданс показан в алфавитном порядке. При нажатии на выбранную букву в списке слов, отображается числовое значение. Если выбрать нужную букву, будет отображаться

[1] https://www.opensourceshakespeare.org

[2] Эти и последующие окна как пример взяты непосредственно из авторского корпуса У. Шекспира.

информация об использовании слов в контексте. При нажатии на название произведения можно будет перейти на контекст, в котором использовано слово. Если ввести фразу в поисковое поле «Полный конкорданс произведений Абдуллы Каххара», на интерфейсе можно будет увидеть конкорданс или частоту использованных слов автора. Нажатие на слово и число указывает на объем произведения в разделе; нажатие на название работы, указывает на контекст. Опыт авторского корпуса У.Шекспира показывает, что доступен поиск одновременно до 6 слов в полном тексте произведений У. Шекспира.

Преимущество системы «Расширенного поиска» состоит в том, что у системы есть несколько возможностей. Когда необходима простая форма слова, желательно выполнить поиск по ключевому слову. Корпус быстро генерирует результаты поиска по части слова; предлагает много вариантов; поисковое окно «словоформа» находит более точные варианты, но относительно медленно. Поиск по ключевым словам не может различить словообразовательные процессы. На странице расширенного поиска можно искать информацию о словоформах или словосочетаниях по жанру, количеству слов, размеру слова, именам героев.

Отдельное поле интерфейса содержит информацию о корпусе, его структуре, назначении, задачах и авторах. В нижней части средней области допустимо размещение окна, такие как «Контакты» и «Форум». Было бы целесообразно предоставить информацию об авторах корпуса, организации, основах программирования.

Надо отметить, что на основе общих принципов создания авторского корпуса должны разрабатываться принципы создания авторского корпуса Абдуллы Каххара. При создании корпуса необходимо учесть бесценный опыт иностранных специалистов этой области, особенности узбекского языка и многогранное творчество Абдуллы Каххара. Решающими факторами являются: правильное размещение содержимого материала, значение содержания корпуса, аннотирование единиц авторского корпуса. Исходя из этого, основные этапы создания Корпуса Абдуллы Каххара можно обобщить следующим образом:

1) проектирование дизайна интерфейса;

2) подбор и обработка материалов для корпуса;

3) разработка принципов аннотирования языковых единиц корпуса;

4) моделирование морфологических, семантических, синтаксических тегов узбекского языка;

5) разработка программного обеспечения.

4. Заключение

Для составления «Корпуса произведений Абдуллы Каххара» необходимо разработать морфологические, семантические, синтаксические принципы разметки, так как нет программного обеспечения для автоматического аннотирования единиц узбекского языка. Также важно выбрать материал для корпуса Абдуллы Каххара: материалы о жизни и творчестве писателя, подборка всех произведений А.Каххара, подборка статей и воспоминаний о жизни писателя. Наличие этой информации в корпусе позволяет использовать ее в качестве электронной библиотеки, аннотирование единиц (текстов) корпуса могут служить многофункциональным источником информации для выполнения различных лингвистических практик над текстами произведений Абдуллы Каххара, по литературоведению, истории, этнографии, лингвистике и лингвокультурологии. Пятитомное издание произведений А.Каххара станет основным источником «Корпуса произведений Абдуллы Каххара».

ЛИТЕРАТУРА

1. Добровольский О. Д., Кретов А. А., Шаров С. А. (2005) Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. – С. 263–296.

2. Захаров В. П., Богданова С. Ю. (2011) Корпусная лингвистика.– Иркутск: ИГЛУ, 2011. – 161 с.

3. Қаххор, А. (1987) Асарлар. 5 ж. / Редкол. Ў. Умарбеков, Саид Аҳмад, Э. Аъзам ва бошқ.; нашрга тайёрловчи К. Қаххорова. Ж.1. Сароб. Роман. Ҳикоялар. – Т. : Адабиёт ва санъат нашриёти, 1987. – 357 б.

4. Қаххор, А. (1987) Асарлар. 5 ж. / Редкол. Ў. Умарбеков, Саид Аҳмад, Э. Аъзам ва бошқ.; нашрга тайёрловчи К. Қаххорова. Ж. 2. Қўшчинор чироқлари. Роман. Ҳикоялар. – Т. : Адабиёт ва санъат нашриёти, 1987. – 368 б.

5. Қаххор, А. (1987) Асарлар. 5 ж. / Редкол. Ў. Умарбеков, Саид Аҳмад, Э. Аъзам ва бошқ.; нашрга тайёрловчи К. Қаххорова. Ж. 3. Ўтмишдан эртаклар: қиссалар. – Т. : Адабиёт ва санъат нашриёти, 1987. – 368 б.

6. Қаҳҳор, А. (1987) Асарлар. 5 ж. / Редкол. Ў. Умарбеков, Саид Аҳмад, Э. Аъзам ва бошқ.; нашрга тайёрловчи К. Қаҳҳорова. Ж. 4. Аяжонларим. Пьесалар. Портретлар. Фельетонлар. – Т. : Адабиёт ва санъат нашриёти, 1987. – 335 б.

7. Қаҳҳор, А. (1987) Асарлар. 5 ж. / Редкол. Ў. Умарбеков, Саид Аҳмад, Э. Аъзам ва бошқ.; нашрга тайёрловчи К. Қаҳҳорова. Ж. 5. Ҳақ «сўзнинг кучи: мақолалар, суҳбатлар, қайдлар. – Т. : Адабиёт ва санъат нашриёти, 1987. – 272 б.

8. Каримов С., Қаршиев А., Исроилова Г. (2007) Абдулла Қаҳҳор асарлари тилининг луғати. Алфавитли луғат. – Т.: ЎзМЭ, 2007. – 434 б.

9. Каримов С., Қаршиев А., Исроилова Г. (2007) Абдулла Қаҳҳор асарлари тилининг луғати. Терс луғат. – Т.: ЎзМЭ, 2007. – 423 б.

10. Каримов С., Қаршиев А., Исроилова Г. (2007) Абдулла Қаҳҳор асарлари тилининг луғати. Частотали луғат. – Т.: ЎзМЭ, 2007. – 420 б.

11. Потемкин С.Б. Авторский корпус и словарь языка Антона Чехова // https: istina.msu.ru.

12. Столяров А.И. (2017) Словарь-конкорданс и его применение в рамках корпусной лингвистики // Гуманитарные научные исследования. –2017. – № 2. http://human.snauka.ru/2017/02/21074.

УДК 81’33

BUILDING A NEURAL NETWORK SYSTEM FOR RUSSIAN-TATAR MACHINE TRANSLATION

Khusainov Aidar[1], Khusainova Albina[3], Suleymanov Dzhavdet[1,2], Gilmullin Rinat[1]
[1]Institute of Applied Semiotics, Academy of Sciences of the Republic of Tatarstan, Kazan, Russian
[2]Kazan Federal University, Kazan, Russia
[3]Innopolis University, Innopolis, Russia

In this paper, we solve the problem of constructing a machine translation system for the Russian-Tatar language pair. A neural network approach based on the Transformer network architecture is used, as well as various algorithms for increasing the volume of training data and the use of monolingual data, e.g. back-translation. For the first time, experiments were conducted for the Russian-Tatar MT on the use of parallel data for other languages (the Kazakh-Russian parallel corpus) in order to transfer knowledge of the neural network (transfer learning). As the main training data the created parallel corpus with a total volume of 983 thousand pairs of Russian-Tatar sentences is used. This corpus includes news, literature, translations of laws and other official documents. The experiments show that the created system is superior in quality to the currently existing Russian-Tatar translators. To summarize, we assess the possibility of using the Transformer architecture of neural network, back-translation algorithm, and transfer learning algorithms to the construction of the machine translation system for the Tatar-Russian language pair. The main result of the work is an increase in quality in terms of BLEU scores to 35 and 39 for RU-TT and TT-RU translation directions respectively.

Keywords: neural machine translation; Tatar language; low-resourced language.

К ВОПРОСУ ПОСТРОЕНИЯ НЕЙРОСЕТЕВОЙ СИСТЕМЫ РУССКО-ТАТАРСКОГО МАШИННОГО ПЕРЕВОДА

Хусаинов Айдар[1], Хусаинова Альбина[3], Сулейманов Джавдет[1,2], Гильмуллин Ринат[1]
[1]Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
[2]Казанский федеральный университет, Казань, Россия
[3]Университет Иннополис, Иннополис, Россия

Данная статья направлена на решение задачи построения системы машинного перевода для русско-татарской языковой пары. За

основу был выбран нейросетевой подход, основанный на архитектуре Transformer, также использовались различные алгоритмы для искусственного увеличения объема параллельных обучающих данных и использования моноязычных данных, например, метод обратного перевода. Впервые для русско-татарской пары были проведены эксперименты по использованию параллельных данных для других языков (казахско-русский параллельный корпус) с целью передачи знаний нейронной сети (transfer learning). В качестве основных обучающих данных использовался созданный параллельный корпус общим объемом 983 тысячи пар русско-татарских предложений. В этот корпус вошли новости, литература, переводы законов и другие официальные документы. Эксперименты показали, что созданная система по качеству превосходит другие общедоступные русско-татарские переводчики. Значение качества перевода по метрике BLEU составило 35 и 39 для направлений перевода с русского на татарский и татарского на русский, соответственно.

Ключевые слова: машинный перевод на нейронных сетях, татарский язык, малоресурсный язык.

1. Введение

Русский и татарский языки являются государственными языками в Республике Татарстан. Этот факт делает актуальной задачу обеспечения населения, государственных и других учреждений возможностью автоматического перевода между этими языками.

Исходя из особенностей татарского и русского языков, в институте прикладной семиотики для решения задачи машинного перевода был выбран статистический подход. Это определило приоритетные задачи, заключающиеся в построении морфологического анализатора татарского языка, способного снизить зависимость моделей от сложности татарской морфологии, и накоплении параллельных русско-татарских предложений, необходимых для обучения статистических моделей.

Результатом работы в рамках направления построения статистического русско-татарского переводчика, основанного на фразах (phrase-based MT), стала общедоступная версия переводчика от компании Яндекс, запущенная в 2015 году. Первая версия Яндекс.Переводчика для данной языковой пары была обучена, в том числе, с использованием морфоанализатора и параллельного корпуса, разработанных институтом прикладной семиотики.

В 2018 году нами был разработана первая версия русско-татарского переводчика на основе нейросетевого подхода [1, 2]. Для обучения системы мы использовали инструментарий Nematus [3]

с усовершенствованиями, предложенными в [4]. В основе подхода была выбрана архитектура сети encoder-decoder-attention, каждая часть которой представляла собой одно- (для случая декодера с механизмом внимания) или двунаправленную (для энкодера) рекуррентную нейросеть. Для решения проблемы большого количества внесловарных слов (OOV problem) мы использовали базовые единицы, построенные на основе алгоритма BPE (byte-pair encoding) [5]. Модель разбиения слов на составляющие части была применена к объединенному русско-татарском корпусу.

В данной работе представлены результаты доработок системы машинного перевода для русско-татарской языковой пары. Используется нейросетевой подход на базе архитектуры сети Transformer [6], а также алгоритм использования моноязычных данных для увеличения объема обучающих данных back-translation. Впервые для русско-татарской пары проведены эксперименты по использованию параллельных данных для других языков (а именно казахско-русского параллельного корпуса) с целью переноса знаний нейросети (transfer learning). В качестве основных обучающих данных используется созданный параллельный корпус общим объемом 983 тысячи пар русско-татарских предложений, включающий тексты новостной тематики, литературу, переводы законов и нормативных актов. Проведенные эксперименты показывают, что созданная система превосходит по качеству существующие на данный момент переводчики.

В разделе 2 данной статьи представлен обзор лингвистических ресурсов, подготовленных и использованных для обучения системы, раздел 3 содержит результаты проведенных экспериментов.

2. Лингвистические ресурсы для обучения

Ключевым фактором, влияющим на качество системы машинного перевода, является объем репрезентативного параллельного корпуса, на котором обучается система. Современные системы перевода для крупных мировых языков обучаются на параллельных данных общим объемом в десятки и сотни миллионов пар предложений. Корпус такого объема позволяет обеспечить устойчивость работы системы на различных текстах.

На текущий момент объем накопленного русско-татарского корпуса составляет 983 тысячи пар предложений, основными источниками для которого были двуязычные книги, законы, нормативные акты, новости, а также тексты, переведенные вручную.

Основные характеристики корпуса представлены в таблице 1.

Таблица 1. Основные характеристики русско-татарского параллельного корпуса

Параметр	Значение
Количество параллельных предложений	983 319
Количество слов в русской части корпуса	15 032 363 (15,3 слов/ предложение)
Количество слов в татарской части корпуса	14 649 484 (14,9 слов/ предложение)
Количество предложений в обучающей/ тестовой/валидационной выборках	977539 / 2499 / 2499

В данной работе наряду с параллельным корпусом также используются моноязычные корпуса для русского и татарского языков. Они необходимы для применения в рамках метода обратного перевода (back-translation), который позволяет на основе предобученной модели переводчика искусственно расширять объем обучающих параллельных данных.

В качестве моноязычных корпусов были использованы корпуса из коллекции Лейпцигского университета [7]. Для русского языка были объединены подкорпуса новостной тематики (news_2010), Интернет-текстов (web_2015), подкорпус Википедии (wiki-2016); для татарского – Интернет подкорпус (web-2018), новостной (news_2015) и смешанный (mix-2015).

Каждый из перечисленных подкорпусов имеет объем равный 1 млн словоформ. Предложения из объединенных коллекций были отфильтрованы с целью удаления дубликатов. Итоговый объем русского корпуса составил 2 999 489 предложений, татарского – 2 355 738 предложений.

Архитектура систем машинного перевода

В рамках данной работы были построены несколько вариантов системы машинного перевода, которые отличались по 3 параметрам:

1. Размеру нейросети (Base/Big);
2. Использованию метода обратного перевода (back-translation) для увеличения объема обучающего корпуса (BT);

3. Использованию русско-казахского параллельного корпуса для метода переноса знаний, transfer learning (TL).

Архитектура нейросети Base имеет следующие параметры: batch size – 2048, hidden size – 512, filter size – 2048, multi-headed attention heads – 8, encoder/decoder's hidden layers – 6, dropout – 0.1, learning rate – 2.0, beam size – 4; вариант Big имеет удвоенные значения для параметров batch size, hidden size, filter size, multi-headed attention heads.

Синтетическая часть параллельного корпуса формировалась путем перевода выборки из моноязычных корпусов с помощью моделей Base. Объем синтетической части был выбран равным объему исходного параллельного корпуса (1 миллион пар предложений).

Для проведения эксперимента с переносом знаний был использован русско-казахский корпус, опубликованный в рамках соревнования WMT-2019 [8]. Его объем составляет 5 миллионов пар предложений.

Для объективной оценки качества перевода различных версий систем использовалась метрика BLEU [9], использование которой остаётся мировым стандартом в области оценки качества систем машинного перевода.

Результаты оценки качества проведенных экспериментов приведены в Таблице 3.

Таблица 3. Значения метрики BLEU для систем машинного перевода

Архитектура нейросети	Количество итераций обучения	Направление перевода	BLEU (без учета регистра)
Base	40	RU-TT	**35.39**
Base	40	TT-RU	38.42
Big	10	RU-TT	34.08
Big	10	TT-RU	37.07
Base_BT	40	RU-TT	34.42
Base_BT	40	TT-RU	**39.21**
Big_TL	10 + 10	RU-TT	34.41
Big_TL	10 + 10	TT-RU	36.08
Yandex [10]	–	RU-TT	15.59
Yandex [10]	–	TT-RU	18.15

В отличие от известных опубликованных результатов аналогичных экспериментов, мы не наблюдаем значительного улучшения качества работы при увеличении объема обучающего корпуса за счет синтетически сформированного параллельного подкорпуса. Для русско-татарского направления перевода значение BLEU 34.42 оказалось меньше, чем для Base – 35.39; прирост метрики для татарско-русского направления перевода составил 0.79 BLEU (рост с 38.42 до 39.21).

Результаты свидетельствуют о росте качества работы переводчика в случае использования алгоритма переноса знаний: значения BLEU выросли на 0.33 и 2.14 для русско-татарского и татарско-русского напарвлений перевода, соответственно.

При этом абсолютно лучшие значений на тестовом подкорпусе были показаны системой Base для русско-татарского перевода (35.39 BLEU) и BASE_BT – для татарско-русского (39.21 BLEU).

4. Заключение

В этой статье мы представили результаты экспериментов по разработке системы русско-татарского машинного перевода, построенной на основе нейросетевой архитектуры Transformer. Был подготовлен обучающий корпус параллельных текстов, применены современные методы машинного обучения. Полученная система перевода значительно превосходит имеющиеся переводчики для данной языковой пары.

СПИСОК ЛИТЕРАТУРЫ

1. Khusainov, A., Suleymanov, D., Gilmullin, R., Gatiatullin, A. Building the Tatar-Russian NMT System Based on Re-translation of Multilingual Data. In. Proc. 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018. P. 163–170. (DOI 10.1007/978-3-030-00794-2_17).

2. А. Хусаинов, Д. Сулейманов, Р. Гильмуллин. Система русско-татарского нейронного машинного перевода. Proceedings of the 6th International Conference on Turkic Languages Processing (TURKLANG-2018). (Tashkent, October 18–20, 2018). – Tashkent, 2018.

3. Open-source neural machine translation in Theano [Electronic resource]. URL: https://github.com/ rsennrich/nematus [Дата посещения: 2019].

4. Sennrich, R., Birch, A., Currey, A., Germann, U., Haddow, B., Heafield, K., Valerio Miceli Barone,A.,Williams,P. The University of Edinburgh's neural mt systems for wmt17. In Proceedings of the Second Conference on Machine Translation. vol. 2: Shared Task Papers. Stroudsburg, PA, USA. 2017.

5. Sennrich, R., Haddow, B., Birch, A. Neural Machine Translation of Rare Words with Sub-word Units. ArXiv e-prints. 2015.

6. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Conference on Advances in Neural Information Processing Systems (NIPS). 2017.

7. Corpora Collection Leipzig University [Electronic resource]. URL: https://corpora.uni-leipzig.de/en?corpusId=tat_web_2018 [Дата посещения: 2019].

8. ACL 2019 Fourth Conference on Machine Translation (WMT19) [Electronic resource]. URL: http://www.statmt.org/wmt19/

9. Papineni,K.,Roukos,S.,Ward,T.,Zhu,W.: BLEU: a method for automatic evaluation of machine translation. In: In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. 2002.

10. Yandex translate. https://translate.yandex.com/ (2017), [Online]

УДК 81’33

REVIEW OF THE CREATED RESOURCES AND SOFTWARE FOR THE TATAR SPEECH SYNTHESIS TASK

Khusainov Aidar[1], Suleymanov Dzhavdet[1,2]
[1]Institute of Applied Semiotics, Academy of Sciences of the Republic of Tatarstan, Kazan, Russia
[2]Kazan Federal University, Kazan, Russia

In this paper, we describe the main stages of creating systems for the synthesis of Tatar speech. This description covers our research starting from the first diphone-based concatenative synthesis system built in 1990s to the last end-to-end neural system built in 2019. Despite the significant difference in technology, the need to create a high-quality corpus of sounding speech remains an unchanged condition for the construction of a synthesizer. We present several single-speaker corpora recorded in sound recording studio, each of the corpus was created for a specific synthesis technology. This fact led to difference in total duration and annotation of corpora. Preliminary experiments showed that the best quality speech synthesizer can be built using several neural approaches, but the only method that also provides real-time synthesis uses Tacotron architecture followed by neural vocoder.

Keywords: speech synthesis; speech corpora, Tatar language; low-resourced language.

ОБЗОР СОЗДАННЫХ РЕСУРСОВ И ПРОГРАММНЫХ СРЕДСТВ ДЛЯ СИНТЕЗА ТАТАРСКОЙ РЕЧИ

Хусаинов Айдар[1], Сулейманов Джавдет[1,2]
[1]Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
[2]Казанский федеральный университет, Казань, Россия

В данной статье описаны основные этапы создания систем синтеза татарской речи. Обзор охватывает результаты исследований, проводимых нами, начиная с первой системы конкатенативного синтеза на основе дифонов, созданной в 1990-х годах, до последней нейронной системы, построенной в 2019 году. Несмотря на значительную разницу в технологиях, необходимость создания высококачественного речевого корпуса остается неизменным условием построения качественного синтезатора. Мы представляем несколько однодикторных корпусов, записанных в профессиональной студии, каждый из которых был создан для определенной технологии синтеза. Это привело к разнице в общей продолжительности корпусов и выбранных подходов к их аннотации. Предварительные эксперименты показали, что синтезатор

речи наилучшего качества может быть построен с использованием нескольких нейросетевых подходов, но единственный подход, который при этом обеспечивает синтез речи в реальном времени, основан на последовательной работе нейросети архитектуры Tacotron и нейросетевого вокодера WaveGlow.

Ключевые слова: синтез речи, речевой корпус, татарский язык, малоресурсный язык.

1. Введение

Задача синтеза речи состоит в генерации речевого сигнала на основе произвольного текстового фрагмента. По мере внедрения речевого интерфейса взаимодействия с различными устройствами, повысилась важность синтезаторов речи, способных эффективно генерировать естественно звучащую и разборчивую речь. Несмотря на значительные успехи последних 5 лет, синтезаторы речи не созданы для подавляющего большинства мировых языков.

В данной статье мы представляем краткий обзор созданных в институте прикладной семиотики АН РТ речевых корпусов и программных средств синтеза татарской речи.

В разделе 2 данной статьи представлен обзор основных подходов к синтезу речи, раздел 3 описывает созданные обучающие корпуса, раздел 4 содержит описание созданных систем синтеза татарской речи.

2. Обзор основных подходов к синтезу речи

Среди требований, предъявляемых к системам синтеза речи, основными являются качество синтезированного сигнала и скорость работы систем. Скорость работы позволяет использовать синтезатор в приложениях реального времени, таких как голосовые ассистенты, при автоматическом озвучивании перевода. Оценка качества автоматического синтеза речи, в свою очередь, также представляет собой сложную задачу. Применяются автоматические методы, в том числе использующие оценки вероятностей, полученные с помощью систем распознавания речи, однако такие характеристики могут лишь косвенно свидетельствовать о качестве речевого сигнала. Общепринятыми на данный момент методиками оценки являются экспертные оценки согласно методикам MOS, MUSHRA [1]. Среди критериев, по которым происходит оценка, основными являются разборчивость (легкость

восприятия произнесенного текста) и естественность звучания синтезированной речи.

Среди подходов, разработанных для решения задач синтеза речи, можно выделить несколько основных классов:

- конкатенативные подходы (дифонный синтез [2], Unit selection [3]). Исходной информацией в данных подходах служат выделенные из речевых записей акустические единицы;
- параметрические подходы, в котором базовыми элементами являются статистические модели звуков языка;
- гибридные модели, совмещающие два предыдущих класса.

На сегодняшний день наилучшие результаты достигаются с помощью моделей, основанных на параметрическом нейросетевом подходе.

Основной сложностью при создании качественных нейросетевых моделей синтеза речи является необходимость учитывать длительные зависимости, существующие в речевых сигналах. Введение в систему большого количества параметров, способных помочь в решении задач, одновременно с этим способно серьезно замедлить процессы как моделирования, так и работы итоговой системы синтеза речи. При этом также следует учитывать, что усложнение системы может приводить к замедлению её работы, что в большинстве практических приложений является недопустимым. Для человеческого уха снижение частоты дискретизации ниже 16 кГц является заметным на слух, что предъявляет к синтезаторам, работающим в режиме реального времени, требования по генерации не менее 16 тысяч отсчетов аудиосигнала в секунду.

Работу всех нейросетевых синтезаторов речи можно поделить на два этапа: подготовку необходимого набора характеристик, выровненных по времени (например, мел-спектрограммы, частота основного тона, различные лингвистические характеристики); преобразование подготовленных характеристик в итоговый аудио файл. При этом количество нейросетей, входящих в систему, их архитектура могут существенно отличаться в различных подходах. Единым для всех подходов остается требование по наличию обучающего корпуса аудиоданных. Разрабатываемые модели тестируются на речевых базах данных, среди которых наибольшей популярностью среди разработчиков пользуются корпуса “The LJ Speech Dataset” [4], VCTK Corpus [5], CMU_ARCTIC [6], а также множество аудиокниг.

3. Лингвистические ресурсы для обучения

Независимо от выбора архитектуры системы синтеза татарской речи, для её построения необходимо наличие корпуса речи. Источником для его создания могут служить аудиокниги, многодикторные корпуса. Однако небольшое количество имеющихся на татарском языке аудиокниг, а также желание получить максимально качественный синтезированный голос, привели нас к необходимости создания собственных речевых баз данных.

По состоянию на сентябрь 2019 года институт прикладной семиотики подготовил 4 речевых корпуса для синтезатора татарской речи. Все записи осуществлялись в профессиональных звукозаписывающих студиях, в качестве дикторов приглашались актеры татарского академического театра, также имеющие опыт в озвучивании книг и ведении теле- и радиопередач.

Основные характеристики корпусов представлены в таблице 1.

Таблица 1. Продолжительность корпусов для задачи синтеза татарской речи

Диктор	Исходный формат записи	Продолжительность записей	Итоговая продолжительность корпуса	Аннотация
Тавзих	22.05 кГц 16 бит моно	Хранится 1009 записей дифонов		
Алмаз	44.1кГц 16 бит стерео	12:08:22	5:30:02	+
Алсу	48 кГц 32 бит моно	7:30:32	6:17:31	+
Рамиль	44.1 кГц 24 бит стерео	22:58:48	16:53:59	–

Корпус «Тавзих» создавался для конкатенативного дифонного синтеза, поэтому его структура значительно отличается от других корпусов [7]. В качестве озвучиваемого текста выступал набор шаблонных фраз, состоящих из трех ритмических групп. Средняя ритмическая группа, в которой озвучивался целевой дифон, представляла трех - или четырехсложное синтетическое слово, начальная и конечная ритмические группы фразы при этом оставались

неизменными. Часть дифонной базы была размечена по периодам основного тона. Разметка проводилась отдельно для обеих полуфонем, входящих в дифон с указанием границы между фонемами. Для размеченных дифонов числовые отсчеты приведены в соответствующих текстовых файлах. Каждое значение представляет собой удвоенное число байтов от начала звукового файла.

В корпусах «Алмаз» и «Алсу» имеются дополнительные уровни аннотации: в аудиофайлах экспертами были вручную размечены все интонационные группы, отмечены заимствованные и акцентные слова, после чего для всего прочитанного текста была построена фонетическая транскрипция и автоматическими средствами определены границы произнесения каждой фонемы. Была реализована модель разметки корпуса, основные элементы которой можно представить следующим образом:

1. Уровень фонем: текущая фонема, две предшествующие, две последующие фонемы.

2. Уровень слогов: тип слога (V, VC, CV, CVC, VCC, CVCC); позиция фонемы в слоге; количество фонем в предыдущем, текущем, последующем слоге; номер текущего слога в слове; гласная в текущем слоге.

3. Уровень слов: часть речи, количество слогов для предыдущего, текущего, следующего слова; количество предшествующих и последующих слов во фразе.

4. Уровень фразы: количество слов/слогов в предыдущей, текущей, последующей фразе.

После аннотирования корпус хранится в формате массива троек файлов: wav (аудиофрагмент) – txt (текстовая аннотация аудиофрагмента) – lab (пофонемное аннотация аудиофрагмента). Содержимое txt-файла представляет собой текстовое представление озвученной фразы, а также следующие специальные символы: * – для обозначения интонации перечисления; |, – для обозначения незавершенной по интонации синтагмы; . – для обозначения завершенной по интонации синтагмы; ! – для восклицательной интонации; ? – для вопросительной интонации; ~ – для обозначения заимствованных слов.

Структура lab-файла построена на основе адаптированного под татарский язык файла аннотации, изначально разработанного для синтеза английской речи на базе скрытых Марковских моделей [8]. Итоговый формат файла аннотации:

t1 t2 p1^p2-p3+p4=p5@p6_p7 /A:a3 /B:b3@b4-b5&b6-b7|b16 /C:c3 /D:d1_d2 /E:e1+e2@e3+e4 /F: f1_f2 /H:h1=h2|h5 /J:j1,

где t1, t2 – временные отсчеты начала и окончания звучания каждой фонемы, полученные автоматически с помощью систем распознавания речи, специально построенных для каждого из голосов.

Описание остальных использованных обозначений приводится в таблице 2.

Таблица 2. Обозначения параметров, использованных в lab-файлах аннотации

Параметр	Описание	Параметр	Описание
p1	фонема за две до текущей	c3	число фонем в следующем слоге
p2	предыдущая фонема	d1	часть речи предыдущего слова
p3	текущая фонема	d2	число слогов в предыдущем слове
p4	следующая фонема	e1	часть речи текущего слова
p5	фонема через одну от текущей	e2	число слогов в текущем слове
p6	позиция текущей фонемы в текущем слоге	e3	позиция текущего слова в текущей фразе
p7	позиция текущей фонемы в текущем слоге, считая с конца слога	e4	позиция текущего слова в текущей фразе, начиная с конца фразы
a3	количество фонем в предыдущем слоге	f1	часть речи следующего слова
b3	количество фонем в текущем слоге	f2	число слогов в следующем слове
b4	позиция текущего слога в текущем слове	h1	число слогов в следующей фразе
b5	позиция текущего слога в текущем слове, начиная с конца слова	h2	число слов в следующей фразе

Продолжение таблицы 2

Параметр	Описание	Параметр	Описание
b6	позиция текущего слога в текущей фразе	h5	Тип синтагмы
b7	позиция текущего слога в текущей фразе, начиная с конца фразы	j1	Заимствованное слово или нет
b16	гласная в текущем слоге		

Корпуса «Алмаз» и «Алсу» для использования при обучении были приведены к формату аудио 16 кГц 16 бит моно, корпус «Рамиль» - 22.05 кГц 16 бит моно. Записи корпуса «Рамиль» также прошли дополнительную предобработку, позволившую улучшить качество итоговой системы синтеза, обработка включала следующие этапы:

- были отфильтрованы короткие (<1 секунды) и длинные (>11 секунд) записи;
- все аудио были нормализованы по громкости;
- были удалены начальные и конечные фрагменты тишины.

Различия в аннотации корпусов также включали формат представления соответствующих текстовых фрагментов: «Алмаз» и «Алсу» содержали фонетическую транскрипцию, «Рамиль» - транслитерацию на латиницу; во всех корпусах были отфильтрованы все знаки пунктуации, кроме знаков точки, запятой, восклицательного и вопросительного знаков.

4. Архитектура систем синтеза речи

Системы синтеза татарской речи были построены на базе следующих подходов:

- конкатенативный дифонный синтез [7];
- HTS-параметрический синтез [9];
- нейросетевой синтез на основе Merlin [10];
- нейросетевой синтез на основе DCTTS [11];
- нейросетевой синтез на основе Tacotron2 [12] / WaveGlow [13].

Система Merlin отличается от двух других нейросетевых подходов тем, что требует пофонемно аннотированного корпуса для

обучения нейросетей: первая нейросеть обучается предсказывать длительности произнесения для каждой фонемы, вторая – акустические характеристики каждой из фонем. Заключительным этапом работы системы является использование вокодера, преобразующего акустические характеристики и данные о длительностях фонем в результирующий речевой сигнал. В качестве вокодера для системы синтеза татарской речи был использован вокодер WORLD [14].

В ходе экспериментов с подходом Merlin были построены синтезаторы речи с использованием различных архитектур нейросетей:

- размер скрытых слоев: [1024, 1024, 1024, 1024, 1024, 1024], тип скрытых слоев: [‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’];
- размер скрытых слоев: [512, 512, 512, 512, 512, 512], тип скрытых слоев: [‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’];
- размер скрытых слоев: [1024, 1024, 1024, 1024, 256], тип скрытых слоев: [‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, LSTM’];
- размер скрытых слоев: [1024, 1024, 1024, 1024, 384], тип скрытых слоев: [‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, LSTM’];
- размер скрытых слоев: [1024, 1024, 1024, 1024, 256], тип скрытых слоев: [‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, ‘BLSTM’];
- размер скрытых слоев: [1024, 1024, 1024, 1024, 384], тип скрытых слоев: [‘TANH’, ‘TANH’, ‘TANH’, ‘TANH’, ‘BLSTM’],

где ‘TANH’ – слой с гиперболическим тангенсом в качестве функции активации нейронов, (B)LSTM – (двунаправленный) слой с нейронами, имеющими дополнительные блоки (gates), позволяющие запоминать (или забывать) длительный контекст.

Была произведена субъективная оценка качества звучания построенных систем синтеза речи, согласно которой в качестве итоговой была выбрана система с заключительным двунаправленным слоем 384 long-short term memory нейронов (BLSTM).

Обучение нейросетей происходило на базе подготовленных lab-файлов, входящих в состав корпусов татарской речи. Обучение системы на подкорпусе голоса “Алсу” длительностью 3 часа 41 минут заняло около 14 часов на видеокарте GTX 1070.

Большинство этапов построения систем синтеза речи на базе описанного подхода было автоматизировано. Наиболее трудозатратной является операция подготовки разметки в виде lab-файлов. Недостатком построенной системы является низкая скорость

генерации речи: с учетом всех необходимых процедур предобработки текста синтез одного предложения с частотой дискретизации 22.05 кГц на видеокарте GTX 1080 занимает около 16 секунд.

Подход, предложенный в DCTTS, является попыткой ускорить обучение нейросетевых моделей синтеза речи. Проблемы, связанные с трудностью обучения систем, аналогичных Merlin, заключаются в использовании рекуррентных слоев, которые, с одной стороны, позволяют моделировать “далекие” закономерности в речи, улучшая качество звучания, но с другой стороны, не позволяют производить распараллеливание вычислений. В DCTTS предполагается обучать две глубокие нейросети, состоящие из сверточных слоев (Deep Convolutional TTS), обучение которых происходит гораздо быстрее.

Основная идея данного метода заключается в обучении нейросетей задаче генерации мел-спектрограмм, на основе которых вокодер тренируется восстанавливать речевой сигнал. Отличительной особенностью является отсутствие необходимости разметки обучающего аудиокорпуса: на вход нейросети достаточно подавать массив аудиофрагментов и соответствующий им текст (или его транскрипцию). Основная сложность заключается в сложности подбора гиперпараметров, а некорректное обучение механизма внимания, входящего в состав нейросетей, может приводить к периодическим повторам или удалениям отдельных частей фраз из итогового аудиосигнала.

Заключительным на момент написания статьи подходом к синтезу речи, реализованным в институте прикладной семиотики АН РТ, является Tacotron2 + WaveGlow. Аналогично DCTTS, Tacotron2 пытается сформировать соответствия между векторами для входных символов (char embeddings) и мел-спектрограммами. Изначально мел-спектрограммы подавались на вход модели нейросетевого вокодера WaveNet. Однако мы использовали архитектуру с альтернативным вокодером WaveGlow, нейросеть которой учится восстанавливать высококачественный аудиосигнал по мел-спектрограмме. Обучение нейросети осуществляется с помощью единственной функции потерь, что делает процедуру обучения более простой и стабильной. Данная архитектура позволила сохранить качество синтеза, достигаемое с помощью нейросетей WaveNet, при значительном ускорении работы синтезатора с 0.11 кГц (для WaveNet) до 520 кГц [13]. Процесс обучения Tacotron2 и WaveGlow может вестись независимо друг от друга. Длительность

обучения зависит от выбора гиперпараметров, текущие версии нейросетей для системы татарского синтезатора были обучены суммарно за 16 дней на 8 видеокартах V100 32GB.

5. Заключение

В данной статье были представлены результаты по созданию систем синтеза татарской речи. Полученные результаты в виде аннотированных корпусов, алгоритмов и программных систем являются первыми для татарского языка. Записанные монодикторные корпуса, а также автоматическая и экспертная разметки позволили сформировать набор обучающих данных, необходимых для обучения систем нейросетевого синтеза речи (например, Merlin). Разработанные системы автоматического синтеза татарской речи позволяют вести работы по внедрению речевого человеко-машинного интерфейса на татарском языке. Система синтеза речи уже внедрена в веб-сервис русско-татарского машинного переводчика [15] и татарского синтезатора [16]. Дальнейшее развитие речевых технологий откроет перспективы совместного использования результатов исследований в области анализа текста на татарском языке, позволит создавать интеллектуальные системы помощи слабовидящим.

СПИСОК ЛИТЕРАТУРЫ

1. International Telecommunication Union. Method for the subjective assessment of intermediate quality levels of coding systems. URL: https://www.itu.int/rec/R-REC-BS.1534/en.

2. Moulines, E., Charpentier, F. “Pitch-sinchronous waveform processing techniques for text-to-speech synthesis using diphones”. In Speech Communication, 9 (5/6), 1990, P. 453–467.

3. Sagisaka, Y. ATR v-talk speech synthesis system. In Proc. ICSLP-92, 1992, Banff, Canada.

4. Keith Ito. The LJ Speech Dataset [Electronic resource]. URL: https://keithito.com/LJ-Speech-Dataset/.

5. English Multi-speaker Corpus for CSTR Voice Cloning Toolkit [Electronic resource]. URL: https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.

6. CMU_ARCTIC speech synthesis databases [Electronic resource]. URL: http://festvox.org/cmu_arctic/.

7. Ибрагимов Т. И. Синтезатор татарской речи. /Т. И. Ибрагимов, Ф. И. Салимов, Д. Ш. Сулейманов //Международная научно-практи-

ческая конференция «Система непрерывного образования инвалидов: опыт, проблемы, тенденции, решения». Казань: Академия управления ТИСБИ, 2006. – С. 91–97.

8. An example of context-dependent label format for HMM-based speech synthesis in English. URL: https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/F0parametrisation/hts_lab_format.pdf.

9. Speech human-machine interface for the Tatar language / A. Khusainov, A. Khusainova // Artificial Intelligence and Natural Language Conference. (Saint-Petersburg, 10-12 November 2016). Helsinki: FRUCT Oy, 2016. P. 60-65.

10. Zhizheng Wu, Oliver Watts, Simon King, «Merlin: An Open Source Neural Network Speech Synthesis System» in Proc. 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA, USA.

11. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. Hideyuki Tachibana, Katsuya Uenoyama, Shunsuke Aihara. https://arxiv.org/abs/1710.08969.

12. Jonathan Shen at al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. arXiv:1712.05884v2.

13. Ryan Prenger, Rafael Valle, Bryan Catanzaro. WaveGlow: A Flow-based Generative Network for Speech Synthesis. arXiv:1811.00002.

14. M. Morise, F. Yokomori, and K. Ozawa. “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, 2016.

15. Русско-татарский переводчик TatSoft.Переводчик. URL: https://translate.tatar.

16. Синтезатор татарской речи TatSoft.Синтез. URL: https://speech.tatar.

УДК 81.27

ON SOCIOLINGUISTIC RESEARCHES IN ALTAI REPUBLIC

A. E. Chumakaev
Research Institute of Altaistics. S. S. Surazakova
Gorno-Altaysk, Altai Republic, Russia
newchae@mail.ru

In this work the sociolinguistic research in the field of the Altai language is given the short review. The main attention is paid to the «Monitoring of a Language Situation in Altai Republic» project which is carried out by the *Altaic Scientific Research Institute* named after *S. S. Surazakov* since 2016. The main results received during implementation of the above-named project are presented.

Keywords: Altai language; sociolinguistics; language situation; language policy; mother tongue.

О СОЦИОЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ В РЕСПУБЛИКЕ АЛТАЙ

А. Э. Чумакаев
НИИ алтаистики им. С. С. Суразакова
г. Горно-Алтайск, Республика Алтай, Россия
newchae@mail.ru

В данной работе дается краткий обзор социолингвистических исследований в области алтайского языка. Основное внимание уделено проекту «Мониторинг языковой ситуации в Республике Алтай», который осуществляется НИИ алтаистики им. С. С. Суразакова с 2016 г. Представлены основные результаты, полученные в ходе реализации вышеназванного проекта.

Ключевые слова: алтайский язык, социолингвистика, языковая ситуация, языковая политика, родной язык.

1. Введение

Социолингвистические исследования в Республике Алтай активно стали проводиться с начала 2000-х гг. Эти исследования в основном осуществлялись среди коренных малочисленных народов республики – тубаларов, чалканцев, кумандинцев [1, 2, 3, 4].

Ученые-лингвисты научно-исследовательской группы алтайского языка НИИ алтаистики им. С. С. Суразакова также прини-

мали и принимают участие в социолингвистических исследованиях. В 2008–2009 гг. группа языковедов института работала по Ведомственной целевой программе «Языковая политика и языковая ситуация в Республике Алтай». В рамках данной программы было проведено анкетирование городского и сельского населения Республики Алтай, осуществлены социолингвистические экспедиции в районы республики. Впервые функционирование алтайского языка было исследовано в самых разных сферах: образовательные организации, семейно-бытовое общение, СМИ и т.д. Результаты работы по названной программе опубликованы в сборнике «Языковая политика и языковая ситуация в Республике Алтай» [5].

В вышеуказанный сборник вошли материалы также по языковой ситуации у теленгитов, которые тоже включены в перечень коренных малочисленных народов республики. В 2012 г. лингвистами института была проведена очередная социолингвистическая экспедиция в места компактного проживания теленгитов – Кош-Агачский и Улаганский районы Республики Алтай [6].

В настоящее время языковая ситуация у тех или иных этнических групп алтайцев, а также у алтайского этноса в целом продолжает привлекать внимание ученых [7, 8].

Проект «Мониторинг языковой ситуации в Республике Алтай»

С 2016 года в НИИ алтаистики им. С. С. Суразакова реализуется проект «Мониторинг языковой ситуации в Республике Алтай». Данный проект является частью той работы, которая выполняется в республике в соответствии с пунктом 2 «а» перечня поручений Президента Российской Федерации по итогам совместного заседания Совета при Президенте Российской Федерации по межнациональным отношениям и Совета при Президенте Российской Федерации по русскому языку 19 мая 2015 г. [9] и поручением Правительства Российской Федерации от 15 июля 2015 г. [10], согласно которым проводится ежегодный мониторинг состояния и развития языков народов Российской Федерации. Указанный мониторинг направлен на выявление состояния языков народов России и выработку комплекса мер по их поддержке и развитию, а также разработку эффективных механизмов реализации государственной языковой политики с учетом конституционного статуса

языков. Мониторинг осуществляется на основе *Методики сбора, обработки и анализа информации о состоянии и развитии языков народов России*.

Проект «Мониторинг языковой ситуации в Республике Алтай» реализуется в институте путем организации социолингвистических экспедиций в районы Республики Алтай. В 2016 г. лингвисты института проводили социолингвистическое анкетирование в Кош-Агачском, Улаганском, Онгудайском и Шебалинском районах республики [11, 12]. Для проведения опроса была разработана анкета, включающая 22 вопроса. В числе прочих вопросов информантам предлагалось также выразить свое отношение по поводу обязательного изучения алтайского языка детьми-алтайцами и детьми не алтайской национальности, возможного преподавания школьных предметов на алтайском языке. Всего было опрошено 719 информантов.

По итогам социолингвистического исследования были получены следующие результаты:

1. Алтайский язык в качестве родного языка указали 605 информантов (97 %) из 623 чел. (86,6 %), идентифицирующих себя как алтайцы.

2. За обязательное изучение детьми-алтайцами алтайского языка высказались 690 информантов (96 %), 14 чел. – против (2 %), 15 – воздержались от ответа.

3. С мнением об обязательном изучении алтайского языка детьми не алтайской национальности согласились 243 чел. (34 %), не согласны – 18 чел. (2,5 %), за изучение по желанию высказались 437 чел. (61 %).

4. Родными языками (алтайским, русским) владеют свободно 655 чел. (91 %).

5. Русским языком свободно владеют 708 информантов (98 %).

6. Мнение о необходимости преподавания школьных предметов на алтайском языке поддерживают 243 чел. (38 %).

Как показывают результаты данного анкетирования, за обязательное изучение алтайского языка детьми-алтайцами выступают 96 % респондентов. Обязательное изучение алтайского языка детьми не алтайской национальности поддерживают 34 % информантов, 61 % не поддерживает, считая, что для данной категории детей изучение алтайского языка должно быть по желанию.

Что касается мнения о необходимости преподавания школьных предметов на алтайском языке, то это мнение поддержива-

ют 38 % респондентов, 62 % выступают против данного мнения, объясняя это тем, что преподавание школьных предметов на алтайском языке может негативно отразиться на уровне владения русским языком, а это, в свою очередь, может создать трудности при дальнейшем поступлении учащихся в вузы и другие образовательные организации.

В 2018 г. ученые-лингвисты научно-исследовательской группы алтайского языка НИИ алтаистики им. С. С. Суразакова провели социолингвистические исследования в селах Улаганского, Кош-Агачского, Усть-Канского и Усть-Коксинского районов Республики Алтай. В Улаганском и Кош-Агачском районах для проведения анкетирования были выбраны те села, которые не были охвачены мониторингом в 2016 году. В Усть-Канском и Усть-Коксинском районах анкетирование проводилось впервые. Всего в перечисленных районах было опрошено 989 респондентов. Получены следующие результаты:

1. Алтайский язык в качестве родного языка указали 879 информантов (88,8 %) из 877 чел. (86,6 %), идентифицирующих себя как алтайцы, и 41 респондента (4,1 %), относящего себя к теленгитам.

2. За обязательное изучение детьми-алтайцами алтайского языка высказались 904 информанта (91,4 %), 11 чел. – против (1,1 %), 46 – воздержались от ответа.

3. С мнением об обязательном изучении алтайского языка детьми не алтайской национальности согласились 342 чел. (34,5 %), не согласны – 90 чел. (9,1 %), за изучение по желанию высказались 469 чел. (47,4 %).

4. Родными языками (алтайским, русским) владеют свободно 838 чел. (84,7 %).

5. Русским языком свободно владеют 915 информантов (92,5 %).

6. Мнение о необходимости преподавания школьных предметов на алтайском языке поддерживают 363 чел. (36,7 %).

Результаты проведенного анкетирования показывают, что в исследованных в 2018 году районах за обязательное изучение алтайского языка детьми-алтайцами выступают 91,4 % респондентов. Обязательное изучение алтайского языка детьми не алтайской национальности поддерживают 34,5 % информантов, 47,4 % не поддерживает, склоняясь к изучение детьми данной категории алтайского языка по желанию.

Мнения о необходимости преподавания школьных предметов на алтайском языке поддерживают 36 % респондентов, 37,6 % выступают против, 20,1 % затруднились ответить. Те, кто были против или затруднились ответить, также считают, что преподавание школьных предметов на алтайском языке может негативно отразиться на уровне владения русским языком.

В 2019 г. сотрудниками института в рамках проекта «Мониторинг языковой ситуации в Республике Алтай» осуществлены социолингвистические экспедиции в Майминский и Чемальский районы республики. Опрос-анкета подверглась некоторой корректировке. Так, например, вопрос *Необходимо ли обязательное изучение алтайского языка в школе детьми-алтайцами?* был заменен на *Вы согласны с мнением о необходимости изучения алтайского языка в школе детьми алтайской национальности?* Материал по экспедициям 2019 г. не включен в данную статью.

Таким образом, исследования, проведенные лингвистами НИИ алтаистики им. С. С. Суразакова в 2016 и 2018 гг., показывают, что высок процент респондентов, высказывающихся за обязательное изучение алтайского языка детьми-алтайцами. Это свидетельствует о том, что алтайское население осознает важность изучения родного языка для сохранения алтайского этноса. Что касается детей не алтайской национальности, то для них, по мнению большинства опрошенных, изучение алтайского языка должно быть по желанию. Анкетирование выявило также хороший показатель по уровню владения родным алтайским языком. Большинство респондентов не поддерживает мнение о необходимости преподавания школьных предметов, кроме алтайского языка и литературы, на алтайском языке, мотивируя это тем, что уровень владения русским языком у детей-алтайцев будет недостаточным для дальнейшего успешного поступления учащихся в вузы и другие образовательные организации.

В целом, следует отметить, что в большинстве исследованных районов, несмотря на определенные сложности, связанные, например, с малым количеством часов, выделяемых для изучения алтайского языка и литературы, ограниченностью сфер использования алтайского языка, сохраняется языковая ситуация, способствующая дальнейшему функционированию алтайского языка.

Заключение

Анализ литературы по социолингвистическим исследованиям в Республике Алтай показал, что основная часть указанных исследований посвящена языковой ситуации у коренных малочисленных народов республики – тубаларов, чалканцев, кумандинцев и теленгитов.

В последние годы в связи с реализацией в НИИ алтаистики им. С. С. Суразакова проекта «Мониторинг языковой ситуации в Республике Алтай» исследования по языковой ситуации в регионе приобрели регулярный характер. Это позволило охватить анкетированием значительное количество населенных пунктов республики и получить новые данные о языковой ситуации у алтайского этноса в целом. Реализация названного проекта в совокупности с другими мероприятиями по мониторингу состояния и развития алтайского языка будет способствовать отслеживанию функционирования алтайского языка в различных сферах, выявлять и обозначать проблемы, намечать пути их решения.

ЛИТЕРАТУРА

1. Сарбашева С. Б. Современная языковая ситуация у тубинцев // Социальные процессы в современной Западной Сибири: философские, политологические, культурологические аспекты. – Горно-Алтайск, 2000. – С. 166–173.

2. Озонова А. А., Николина Е. В., Кокошникова О. Ю., Тазранова А. Р. Социолингвистическая ситуация у тубаларов и чалканцев // Языки коренных народов Сибири. Вып. 7. Часть 1. Экспедиционные материалы. – Новосибирск: изд-во НГУ, 2003. – С. 3–9.

3. Уртегешев Н. С. Социолого-лингвистическая ситуация у кумандинцев // Вестник Казахского национального университета им. аль-Фараби. – 2005. – N 5 (87). – С. 105–107.

4. Бельгибаев Е. А., Назаров И. И., Николаев В. В. Современные этноязыковые процессы у северных алтайцев // Этнография Алтая и сопредельных территорий: материалы Международной научно-практической конференции. – Барнаул: Изд-во Барнаул. гос. пед. ун-та, 2005. – Вып. 6. – С. 155–157.

5. Языковая ситуация и языковая политика в Республике Алтай: сборник научных статей / Отв. ред. Н. Н. Тыдыкова. – Горно-Алтайск: ГНУ РА «НИИ алтаистики им. С. С. Суразакова»; Горно-Алтайская республиканская типография, 2010. – 228 с.

6. Чумакаев А. Э. Языковая ситуация в Кош-Агачском и Улаганском районах Республики Алтай (по данным социолингвистической экспедиции) // Сибирь в исторической перспективе и проблемы сохранения народов и культур: тезисы докладов Всероссийской конференции, приуроченной к Году истории России. – Новосибирск, 2012. – С. 100.

7. Чемчиева А. П. Языковая ситуация в среде коренных малочисленных народов города Горно-Алтайска // Исторические, философские, политические и юридические науки, культурология и искусствоведение. Вопросы теории и практики. – Тамбов: Грамота, 2017. – № 12(86). В 5-ти ч. – Ч. 5. – С. 237–240.

8. Екеев Н. В. Современная этноязыковая ситуация в Республике Алтай (по материалам Всероссийской переписи населения 2010 г.) // Актуальные вопросы алтайского языкознания: проблемы развития литературного языка, совершенствование современной орфографии: Материалы Всероссийской научно-практической конференции, посвященной 115-летию Т. М. Тощаковой / БНУ РА «НИИ алтаистики им. С. С. Суразакова»; Редакционная коллегия: М. С. Дедина, Н. В. Екеев, А. Н. Майзина, А. Э. Чумакаев (отв. ред.). – Горно-Алтайск: Горно-Алтайская типография, 2017. – 336 с. 114–120.

9. Перечень поручений по итогам совместного заседания Совета по межнациональным отношениям и Совета по русскому языку [Электронный ресурс] http://www.kremlin.ru/acts/assignments/orders/49877) (дата обращения: 24.08.2019 г.).

10. Об обеспечении выполнения поручений Президента России по итогам совместного заседания Совета по межнациональным отношениям и Совета по русскому языку [Электронный ресурс] http://government.ru/orders/selection/404/18910/ (дата обращения: 24.08.2019 г.).

11. Чумакаев А. Э. О результатах реализации проекта «Мониторинг языковой ситуации в Республике Алтай» в 2016 г. // Актуальные вопросы алтайского языкознания: проблемы развития литературного языка, совершенствование современной орфографии: Материалы Всероссийской научно-практической конференции, посвященной 115-летию Т. М. Тощаковой. Редакционная коллегия: М. С. Дедина, А. Н. Майзина, А. Э. Чумакаев (отв. ред.). – Горно-Алтайск: Горно-Алтайская типография, 2017. – С. 152–158.

12. О результатах реализации проекта «Мониторинг языковой ситуации в Республике Алтай» // История повседневности населения Западной Сибири и сопредельных регионов как форма цивилизационной идентичности Евразии [Электронный ресурс]: материалы Всероссийской с международным участием научной конференции (г. Бийск, 21–23 июня 2018 г.) / Отв. ред. А. В. Литягина. – Бийск: АГГПУ им. В. М. Шукшина, 2018. – С. 131–134.

УДК 81’33

CONSTRUCTING MINIMALIST MONOLINGUAL DICTIONARIES FOR TURKIC LANGUAGES

Barış Can Erkoç, Ercan Solak
Işık University, Meşrutiyet mah. Şile, İstanbul 34980, Turkey
ercan.solak@isikun.edu.tr

Making monolingual dictionary has a long tradition riddled with changing and often incoherent approaches. Each dictionary reflects the peculiar methods used by its creators and maintainers. From the coverage of lemmas to the inclusion of semantic relationships among them, each dictionary has its own set of assumptions, which are often implicit and change as the dictionary evolves. Turkic dictionaries are no different but with additional set of challenges, due to mainly the agglutinative nature of their morphologies.

Having a consistent lexicon is crucial in every stage of computational linguistics. In this paper, we provide the details of LexEdit, an open-source tool we created to manipulate the lexicon of Turkic languages. Using LexEdit, multiple annotators can independently edit the lexicon to inspect, add and remove lemmas, or collapse multiple senses into one, and write sense-definitions with the help of templates.

We also provide the details of DicJSON, a simple file format we propose for storing lexicons for the permanent data in LexEdit. LexEdit, thus, is a novel modular application with a service and delegation integration to morphological analyzers which can be configured at run-time.

Keywords: lexicography; LexEdit; dictionary; lexicon; Turkic Languages.

СОЗДАНИЕ МИНИМАЛЬНЫХ ОДНОЯЗЫЧНЫХ СЛОВАРЕЙ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ

Барыш Джан Эркоч, Эрджан Солак
Университет Ышык, Стамбул, Турция
ercan.solak@isikun.edu.tr

Создание одноязычного словаря имеет давнюю традицию пронизанную изменяющимися и часто непоследовательными подходами. Каждый словарь отражает специфические методы, используемые его авторами и составителями. От описания самих лемм до семантических отношений между ними, каждый словарь имеет свой набор характеристик, которые часто имплицитны и изменяются по мере развития словаря. Тюркские словари ничем не отличаются, кроме как допол-

нительным набором сложностей, в основном из-за агглютинативного характера их морфологии.

Наличие последовательной лексики имеет решающее значение на каждом этапе компьютерной лингвистики. В данной статье мы подробно описываем инструмент LexEdit, находящийся в открытом доступе, который мы создали для работы с лексической базой/ словарным лексиконом тюркских языков. Используя LexEdit, разные авторы могут независимо друг от друга редактировать лексическую базу, проверяя, добавляя или удаляя леммы, или объединяя множество смыслов в один, или записывая определения смыслов с помощью шаблонов.

Мы также предоставляем подробную информацию о DicJSON, файле простого формата, который мы предлагаем для хранения лексических баз в постоянной базе данных LexEdit. Таким образом, LexEdit – это новое модульное приложение с функциями и возможностями включения в морфологические анализаторы, которые можно настраивать во время выполнения.

Ключевые слова: лексикография; LexEdit; словарь; лексическая база; тюркские языки.

1. Introduction

Having a reliable monolingual electronic dictionary is essential for many tasks in natural language processing (NLP). At the most basic level, a comprehensive dictionary provides the list of root lexicon which is used in the initial processing steps such as tokenization, stemming and morphological analysis. At the other end of the spectrum, senses for headwords constitute the reference data for semantic analyses. When the dictionary provides further features such as part-of-speech (POS) tags, domains and usage frames, these become useful when building parsers of syntactic relations.

Lexicography, the discipline of writing dictionaries, is a craft as much as a scientific endeavor. As a descriptive snapshot of a language and its actual use, a dictionary contains its dynamism as well as its hazy boundaries. It is dynamic because the languages evolve albeit at different paces depending on the social context. A dictionary is not a mathematical construction because the actual language has many fluid boundaries and peculiarities living along those boundaries.

Dictionary making has a long history. First dictionaries were compiled by the people of the ancient city of Ebla around 2300 BC (Dumper, Stanley, 2007). Since then, evolutions of dictionaries have become mostly path dependent as each new dictionary in a language used the existing dictionaries as leverage for the simple reason of saving as much manual labor as possible. This evolution is particularly

emphasized in the development of a single dictionary through its different editions. Each new edition begins with the previous one and modifies only a small portion of it by adding, removing a small subset of its headwords and their definitions. On rare occasions, we see the constructions of a dictionary from scratch. Considering the manual expert effort that goes into making dictionaries, it is no surprise that we see only a few comprehensive dictionaries. For example, the Contemporary Turkish Dictionary (CTD) published and maintained by the Turkish Language Association is the most common reference dictionary in Turkish.

Until recently dictionaries have been available only in print form. Their electronic versions are mostly direct ports of print versions augmented with the ease of co-referencing and searching. Hence, when an electronic dictionary is used in NLP research, the artefacts of its evolution are reflected in the pipeline.

If the dictionary has evolved to contain redundant entries, the spurious lexicon may end up creating a bias in the results. As an extreme example, consider a lexicon that contains infected forms as root lexemes. This will result in more analyses in the morphological analysis step, thus increasing the ambiguity in parsing.

Another artefact is the fine granularity of the senses of a single headword. For example, lexicographers might designate a distinct sense for a metaphorical use of a lexeme. Another equally common reason for redundant senses is the identification of a common context as indicating a new sense when, in fact, it is not. Such sense inflation makes the data sparse in sense disambiguation.

In this paper, we propose an approach to edit existing dictionaries by pruning the list of headwords and grouping apparently distinct yet actually same senses given in the definition of a headword. In order to streamline the edit, we introduce DicJSON, a dictionary structure we propose and LexEdit, an open-source tool we created to make it easier to visually and consistently edit dictionaries.

The rest of the paper is organized as follows. Section 2 gives a very brief account of making dictionaries as well as comparison of corpus based dictionaries with traditional ones and states the problem. Section 3 proposes a simple method to improve effectiveness of an existing online dictionary for its use in NLP. Section 4 and 5 present a dictionary editor, LexEdit, and data structure it uses. Finally, Section 6 gives some concluding remarks.

2. Making dictionaries

Contemporary dictionaries are based on the premise that a dictionary must reflect the actual uses of language in different contexts such as official documents, speech, news, web and social media. In that respect, dictionaries are descriptive snapshots of languages with all their dynamism and peculiarities. A natural implication of this approach is that, dictionary makers should base their efforts on representative corpora rather than relying solely on the intuition and the expertise of the lexicographers on the particular language. Normally, we need to have very large amount of text to have adequate representation for the rarer words and their rarer usages, (Atkins and Rundell, 2008). While a pure representative corpora is not possible, current corpora for lexicography aim for balanced one with a fair amount of representation that includes most usage patterns. Hence, modern lexicographers employ large corpora collected from various sources.

The first corpus based dictionary was Collins COBUILD English Language Dictionary (Sinclair, Hanks, Fox, Moon, Stock, 1987). Since then other corpus based dictionaries have been created for different languages, for German (Storjohann, 2017), French (Lonsdale, Le Bras, 2009) and Spanish (Davies, 2006). The availability of huge amounts of digital text on the web as well as the digitization of books and magazines made the construction of raw corpora much easier.

On the other hand, traditional dictionaries relied mostly on the intuition of a group of lexicographers and the corpora they used were quite small and generally comprised of literary texts. Also, when a dictionary evolved over a long period of time, it tends to accumulate inconsistent approaches brought in by different editorial teams. For example, the way the definitions are written out might have changed, contextual uses might be introduced as new senses, redundant headwords might have been added.

Despite these shortcomings, electronic versions of traditional dictionaries are all we have got when working in the NLP pipeline. In the sequel, we describe one way these traditional dictionaries might be edited to become more consistent and therefore useful for NLP practitioners.

3. Editing Dictionaries

In this section, we describe a simple yet effective method to leverage the existing traditional electronic dictionaries in order to

obtain a canonical, consistent dictionary that can be used in the NLP practice.

We consider two main editing categories. The first category is the revision of the list of headwords and the second is the revision of the senses given for a particular headword.

In the revision of the headword list, the annotator may mark a headword to be removed from the list. This happens, for example, when an inflected form of a root headword has been included in the dictionary as a separate entry. In this particular case, the annotator reviews the connection between the inflected form and the root headword and marks the inflected form to be removed.

Another common case of headword removal might occur when the semantics of a derived wordform can be unambiguously predicted from its constituent parts. In this case, the annotator might remove the derived wordform from the list. Naturally, the annotator must make sure that the derived wordform has no other sense that cannot be directly inferred from the roots and the derivational morphemes. For example in CDT, meaning of headword «aceleci» (hasty) can be predicted from the semantics of its root «acele» (haste) and the semantics of the derivational suffix -CI which designates a person who performs (produces or sells) the action entailed in the semantics of the root it is attached to.

Additions to the list might also occur in the rare cases where a derived wordform is in the list but its root lemma is missing. In that case, the root lemma is added to the list as a new headword. Of course, this also requires re-evaluating the derived wordform for removal and writing a definition for the inserted root headword.

For the revision of the senses of a lemma, the most common operation is grouping of the senses, thus decreasing the granularity in the definitions. This happens when different senses actually correspond to the uses of the same sense in different contexts. For example in CDT, lemma «yüz» have four different homonyms which roughly correspond to the semantics of «hundred», «face», «swim» and «skinning». The «face» semantics contains eleven different senses and three of them are listed as

- Surface
- Side
- Each surface of an object that facing outwards

Obviously, these three senses can be grouped under the general sense of «surface».

In the next section, we introduce LexEdit, an open-source tool that we developed specifically to perform these two kinds of revisions on an existing dictionary.

4. LexEdit

Editing a dictionary as described in the previous section can be made easier and faster by using a visual editor with the familiar interactions through mouse and keyboards shortcuts. Also, while editing, an annotator might need to refer to different parts of the dictionary, search for particular lemmas and look up possible morphological analyses of wordforms. LexEdit provides these utilities in a compact user interface which is given in Figure 1.

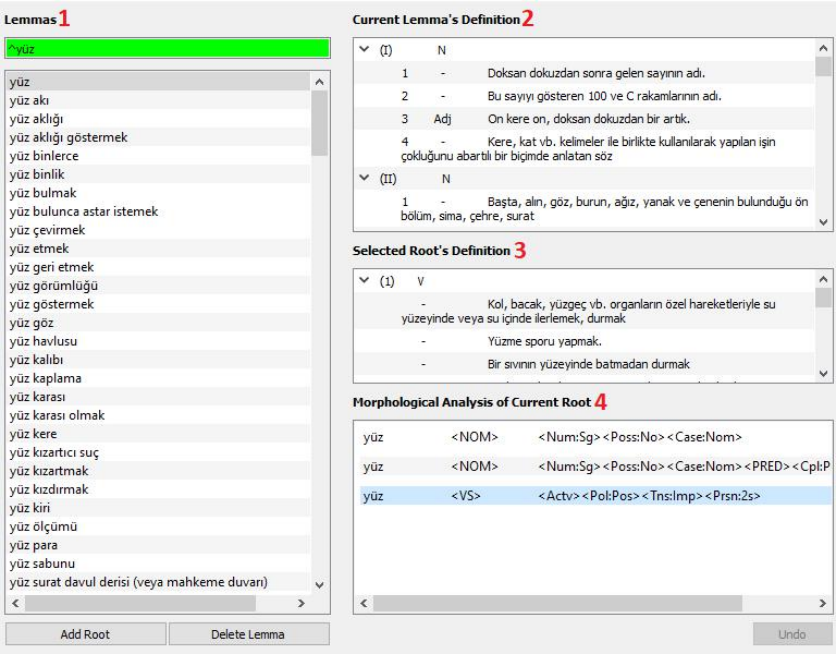

Fig. 1. The user interface of LexEdit

LexEdit works with a master dictionary stored in DicJSON format which is explained in the next section. All the edits done by the annotators are stored in separate folders, one for each annotator. This way, all annotators refer to the same master dictionary yet can

work independently. A master editor can consolidate the edits of all annotators to generate a new and revised dictionary.

LexEdit assumes that master dictionary and the edit histories of different annotators are stored in a common location (e.g. in a Dropbox or Google Drive folder) accessible by all of the annotators.

Pane 1 (numbered in red) in Figure 1 is a searchable list of headwords in the dictionary. The search strings can be literals or regular expressions and the search results are displayed in real-time.

Pane 2 in Figure 1 displays the homonyms and the sense definitions in each homonym as a numbered list.

Pane 4 displays the morphological analyses of the headword selected in Pane 1. The analyses are retrieved through a web service which can be configured at run-time. The web service is a simple interface to a morphological analyzer that returns the list of analyses as in a JSON response. When the annotator selects a particular analysis in Pane 4, the definition of its root lemma is retrieved from the dictionary and is displayed in Pane 3.

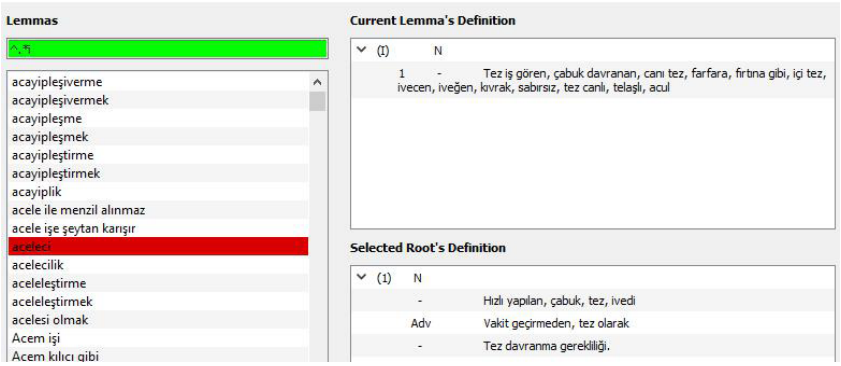

Fig. 2 Annotator deciding to remove a redundant headword

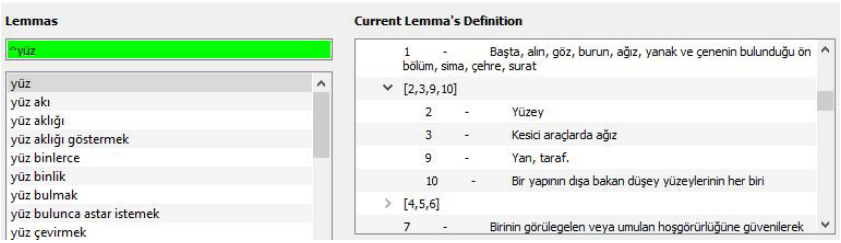

Fig. 3. Grouping of senses

By comparing the senses given in Panes 2 and 3, the annotator judges the whether the headword is redundant or not. In the example in Figure 2, the annotator judges that the senses of the lemma «aceleci» can be unambiguously inferred from the senses of its root «acele» and thus marks the headord «aceleci» for removal.

For the grouping of senses of a headword selected in Pane 1, only Pane 2 is used. The annotator can use simple drag and drop operations to group or de-group senses. In the example in Figure 3, the annotator has grouped the senses 2, 3, 9, 10 and the senses 4, 5, 6 of the lemma «yüz».

5. DicJSON format

The underlying file format of the dictionaries used in LexEdit is a custom JSON file which we named DicJSON. Below is a segment from the DicJSON file of CDT.


```
{

    "homonyms": [
      {
        "senses": [
          {
            "inferred_pos": null,
            "gloss": "Yüzmesini sağlamak veya yüzme işini yaptırmak",
            "pos": "(-de)",
            "example": "Burada değil, karşı kıyıda yüzdürüyorlar."
          },
          {
            "inferred_pos": "V",
            "gloss": "Batmış veya oturmuş tekneyi suyun yüzüne çıkarıp yüzer duruma getirmek",
            "pos": "(-i)",
            "example": "Batık gemileri yüzdürdüler."
          }
        ],
        "lemma": "yüzdürmek",
        "default_pos": "(-de)"
      },
      {
        "senses": [
          {
            "inferred_pos": "V",
            "gloss": "Derisini çıkarttırmak, derisini soydurtmak.",
            "pos": "(-i)",
            "example": null
          }
        ],
        "lemma": "yüzdürmek",
        "default_pos": "(-i)"
      }
    ],
    "lemma": "yüzdürmek"

  }
```


Each headword is represented as a key in the DicJSON. The value is a list of its homonyms. Each homonym is a key-value dictionary that stores the POS tag(s), list of senses and example sentences for each sense when it is provided.

6. Conclusion

In many natural language processing tasks, having a consistent and minimal dictionary is crucial. However, existing electronic dictionaries are usually the digital versions of print dictionaries which are the products of a historical accumulation of often inconsistent and redundant additions. In this paper, we described LexEdit, an open source tool we created to easily edit existing dictionaries to make them more suitable for NLP tasks. This effort brings an interim solution when we lack the means to undertake the immense task of constructing a corpus based electronic dictionary from scratch.

REFERENCES

Dumper, M., & Stanley, B. E. (2007). Cities of the Middle East and North Africa: A Historical

Encyclopedia(pp. 141–142). Santa Barbara, Calif: ABC-CLIO.

Atkins, B. T. S., & Rundell, M. (2008). The Oxford Guide to Practical Lexicography (pp. 61–69). Oxford:

Oxford University Press.

Sinclair, J.M., Hanks, P., Fox, G., Moon, R., Stock, P. (eds.) (1987). Collins COBUILD English Language Dictionary. London: Collins ELT.

Storjohann, P. (2017). elexiko: A Corpus-Based Monolingual German Dictionary. HERMES – Journal of Language and Communication in Business, 18(34), (pp. 55–82).

Lonsdale, D., & Le Bras, Y. (2009). A frequency dictionary of French: Core vocabulary for learners.

Abingdon [England]: Routledge.

Davies, M. (2006). A frequency dictionary of modern Spanish: Core vocabulary for learners. New York: Routledge.

СОДЕРЖАНИЕ

СЕДЬМАЯ
МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ
ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ
«TURKLANG 2019»

Труды конференции

В авторской редакции

Подписано в печать 30.10.2019 г.
Формат 60×84 $^{1}/_{16}$. Бумага офсетная.
Гарнитура «Таймс». Усл.-печ. л. 20,4.
Тираж 100 экз. Заказ 1019.

Отпечатано в ООО «Фолиант»
420111 г. Казань, ул. Профсоюзная, д. 17в

ISBN 978-5-9690-0548-8