TurkLang.2018

VI International Conference on Computer Processing of Turkic Languages

Tashkent, Uzbekistan
18-20 October, 2018

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

АКАДЕМИЯ НАУК РЕСПУБЛИКИ ТАТАРСТАН

ИНСТИТУТ ПРИКЛАДНОЙ СЕМИОТИКИ

ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ Л. Н. ГУМИЛЁВА

МИНИСТЕРСТВА ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН

НИИ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

VI МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ

«TURKLANG-2018»

(труды конференции)

ТАШКЕНТ
ИЗДАТЕЛЬСКО-ПОЛИГРАФИЧЕСКИЙ
ДОМ «NAVOIY UNIVERSITETI»
2018

С 23
УДК 811.512.1 (063)
ББК 81.2 ТЮРК (я43)

Шестая Международная конференция по компьютерной обработке тюркских языков «TurkLang-2018». (Труды конференции) –Ташкент: Издательско-полиграфический дом «NAVOIY UNIVERSITETI», 2018. – 390 с.

Научные редакторы: **PhD Н.З. Абдурахмонова;**
к.т.н. А.Р. Гатиатуллин

Сборник содержит материалы Шестой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2018» (Ташкент, Узбекистан, 18–20 октября 2018 г.)

Данная публикация предназначена для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной лингвистики и ее приложений.

(Издается в авторской редакции)

ISBN 978-9943-5635-1-3

Издание рекомендовано к публикации Постановлением №8 Ученого совета Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои от 28 декабря 2018 года.

ПРЕДИСЛОВИЕ

В этом году на базе Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои прошла уже шестая международная конференция по компьютерной обработке тюркских языков TurkLang-2018. Предыдущие конференции прошли в Астане (2013), Стамбуле (2014), Казани (2015, 2017), Бишкеке (2016). География проведения, представленные труды и состав участников конференции подтверждают, что в настоящее время тематика конференции продолжает оставаться весьма актуальной.

Целью серии международных конференций TurkLang является создание пространства совместных компьютерных лингвистических исследований для тюркских языков. На конференции представляются качественно новые результаты, связанные с разработкой компьютерных лингвистических приложений для тюркских языков. Как известно, тюркские языки обладают сложной и самобытной грамматической и семантической системами, поэтому простой перенос решений, полученных на материале других языков (в том числе английского и русского) порой практически не возможен.

Мы надеемся, что проведение этой конференции на базе Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои послужит очередным толчком для развития в стенах нашего университета направления компьютерной лингвистики.

В сборник трудов включены статьи участников VI Международной конференции по компьютерной обработке тюркских языков «TurkLang-2018» (Ташкент, Узбекистан, 18–20 октября 2018 г.). Участниками конференции, учеными и специалистами из Узбекистана, Казахстана, Кыргызстана, Турции, Азербайджана, России (Татарстан, Башкортостан, Москва, Саха (Якутия), Чувашия, Тува, Крым и др.), Португалии, были представлены доклады, посвященные актуальным проблемам компьютерной и когнитивной лингвистики в плане разрешения их в контексте тюркских языков. В ходе конференции активно и плодотворно обсуждались вопросы разработки формальных лингвистических моделей, электронных корпусов, систем машинного перевода, речевых технологий, а также проблемы, связанные с функционированием национальных языков в Интернет-технологиях. Участники отметили конструктивность обсуждения на секциях и круглых столах проблем разработки общей терминологии, общей системы обозначений лексико-грамматических категорий, использования для реализации своих национальных проектов аналогичных подходов, методов и технологий, особенно с учетом близости тюркских языков практически во всех компонентах, включая лексику, морфологию, синтаксис и семантику.

Тематика конференции находится в постоянном развитии. В список новых обсуждаемых тем в 2018 году включен совместный проект по созданию компьютерной онтологии тюркской грамматики. Участниками проекта являются ученые Евразийского национального университета имени Л.Н.Гумилева, Кыргызского государственного технического университета имени И.Раззакова, Стамбульского технического университета, Академии наук Республики Татарстан, а также нашего университета. Как показывает обсуждение реализации проекта и задач по этой проблематике и путей их решения, унификация систем понятий и терминов не является тривиальной практической задачей и требует теоретического пересмотра многих традиционных грамматических описаний.

Ректор Ташкентского государственного университета
узбекского языка и литературы имени Алишера Навои,
профессор Ш. С. Сирожиддинов

ПРОГРАММНЫЙ КОМИТЕТ

1. Сирожиддинов Шухрат Самариддинович (Ташкент, Узбекистан) – председатель
2. Сулейманов Джавдет Шевкетович (Казань, Татарстан, Россия) – сопредседатель
3. Шарипбаев Алтынбек Амирович (Астана, Казахстан) – сопредседатель
4. Ешреф Адалы (Стамбул, Турция)
5. Алтынбек Гулила (Урумчи, Китай)
6. Гатиатуллин Айрат Рафизович (Казань, Татарстан, Россия)
7. Дыбо Анна Владимировна (Москва, Россия)
8. Желтов Валериан Павлович (Чебоксары, Чувашия, Россия)
9. Исраилова Нелла Амантаевна (Бишкек, Кыргызстан)
10. Кубединова Ленара Шакировна (Симферополь, Крым, Россия)
11. Мамедова Масума Гусейновна (Баку, Азербайджан)
12. Офлазер Кемаль (Доха, Катар)
13. Садыков Ташполот (Бишкек, Кыргызстан)
14. Салчак Аэлита Яковлевна (Кызыл, Тыва, Россия)
15. Сиразитдинов Зиннур Амирович (Уфа, Башкортостан, Россия)
16. Татевосов Сергей Георгиевич (Москва, Россия)
17. Торотоев Гаврил Григорьевич (Якутск, Саха, Россия)
18. Арипов Мирсаид Мирсиддикович (Ташкент, Узбекистан)

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

1. Сирожиддинов Шухрат
2. Мухамедова Саодат
3. Абдурахмонова Нилуфар
4. Гатиатуллин Айрат
5. Хакимов Муфтах
6. Аббасова Татьяна
7. Участники лаборатории «Компьютерная лингвистика» в ТашГУУЛ

Рецензенты:
д.ф.н., профессор Б. Менглиев (Узбекистан)
PhD, профессор Карлоз Гомез (Испания)

СЕКЦИЯ 1.
ФОРМАЛЬНЫЕ И КОНЦЕПТУАЛЬНЫЕ МОДЕЛИ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ

QUESTION AND ANSWERING SYSTEM FOR TURKIC LANGUAGES

E. ADALI, Technical University of Istanbul, Istanbul, Turkey, adali@itu.edu.tr

One of the application of NLP is Question and Answering (QA) system. In this paper we will introduce a QA system which is based on knowledge. QA system uses their own database such as Weather reporting, scores of games, booking a room at the hotel, booking a flight and e-commerce application. There are question and answer database which are working in parallel. The data in two database increases by the time. The proposed system is also a learning system. The result of the learning process new data is stored in two database. Some answer needs additional information such as room availability or weather information. For these reason an information database is attached to system. The components of the proposed system are: analyzing of question, interpretation, learning and answering.

Key words: *Question and Answering system, Natural Language Processing, Information Retrieval.*

ВОПРОСНО-ОТВЕТНЫЕ СИСТЕМЫ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ

Е. Адалы, Стамбульский технический университет, Стамбул, Турция, adali@itu.edu.tr

Одним из применений НЛП является система вопросов и ответов (QA). В этой статье мы представим систему обеспечения качества, основанную на знаниях. Система QA использует свою собственную базу данных, такую как отчеты о погоде, множество игр, бронирование номеров в отеле, бронирование авиабилетов и приложение электронной коммерции. Есть база данных вопросов и ответов, которые работают

параллельно. Данные в двух базах данных с течением времени увеличиваются. Предлагаемая система также является системой обучения. В результате процесса обучения новые данные хранятся в двух базах данных. Некоторые ответы требуют дополнительной информации, такой как наличие комнаты или информация о погоде. По этой причине к системе прикреплена информационная база данных. Компоненты предлагаемой системы: анализ вопроса, интерпретация, обучение и ответы.

Ключевые слова: *система вопросов и ответов, обработка естественного языка, поиск информации.*

INTRODUCTION

One of the application of NLP is Question and Answering (QA) system. In this field there are many researches and applications. Some competition are organized to increase the performance of the systems. The first example of QA is ELISA which is developed by Weizenbaum at MIT in 1964 [1]. As far as the coverage of the QA systems are concerned they can be classified as general and specific. Some question are short and some of them are long. There are two type of QA systems: 1º Information Retrieval (IR) based QA and; 2º Knowledge based QA.

In this paper we will introduce a QA system which is based on knowledge, but some brief information will be given about IR based QA system, first.

Companies, institution and people are putting information on the Web therefore we can find answer of any question on the Web or some other collections. In order to get any information from the Web some IR systems had been developed which have crawler, indexing and search engine. An IR system implement the following steps:

- Download the pages,
- Discover the links in page and download them,
- Discover the keywords of pages,
- Put the keywords into index,
- Indexing of word and page.

Indexing of pages written in Turkish is somewhat difficult than any India-European pages. Although an English words can get one prefix and one suffix Turkish words can take many suffixes. Therefore we need to discover the root or

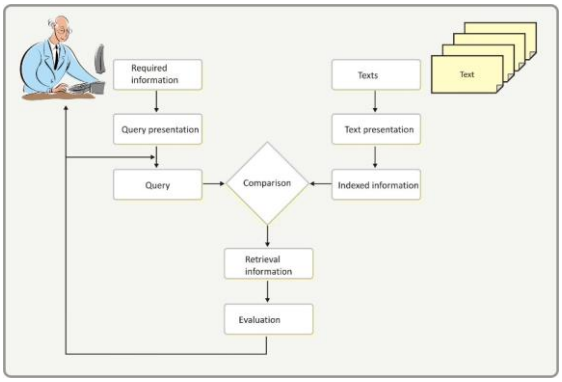


Figure-1: The processing of information retrieval form the Web

stem of the word.

The working way of the IR system is given in Figure-1. In order to get required text logical or vector based methods are used.

Logical Method: In logical methods, the similarity between query and retrieval text are logical. The basic features of the logical methods are:

- Logical methods look into match up with query and retrieval therefore cannot find similarities.
- Retrieval text cannot be graded.
- The weight and impotency of the keywords are the same.
- The usage of the logical operator are more effective than keywords.

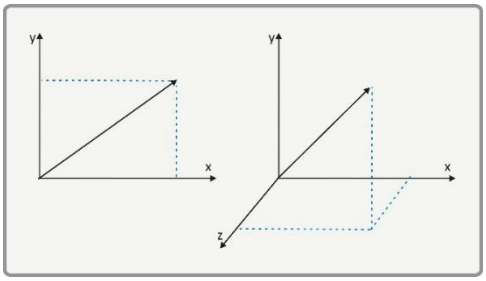


Figure-2: Two or three dimensional vectors

Vector Representation: In this method keywords, query and retrieval text are represented by two or three dimensional vectors, Figure-2.

Key words: <t_1, t_2, t_3, …,t_n>

Query: $\overrightarrow{Si} = (b_{i,1}, b_{i,2}, , , b_{i,n})$

Text : $\overrightarrow{bi} = (a_{i,1}, a_{i,2}, , , a_{i,n})$

a_i is the weight of t_i in query

Terms and text are shown as matrix form in Table-1 and an example of matrix representation is given in Table-2 respectively.

Table-1: Matrix presentation

		Terms space				
		t_1	t_2	t_3	...	t_n
Text space	**B1**	a_{11}	a_{12}	a_{13}	...	a_{1n}
	B2	a_{21}	a_{22}	a_{23}	...	a_{2n}
	B3	a_{31}	a_{32}	a_{33}	...	a_{3n}
	...					
	Bm	a_{m1}	a_{m2}	a_{m3}	...	a_{mn}
		b1	b2	b3	...	bn
		Query space				

Table-2: Example for matrix presentation: 1 means this word is shown in the text, 0 is not shown.

		Terms space						
		ak	al	alaca	ama	anlık	...	zaman
Text space	**1th text**	1	0	0	0	0	…	0
	2nd text	0	1	0	0	0	…	1
	3rd text	0	0	0	0	0	…	0
	...	…	…	…	…	…	…	…
	mth text	0	0	1	0	1	…	0
		b1	b2	b3			...	bn
				Query space				

Although matrix representation do not show the number of words in text, vector representation shows the number of roots, stems and word phrases. In Figure-3 an example is given.

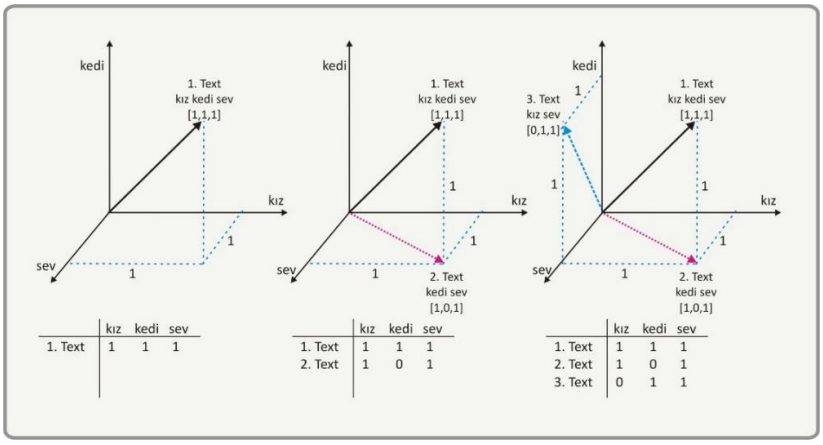


Figure-3: Three terms and three texts example

In vector representation method queries are also represented as vector. The similarity of query and text vector is calculated internal multiplication or cosine similarity

$sim(b_i, s) = \sum_{j=1}^{t} b_{i,j} w_{s,j}$ (internal multiplication)

$\text{sim}(b_i, s) = \cos \acute{e}$

Information Retrieval Based Question and Answering

Information retrieval based answering systems use information on the Web. The principle of the system is depicted in Figure-4

In order to find the appropriate answer a question logical and vector space methods are used. Solving a question is related the language. Some example of short question and answer are given in Table-3.

Table-3: Examples of short question and answers

Question	Answer
What is the name of tallest mountain name?	Ağrı Mountain
What is the name of drink has Turkish?	Turkish Coffee
What is the name of name of famous Turkish dessert?	Baklava
What is the abbreviation of İstanbul Technical University?	İTÜ
How tall is Ağrı Mountain?	5.137
Who is the founder of Turkey?	Atatürk
Where is the biggest lake of Turkey?	Van
What is the name of Turkish folk instrument?	Bağlama

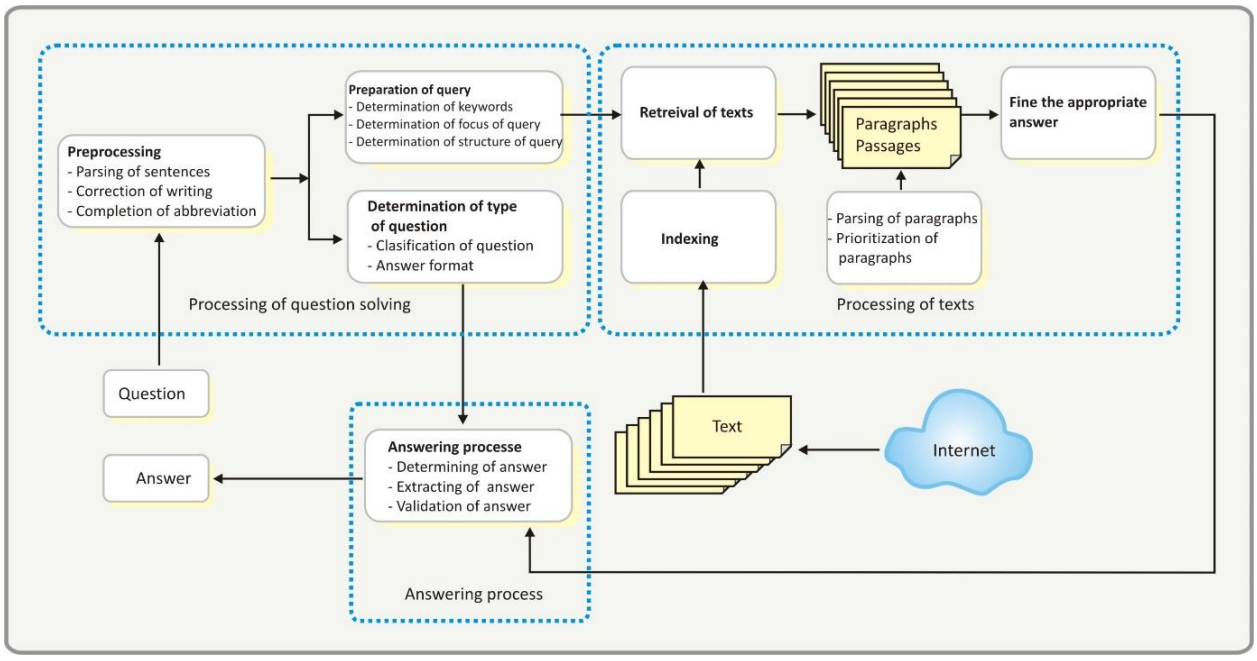


Figure-4: Basic components of question answering system

2.1 Question Solution Process

Some question may consists of more than one sentence therefore question is firstly separated sentences. In a sentence there are some typing errors and abbreviations, eg: *slm: selam, tşk: teşekkür*. This problems must be solved in preprocessing step. The steps of the question solution process are as follows:

Solving of Question: This step is the understanding phase of the question in other word to find the keywords, focus of the question and type of the question. Keywords will be categorized as follows:

- Person name
- Abbreviation
- Institution
- Location (state, city)
- Article (food, plant..)
- Time (century, year, month, period)
- Numerical value

Keywords can determine the type of the question. For example, the spirit of the question «*What is the name of name of famous Turkish dessert*?» is Turkish dessert. Some words and word phrase can determine the focus of the question. For example, the focus of the question «*What is the name of tallest mountain name?*» is tallest mountain. The last step is finding type of the question and answer. If we evaluate the question *What is the name of tallest mountain name?*

- Type of answer : Location
- Query: Turkey, the tallest
- Focus: mountai

Classification of Questions: In order to answer a question correctly, the class of the question must be known. Some Word in the question can help to find the class of the question. Finding of question class help to determine class of answer as well. Li and Roth give a table for question class [2,3]. Machine Learning (ML), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Naïve Bayes, Bag of Word (BoW) methods are used for question classification.

Morphological Parsing: After solving the question, the keywords are obtained and can be used for India-European languages but it is necessary to find the root or stem of the word, eg.

- özelliğindeki → özellik
- niteliğindeki → nitelik
- kurallarından → kural
- arananlar → ara
- çalışmaların → çalışma
- anlamlandırma → anlam

Therefore morphological analysis is necessary.

Structure of Query: The structure of the query must be organized depend on the application.

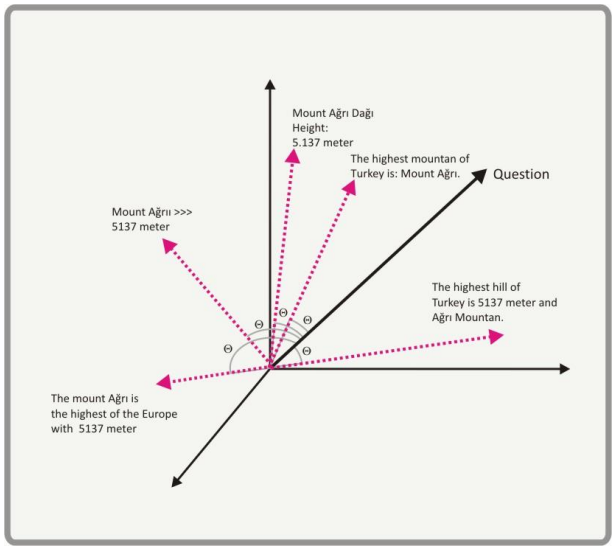


Figure-5: Vector representation of question and possible answers.

2.2 Text Processing

All collected texts firstly indexing then parsing of paragraphs. Passage have to be find in text. Finally finding an appropriate answer.

2.3 Answering Process

Answering steps consists of determining, extraction and validation of answer. The find the best answer vector methods can be used as shown in Figure-5.

Knowledge Based Question and Answering

Some QA system uses their own database. Some examples of such systems are:

- Weather reporting,
- Scores of games,
- Booking a room at the hotel,
- Booking a flight,
- E-commerce application

Some examples are shown in Table-4

Question	Answer
What is the score of Tukey and Uzbekistan?	1-1
Do you have a room at April 23?	Yes we have
What will be weather tomorrow?	Cloudy, rainy and 19 degree
I am looking for a solitaire ring, do you have it?	We are sorry, we do not.
Do you have a flight to Tashkent?	Yes we do.
I want to take two ticket for Swan lake	Which performance?
May 19, soiree	Which place?
Front row please	I can give you in third row.

In the knowledge based system there are question and answer database which are working in parallel. The data in two database increases by the time. The proposed system is also a learning system. The result of the learning process new data is stored in two database. Some answer needs additional information such as room availability or weather information. For these reason an information database is attached to system. The learning process and parallel database system is given in Figure-6.

Same examples data of the information database is given in Table-5 and initial data of parallel database is shown in Table-6.

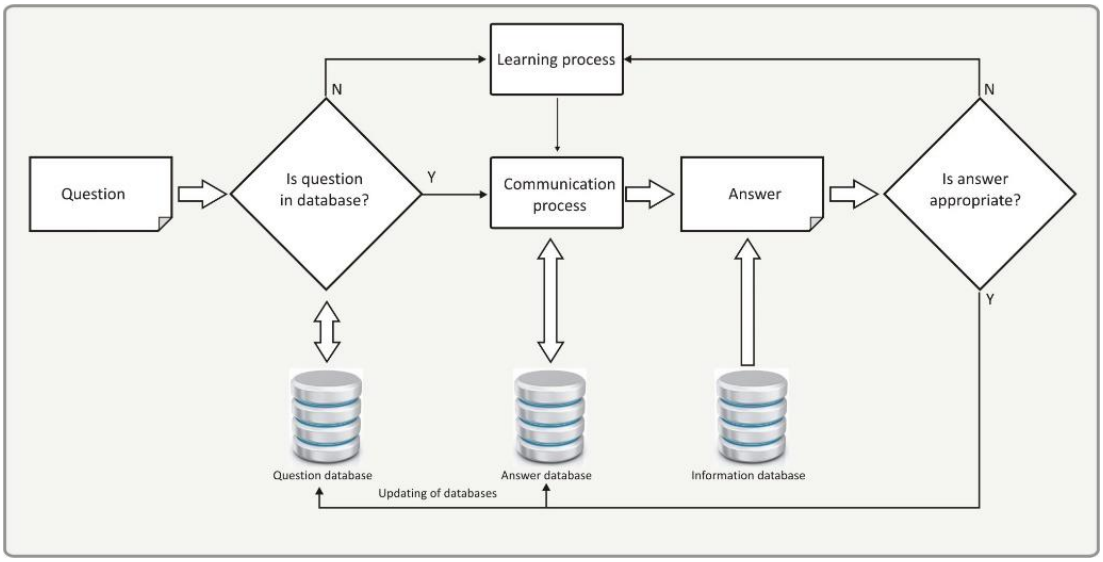


Figure-6: Learning process and parallel database.

Table-5: Initial state of information database

Time check-in	Time check-out	Type of rooms	Number of rooms	Room price
18 October 2018	25 October 2018	Standard	5	125
19 October 2018	21 October 2018	Standard	2	125
18 October 2018	28 October 2018	Suit	3	150

Table-6: Initial state of parallel databases

No:	Question	Subject	Object	Complement	Predicate	Answer
1	Do you have room?	2nd plural	Vacant room		Is there?	Yes we have
2	Do you have room at January 20th	2nd plural	Vacant room	January 20th	Is there?	Yes we have room at January 20th
3	What is the room price?	3rd single	Room	price	What is?	125 TL
4						

Question : Do you have a room at October 18th?

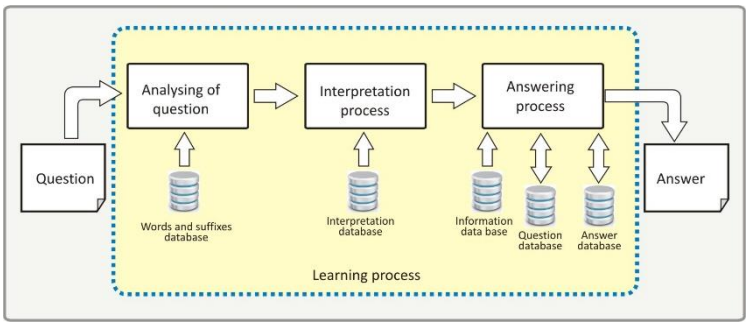
- Predicate: Looking for a place to stay; is there?
- Object : Vacancy, vacant room
- Complement: October 18th

The question and the answers are searched in database. The second question is similar to the question but the complement is different. In this case this question is asked to customer:

We have room at October 18th. Is it good for you?

If the answer is «OK» the following question and the answer are written into parallel databases:

- Question: Do you have room at October 18th?
- Answer: Yes we have room at October 18th

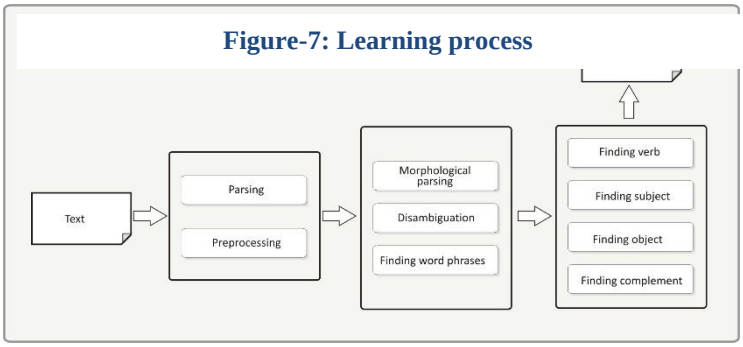


For agglutinative language the learning process consists of analyzing, interpretation and answering steps which are depicted in Figure-7

3.1 Analyzing of Question

A question consists of one or more sentences so firstly sentences must be find. All spelling error must be corrected and some abbreviations are converted to regular form. This step is called preprocessing phase.

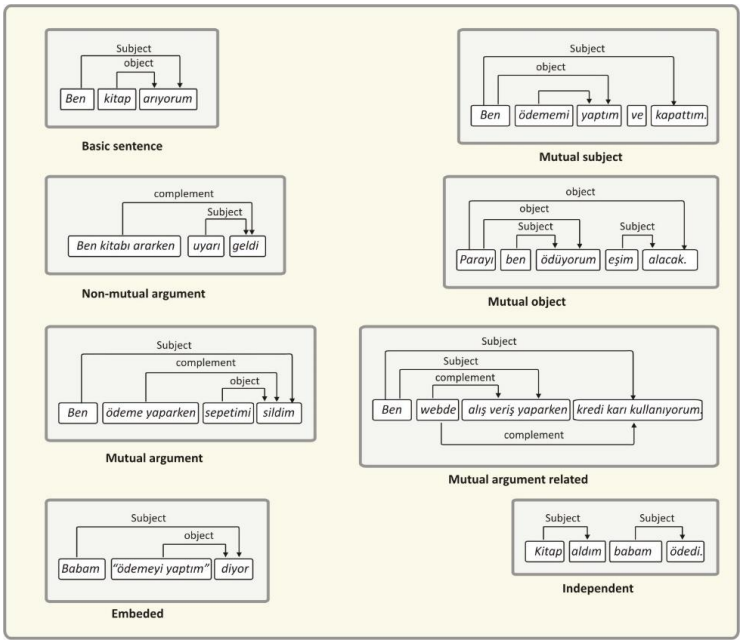
Figure-7: Learning process



In the second step the arguments of the sentence are discovered. A Turkish sentence syntax is SOV (Subject, Object, Verb).

The detail of the analyzing of a question is drawn in Figure-8. After parsing sentence and preprocessing of the sentence, morphological parsing, disambiguation and finding word phrases steps have to be done. After these steps predicate, subject, object and complement of the sentence can be found. After all we can have solved text.

Figure-8: Analyzing of question



The structure of Turkish sentence can be classified in four set:

- Basic
- Combined
- Sequential
- Complex (embedded)

All possible sentence structures are given with examples in Figure-9.

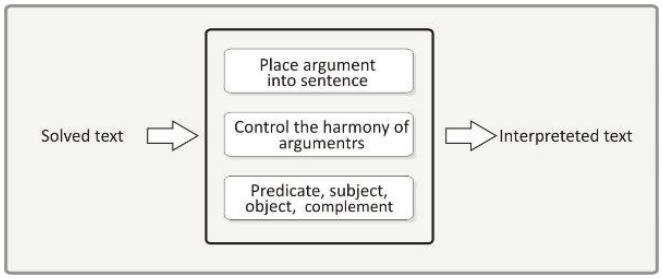
Each argument can be find by using of grammar rule of the language.

3.2 Interpretation

After finding the arguments of the sentence, sentence can be interpreted. In order to do this;

Figure-9: Possible structure of Turkish sentence

- All arguments are placed in their appropriate place.
- The harmony of the arguments must be checked.



- Predicate, subject, object and complement must be defined.

All this steps are shown in Figure-10.

3.3 Learning and Answering

In order to prove the interpretation of the question is correct, some questions are asked to user and get his/her approval. This process is shown in Figure-11.

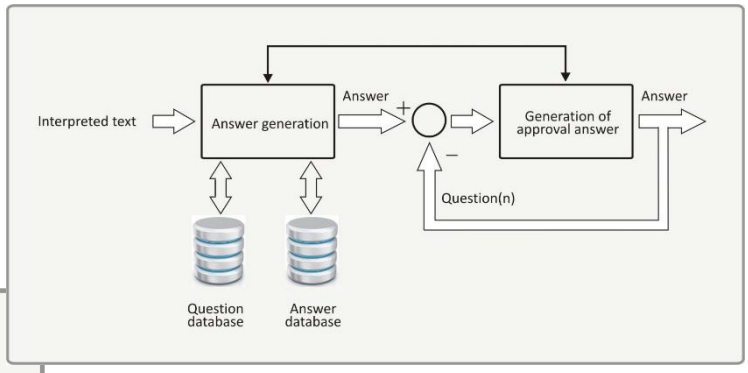
Question and answer are represented as vector and similarity is calculated as follows:

Figure-10: Interpretation process

Query: $\overrightarrow{bi} = (w_{i,1}, w_{i,2}, , , w_{i,n})$

Answer : $\overrightarrow{ai} = (w_{i,1}, w_{i,2}, , , w_{i,n})$

$\text{sim}(a_i, b_i) = \cos 6$



Do you have room at October 18th

Do you have room at January 20th

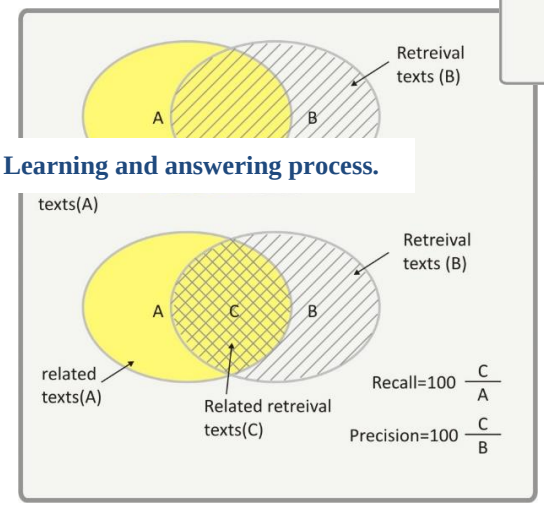


Figure-11: Learning and answering process.

$$Sim = \frac{number\ of\ mutual\ words}{number\ of\ in\ bag\ of\ words} = \frac{5}{7}$$

In order to generate answer, first the structure of sentence is installed. The structure of a Turkish sentence is:

Subject + object + complement + adverb + verb

we + vacancy + January 20 + — + have

- + boş oda + 18 Ekim + — + var

Figure-12: Recall and precision of IR

Performance of QA System

The recall and precision of IR methods is depicted in Figure-12.

REFERENCES:

[1] J. Weizenbaum, ELIZA A Computer Program For the Study of Natural Language Communication Between Man And Machine MIT Cambridge, Communication of the ACM Volume 9 / Number 1 / January, 1966

[2] X. Li. D. Roth, *Learning Question Classifiers*, COLING '02 Proceedings of the 19th international conferenceon Computational linguistics — Volume 1 Pages 1-7, 2002

[3] X. Li. D. Roth, *Learning Question Classifiers: The Role of Semantic Information*, Natural Language Engineering 1 (1): 000–000. Cambridge University Press. 2004

[4] *Scoring, term weighting and the vector space model*, Cambridge University Press. 2009

[5] *Probabilistic information retrieval*, Cambridge University Press. 2009

[6] M. Sahlgren, *The Word-Space Model, Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*, Doktora Tezi, Stockholm University, 2006

[7] A. Delibaş, *Doğal Dil İşleme ile Türkçe Yazım Hatalarının Denetlenmesi*, Yüksek Lisans Tezi, İTÜ Fen Bilimleri Ens. 2008

[8] N. Coşkun, *Türkçe Tümcelerin Ögelerinin Bulunması,* Yüksek Lisans Tezi, İTÜ Fen Bilimleri Ens. 2013

УДК 004 934

ПРИМЕНЕНИЕ НЕПРЕРЫВНОГО ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ДЛЯ ФИЛЬТРАЦИ СИНТЕЗИРОВАННОГО РЕЧЕГОГО СИГНАЛА

В. И. Семенов, А. К. Шурбин, Чувашский государственный университет им. И. Н. Ульянова, Чебоксары, Россия, syundyukovo@yandex.ru; shurti@mail.ru

В работе представлен алгоритм обратного непрерывного вейвлет-преобразования в частотной области с применением быстрого преобразования Фурье. Алгоритм позволяет на четыре порядка увеличить скорость вычисления обратного непрерывного вейвлет-переобразования по сравнению с прямым численным интегрированием.

***Ключевые слова**: вейвет-преобразование, масштабный коэффициент, синтез речи, Фурье преобразование, речевой сигнал.*

ADOPTION OF CONTINUOUS WAVELET TRANSFORM FOR SYNTHESIZED SPEECH SIGNAL FILTERING

V.I. Semenov, A.K. Shurbin, Chuvash State University named after I.N. Ulyanov", Cheboksary, Russia, syundyukovo@yandex.ru, shurti@mail.ru

The paper presents an algorithm for the inverse continuous wavelet transform in frequency domain using the fast Fourier transform. The algorithm allows to increase the speed of calculating the inverse continuous wavelet transform by four times compared with direct numerical integration.

***Key words:** wavelet transform, scale factor, speech synthesis, Fourier transform, speech signal.*

При синтезе речевого сигнала в местах соединения фонем возникают значительные перепады частот и амплитуды сигнала, что негативно сказывается на качестве синтезируемой речи. Разрыв в точках соединения отчетливо ощутим на слух. Чтобы устранить этот недостаток, авторами используется локальная фильтрация. Вблизи точек соединения фонем высокочастотные коэффициенты приравниваются нулю, а остальные участки остаются без изменений. В результате реконструкции синтезированного сигнала в местах соединения фонем разрывы сглаживаются, что положительно сказывается на качестве синтезированной речи. Для фильтрации используется быстрое непрерывное вейвет-

преобразование (ВП). Непрерывное ВП одномерного сигнала $S(t)$ производится по формуле:

$$W(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} S(t)\ \psi\left(\frac{t-b}{a}\right)\ dt, \qquad (1)$$

где первый аргумент a (временной масштаб) аналогичен периоду осцилляций, а второй b – смещению сигнала по оси времени. Реконструкция выполняется с применением формулы обратного непрерывное ВП:

$$S(t)=C_{\psi}^{-1}\int_{0}^{\infty}\int_{-\infty}^{\infty}\psi\left(\frac{t-b}{a}\right)\ W(a,b)\frac{dadb}{a^{3+k}}, \qquad (2)$$

где C_{ψ} – нормализующий коэффициент:

$$C_{\psi}=\int_{-\infty}^{\infty}|F_{\psi}(\omega)|^2\cdot\omega^{-1}d\omega<\infty,$$

F_{ψ} (ω) – Фурье–спектр базисной функции, ω — циклическая частота, k – показатель степени масштабного множителя.

Вычисление ВП прямым численным интегрированием для больших временных последовательностей по формулам (1) и (2) занимает длительное время. Для увеличения быстродействия, авторами разработан алгоритм непрерывного быстрого ВП в частотной области с использованием быстрого преобразования Фурье (БПФ). Вейвлет-преобразование используется не только для синтеза речи, но и для распознавания речи [1,2,3]. Алгоритм численного вычисления прямого быстрого непрерывного вейвлет-преобразования в частотной области приведен в работах [4,5]. Нормализующий коэффициент в формуле (2) непрерывного обратного вейвлет-преобразования $C = C_{\psi}$ в разработанном алгоритме вычисляется из аналога теоремы Парсеваля для вейвлет-коэффициентов:

$$\int S(t)S^{*}(t)dt=C^{-1}\iint W(a,b)W^{*}(a,b)\frac{dadb}{a^{2}}\ . \qquad (3)$$

После определения нормализующего коэффициента C из (1) он подставляется в формулу:

$$S(t)=C^{-1}\int_{0}^{\infty}\int_{-\infty}^{\infty}\psi\left(\frac{t-b}{a}\right)\ W(a,b)\frac{dadb}{a^{2}}\ . \qquad (4)$$

Теоретической основой вычисления обратного непрерывного быстрого вейвлет-преобразования сигнала $S(t)$ в частотной области является использование формул (4) и (3). Обратным преобразованием произведения спектров вейвлет-спектра $W(a,b)$ и вейвлета $\psi(t)$ вычисляется интеграл по переменой b. Суммированием полученного интеграла по масштабному коэффициенту a рассчитывается реконструированный сигнал $S(t)$.

Алгоритм численного вычисления обратного непрерывного вейвлет-преобразования по формуле (4) в частотной области включает следующие шаги.

1. Вычисляются коэффициенты тригонометрического ряда $a_1(n)$ вейвлет-спектра *W(a,b)* с использованием прямого БПФ по формуле:

$$a_1(n) = \frac{1}{N}\sum_{k=0}^{N-1} W(a,k)\cos(\frac{2\pi nk}{N}).$$

2. Вычисляются коэффициенты тригонометрического ряда $b_1(n)$ вейвлет-спектра *W*(*a*,*b*) с использованием прямого БПФ по формуле:

$$b_1(n) = \frac{1}{N}\sum_{k=0}^{N-1} W(a,k)\sin(\frac{2\pi nk}{N}).$$

3. Вычисляются коэффициенты тригонометрического ряда $a_2(n)$ вейвлета $\psi(t)$ с использованием прямого БПФ по формуле:

$$a_2(n) = \frac{1}{N}\sum_{k=0}^{N-1} \psi(k)\cos(\frac{2\pi nk}{N}).$$

4. Вычисляются коэффициенты тригонометрического ряда $b_2(n)$ вейвлета $\psi(t)$ с использованием прямого БПФ по формуле:

$$b_2(n) = \frac{1}{N}\sum_{k=0}^{N-1} \psi(k)\sin(\frac{2\pi nk}{N}).$$

5. Вычисляется комплексно сопряженный спектр по формулам,

$$c_1(n) = a_1(n)\cdot a_2(n) + b_1(n)\cdot b_2(n),$$
$$c_2(n) = b_1(n)\cdot a_2(n) - a_1(n)\cdot b_2(n).$$

Большинство непрерывных вейвлетов – либо четные, либо нечетные функции. Для четных вейвлетов ряд составлен из одних косинусов, а для нечетных – из одних синусов. Для четных вейвлетов $b_2(n) = 0$ и

$$c_1(n) = a_1(n)\cdot a_2(n), \quad (5)$$
$$c_2(n) = b_1(n)\cdot a_2(n). \quad (6)$$

Для нечетных вейвлетов, $a_2(n) = 0$ и

$$c_1(n) = b_1(n)\cdot b_2(n), \quad (7)$$
$$c_2(n) = -a_1(n)\cdot b_2(n). \quad (8)$$

6. Для четного вейвлета путем $M + 1$ обратных преобразования Фурье от комплексно сопряженного спектра (5), (6) вычисляется функция $s'_m(n)$:

$$s'_m(n) = \sum_{k=0}^{N-1} c(k)\exp(i\frac{2\pi nk}{N}).$$

7. Для нечетного вейвлета путем $M + 1$ обратных преобразований Фурье от комплексно сопряженного спектра (7), (8) вычисляется функция $s'_m(n)$:

$$s'_m(n) = \sum_{k=0}^{N-1} c(k)\exp(i\frac{2\pi nk}{N}),$$ (обозначение ‘ не означает дифференцирование).

8. По формуле (3) вычисляется нормализующий коэффициент *C*.

9. По формуле $S(n) = C\sum_{m=0}^{m} s'_m(n)$,

реконструируется сигнал, где m — уровень декомпозиции. Постоянную C, можно определить проще, используя следствие формулы (3) (теоремы Парсеваля). В пространстве действительных функций плотность энергии сигнала:

$$E_W(a,b) = W_l^2(a,b).$$

Локальная плотность энергии в точке t_0 $E_\delta(a,t_0) = W_l^2(a,t_0)$.

Тогда, $$S(t_0) = C\sum_{m=0}^{m} s_m^{'}(t_0). \qquad (9)$$

Постоянная C вычисленная по формуле (9), совпадает с постоянной, найденной по формуле (9). Чтобы при вычислении по формуле (9) не было деления на ноль или умножения на отрицательное число, постоянную C лучше вычислить для функции в максимуме.

ЛИТЕРАТУРА:

1. Патент на изобретение №2403628 РФ Способ распознавания ключевых слов в слитной речи. Семенов В.И. Желтов П.В., № 2008141557/09(053961). Заявл. 20.10.2008. Опубл. 10.11.2008 г.

2. Семенов В.И., Желтов П.В. Распознавание речи на основе вейвлет-преобразования./ Чуваш. ун-т. Чебоксары, 2008. 16с. Деп. в ВИНИТИ РАН 29.02.08, №174-В2008.

3. Семенов В.И., Желтов П.В. Вейлет-преобразование акустического сигнала/ КГТУ им. А.Н. Туполева. Казань, 2008. 102 с..

4. Семенов В.И., Шурбин А.К., Михеев К.Г., Михеев Г.М. Фильтрация изображений, полученных с помощью оптического микроскопа, с применением кратномасштабного анализа. Химическая физика и мезоскопия том 16 №3, Ижевск, 2014. С. 399-404.

5. В.П. Желтов, П.В. Желтов, В.И. Семенов, А.И. Трофимова, А.К. Шурбин. Распознавание слитной речи с использованием вейвлет-преобразования. Труды Казанской школы по компьютерной и когнитивной лингвистике, TEL-2014, Казань, 2014. С. 9-13.

ЛОГИКО-ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ РУССКОГО ЯЗЫКА

М. Хакимов, Национальный университет Узбекистана им. Мирзо Улугбека, Ташкент, Узбекистан, muftah@mail.ru

В данной работе приведены логико-лингвистические модели слов и предложений по типам русского языка для системы машинного перевода «Tarjimon — LMX». Система «Tarjimon — LMX» разрабатывается как многоязычная на основе технологии моделируемого компьютерного переводчика разработанной для машинного перевода. Логико-лингвистические модели описываются с помощью входного языка, разработанного для моделирования естественных языков.

Ключевые слова: *логико-лингвистическая модель слов, существительное, прилагательное, глагол, местоимение, наречие, числительное, логико-лингвистическая модель предложений, повествовательное предложение, вопросительное предложение, восклицательное предложение, русский язык, расширяемый входной язык, система машинного перевода.*

LOGIC-LINGUISTIC MODELS OF RUSSIAN

M. Khakimov, National university of Uzbekistan after Mirzo Ulugbek, Tashkent, Uzbekistan, muftah@mail.ru

In the given work logic-linguistic models of words and offers on types of Russian for the machine translation system «Tarjimon — LMX» are resulted. The system «Tarjimon — LMX» is developed as multilingual on the basis of technology of the modelled computer translator developed for machine translation. Logic-linguistic models are described by means of the source language developed for modelling of natural languages.

Key words: *logico-linguistic model of words, a noun, an adjective, a verb, a pronoun, an adverb, the numeral, logico-linguistic model of offers, the narrative offer, a question, the exclamatory offer, Russian, the expanded source language, the machine translation system.*

ВВЕДЕНИЕ

Каждый естественный язык (ЕЯ) является сложной системой, состоящих математически неструктурированных и не формализованных составных частей. Однако проведенные исследования над ЕЯ показывают, что не структурированность и не формализованность ЕЯ, можно привести к структурированному и формализованному виду, используя линейную

методологию – выявление состава слова и построением логико-лингвистических (семантических) моделей по типам слов и предложений, и далее построением математических моделей с помощью входного языка [1]. Данную методологию можно определить как степень формализации языка. Степень формализации в свою очередь определяет степень формализации семантики ЕЯ и точность алгоритма. Поверхностное понимание степени формализации ЕЯ, что формализованный язык – абстрактная, полностью оторванная от содержания конструкции с простой логической структурой приводит к низкой технологии машинного перевода [2]. Формализация позволяет выделить различные части ЕЯ, исследовать динамику их связей и главным образом даст возможность описания семантической структуры. Все эти качества очень существенны, когда используется общее ядро системы, т.е. когда над всеми ЕЯ входящие в данную среду перевода применяется единый системный подход, независимо с какого на какой ЕЯ осуществляется перевод.

Так как, русский язык (РЯ) также предназначен для включения в систему машинного перевода «Tarjimon — LMX», то он должен быть исследован с точки зрения формализации на принципах моделируемой компьютерной технологии. Следовательно, необходимо построение логико-лингвистических моделей слов и предложений по типам РЯ.

1. ЛОГИКО-ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ СЛОВ ПО ТИПАМ

1.1. Общие логико-лингвистические модели слов

Лексический анализ словообразования русского языка показывает, что русская словоформа может состоять из пяти частей: корня, приставки, суффикса, окончания, частицы и союза. Опираясь на данную структуру русского языка, для словообразования мы получаем следующие общие логико-лингвистические модели. Здесь и далее все логико-лингвистические модели излагаются на расширяемом входном языке [1]**.**

1. ⇓ частица $\oplus\Downarrow$ приставка $\oplus$ корень $\oplus\Downarrow$ корень $\oplus$ суффикс $\oplus\Downarrow$ суффикс $\oplus\Downarrow$ окончание (сложное слово)
2. корень $\oplus$ частица
3. корень $\oplus$ союз
4. корень $\oplus$ морфема (о/е) $\oplus$ корень (сложное слово)
5. корень $\oplus$ суффикс числительного $\oplus\Downarrow$ числительное
6. числительное $\oplus$ суффикс числительного $\oplus$ суффикс числительного
7. числительное $\oplus$ числительное

1.2. Логико-лингвистические модели имен существительных

При составлении имени существительного подкоренным словом может являться существительное, прилагательное, глагол и местоимение. При

присоединении к подкоренному слову приставки, суффикса и окончания образуется существительное. А в случае, когда между двумя подкоренными словами имеется морфема – о/е, можно строить сложное существительное. Определяем семь разных видов такого случая. Логико-лингвистические модели составления **имен существительных** напишем в нижеследующих вариантах.

1. ⇓ приставка ⊕ корень ⊕ суффикс ⊕⇓ суффикс ⊕⇓ окончание (пример: преобразование).
2. корень прилагательного ⊕ корень существительного ⊕⇓ суффикс ⊕⇓ окончание (пример: краснодеревщик).
3. корень существительного ⊕ окончание (пример: стекло, книга)
4. ⇓ корень прилагательного ⊕ корень существительного ⊕⇓ корень глагола ⊕⇓ суф фикс ⊕⇓ окончание (пример: длинноусый)
5. корень местоимения ⊕ корень глагола (пример: ничего неделание)
6. частица ⊕ корень существительного (пример: небыль, неприятель)
7. корень существительного ⊕ морфема (о/е) ⊕ корень существительного (пример: словообразование).

Исключения I.

Связанные корни: свергнуть, отвергнуть, низвергнуть, добавить, убавить, отбавить, прибавить, добавка, прибавка, прибавление, вонзить, пронзить. В этих словах корни не являются полноценными корнями, т.е. не могут выступать в роли отдельных слов. Такие корни называют радиксоидами.

Исключения II.

1) Радиксоиды вверг/верис (свергнуть, изверисение).
2) –у- (обуть, разуть, обувь). Слово обувь – не членимое простое.
3) –н- (поднять, отнять, унять)
   - ировать (агитировать)
   - ация (агитация)
   - атор (агитатор)
   - изм (атеизм)
   - ист (атеист)
   - янт (шекулянт)

1.3. Логико-лингвистические модели прилагательных

При построении **имён прилагательных** подкоренным словом может служить прилагательное и числительное. При присоединении к этим подкоренным словам частицы, приставки, суффикса и окончания можно образовать прилагательное. И в случае, когда между двумя подкоренными словами образуется морфема – (о/е), можно построить прилагательное.

Определяем пять разных видов такого случая. Логико-лингвистические модели составления прилагательных опишем в следующих видах:

1. ⇓ приставка ⊕ корень ⊕ суффикс ⊕⇓ окончание (пример: сверхсильный, лимонный)
2. ⇓ приставка ⊕ корень прилагательного ⊕⇓ корень существительного ⊕⇓ корень прилагательного ⊕ суффикс ⊕ окончание (пример: наиумнейший)
3. частица ⊕ корень прилагательного ⊕⇓ суффикс ⊕⇓ окончание (пример: некрасивый)
4. корень прилагательного ⊕ морфема (о/е) ⊕ корень прилагательного ⊕⇓ суффикс ⊕⇓ окончание (пример: победоносный)
5. ⇓ приставка ⊕ числительное ⊕ корень прилагательного ⊕ суффикс ⊕ окончание (пример: двойственный, одноглазый).

1.4. Логико-лингвистические модели глагола

При построении **глагола** подкоренным словом может служить существительное, глагол и прилагательное. При присоединении к этим подкоренным словам приставки, суффикса, постфикса и окончания образуется глагол. Определяем шесть разных случаев построения глагола. Эти логико-лингвистические модели напишем в следующих формах:

1. ⇓ частица ⊕⇓ приставка ⊕ корень ⊕ суффикс ⊕⇓ суффикс ⊕⇓ окончание (пример: не приходить)
2. ⇓ приставка ⊕ корень существительного ⊕ суффикс ⊕⇓ окончания (пример: безобразничать)
3. частица ⊕ корень глагола (пример: не брать)
4. корень глагола ⊕ постфикс (пример: здороваться)
5. корень существительного ⊕ суффикс ⊕⇓ суффикс ⊕⇓ окончание (пример: хулиганить)
6. ⇓ приставка ⊕ корень прилагательного ⊕ суффикс ⊕⇓ суффикс ⊕⇓ окончание (пример: подсинить, разозлить)

1.5. Логико-лингвистические модели местоимений

При построении **местоимения** подкоренным словом могут служить местоимения и существительное, местоимения и прилагательное, местоимения и числительное. Они вместе образуют местоимение. Определяем шесть разных случаев построения местоимения. Эти логико-лингвистические модели напишем в следующих вариантах:

1. предлог ⊕ местоимение (пример: у меня, ко мне)
2. частица ⊕ местоимение (пример: никто)
3. местоимение ⊕ частица (пример: кто либо)

4. корень местоимение $\oplus$ корень существительного
5. корень местоимение $\oplus$ корень прилагательного
6. корень местоимение $\oplus$ корень числительного.

1.6. Логико-лингвистические модели наречий

При построении **наречия** подкоренным словом могут служить существительное, числительное, наречия и прилагательное. При присоединении к этим подкоренным словам частицы, приставки, суффикса и окончания образуется наречие. В случаях, когда встречаются два наречия вместе, образуется новое наречие. Определяем шесть разных видов построения наречий. Эти логико-лингвистические модели напишем в следующих формах:

1. приставка $\oplus\Downarrow$ корень числительного $\oplus\Downarrow$ корень $\oplus\Downarrow$ суффикс $\oplus\Downarrow$ окончание (пример: во вторых, по русск.)
2. корень $\oplus\Downarrow$ корень существительного $\oplus\Downarrow$ корень прилагательного $\oplus\Downarrow$ суффикс $\oplus\Downarrow$ окончание (пример: вечером, летом)
3. приставка $\oplus\Downarrow$ корень существительного $\oplus\Downarrow$ корень прилагательного $\oplus\Downarrow$ корень числительного $\oplus\Downarrow$ корень местоимение (пример: дважды, по-осеннему)
4. частица $\oplus$ корень наречие (пример: некрасиво)
5. корень наречие $\oplus$ частица (пример: как небудь)
6. наречие $\oplus$ наречие (пример: чисто пречисто)

1.6. Логико-лингвистические модели числительных

При построении **числительного** подкоренным словом служить само числительное. При присоединении к подкоренному слову суффикса и числительного образуется новое числительное. Определяем три разных вида построения числительного. Эти логико-лингвистические модели опишем в следующих формах:

1. корень числительного $\oplus$ суффикс $\oplus\Downarrow$ числительное (пример: второй, двадцатый)
2. корень числительного $\oplus$ суффикс $\oplus\Downarrow$ суффикс (пример: пятисотый)
3. числительное $\oplus$ числительное (пример: двадцать пятый)

2. ЛОГИКО-ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ ПРЕДЛОЖЕНИЙ

2.1. Логико-лингвистические модели повествовательных предложений

В русском языке выявлено 11 видов образования **повествовательных предложений** и их можно описать в виде следующих логико-лингвистических моделей:

1. существительное $\oplus\Downarrow$ наречия $\oplus\Downarrow$ местоимение $\oplus\Downarrow$ числительное $\oplus\Downarrow$ прилагательное $\oplus\Downarrow$ глагол $\oplus\Downarrow$ прилагательное $\oplus$

существительное ⊕⇓ местоимение ⊕⇓ прилагательное ⊕⇓ глагол ⊕⇓ прилагательное ⊕⇓ числительное

2. местоимение ⊕ существительное ⊕ глагол ⊕⇓ прилагательное ⊕⇓ существительное ⊕ глагол
3. ⇓ местоимение ⊕⇓ прилагательное ⊕⇓ существительное ⊕⇓ местоимение ⊕⇓ прилагательное ⊕⇓ существительное ⊕⇓ прилагательное ⊕⇓ существительное ⊕⇓ глагол ⊕⇓ местоимение ⊕⇓ прилагательное
4. прилагательное ⊕ существительное ⊕⇓ наречие ⊕⇓ глагол ⊕⇓ прилагательное ⊕⇓ существительное.
5. существительное ⊕ глагол ⊕⇓ местоимение ⊕⇓ предлог ⊕⇓ прилагательное ⊕⇓ существительное ⊕⇓ союз ⊕⇓ местоимение.
6. местоимение ⊕⇓ модальное слово ⊕⇓ местоимение ⊕⇓ частица ⊕⇓ наречие ⊕ глагол ⊕⇓ местоимение ⊕⇓ прилагательное существительное.
7. предлог ⊕ местоимения ⊕ глагол ⊕ наречие ⊕ предлог ⊕ местоимение ⊕ существительное.
8. числительное ⊕ предлог ⊕ существительное ⊕⇓ прилагательное ⊕ существительное ⊕ числительное ⊕ предлог
9. местоимение ⊕⇓ существительное
10. модальное слово ⊕⇓ местоимение ⊕ существительное ⊕ частица ⊕⇓ модальное слово ⊕⇓ местоимение ⊕ прилагательное ⊕⇓ глагол.
11. частица ⊕⇓ глагол ⊕ местоимение ⊕ союз ⊕⇓ частица ⊕ глагол ⊕ модальное слово ⊕ наречия ⊕ местоимение ⊕ союз ⊕ частица ⊕ прилагательное ⊕ глагол ⊕ существительное.

2.2. Логико-лингвистические модели вопросительных предложений

Определив 9 видов образования **вопросительных предложений**, создадим для них следующие логико-лингвистические модели:

1. вопросительное слово ⊕⇓ местоимение ⊕⇓ существительные ⊕⇓ прилагательные ⊕⇓ глагол ⊕⇓ существительные.
2. существительные ⊕ наречие ⊕ глагол.
3. местоимение ⊕⇓ наречие ⊕⇓ существительные ⊕⇓ прилагательные ⊕⇓ глагол ⊕ существительные ⊕ глагол ⊕⇓ наречие ⊕⇓ местоимение ⊕⇓ глагол.
4. ⇓ существительные ⊕ наречие ⊕ прилагательные ⊕ глагол ⊕⇓ существительные.
5. наречие ⊕⇓ числительное ⊕⇓ существительные ⊕ местоимение ⊕⇓ местоимение ⊕ глагол ⊕ модальное слово ⊕ местоимение ⊕ предлог ⊕ местоимение.

6. местоимение $\oplus\Downarrow$ существительные $\oplus\Downarrow$ местоимение $\oplus$ прилагательные $\oplus$ глагол $\oplus\Downarrow$ предлог $\oplus\Downarrow$ существительные.
7. модальное слово $\oplus$ местоимение $\oplus\Downarrow$ местоимение $\oplus\Downarrow$ модальное слово $\oplus$ глагол.
8. союз $\oplus\Downarrow$ существительные $\oplus$ местоимение $\oplus$ частица $\oplus$ глагол.
9. предлог $\oplus\Downarrow$ существительные $\oplus$ местоимение $\oplus$ глагол.

2.2. Логико-лингвистические модели восклицательных предложений

В ходе грамматического анализа русского языка выявлено шесть видов **восклицательных предложений.** Их логико-лингвистические модели имеют следующие варианты:

1. существительные $\oplus$ глагол $\oplus\Downarrow$ существительные $\oplus\Downarrow$ наречие
2. местоимение $\oplus\Downarrow$ существительные $\oplus\Downarrow$ наречие + глагол $\oplus\Downarrow$ существительные
3. числительное $\oplus$ глагол
4. глагол $\oplus$ существительные $\oplus\Downarrow$ частица $\oplus\Downarrow$ предлог $\oplus\Downarrow$ местоимение $\oplus\Downarrow$ существительные $\oplus$ глагол
5. союз $\oplus\Downarrow$ местоимение $\oplus\Downarrow$ прилагательные $\oplus\Downarrow$ существительные $\oplus\Downarrow$ глагол
6. модальное слово $\oplus\Downarrow$ существительные $\oplus$ глагол

ЛИТЕРАТУРА:

1. Хакимов М.Х. Расширяемый входной язык математического моделирования естественного языка для многоязычной ситуации машинного перевода. ЎзМУ хабарлари, № 1, 2009, с.75-80.
2. Хакимов М.Х. Формальные системы машинного перевода в многоязычной ситуации. Материалы республиканской научной конференции «Современные проблемы математики, механики и информационных технологий», НУУз, Институт математики и ИТ АН РУз, Т, 2008, с.297-301

О РАЗРАБОТКЕ МНОГОФУНКЦИОНАЛЬНОГО МНОГОЯЗЫЧНОГО ИНТЕРНЕТ-СЕРВИСА НА БАЗЕ ТЮРКСКОЙ МОРФЕМЫ

А.Б. Альменова[1], А.Р. Гатиатуллин[2], А.М.Баширов[2],
[1]Кызылординский государственный университет им. Коркыт Ата, г. Кызылорда, Казахстан;[2]
Академия наук Республики Татарстан, г. Казань, Россия,
pr.almenova@mail.ru, ayrat.gatiatullin@gmail.com, a.basheerov@gmail.com

В статье описывается разработка многофункционального, многоязычного лингвистического интернет-сервиса на базе структурно-параметрической функциональной модели тюркской морфемы. Основными задачами сервиса являются формирование ресурсной базы для программных продуктов, осуществляющих компьютерную обработку тюркских языков, таких как системы машинного перевода, информационно-поисковые системы, системы разметки электронных корпусов, извлечения данных и др.

В настоящее время реализуется языконезависимая часть сервиса для описания значений языковых единиц. Семантическая часть реализуется в виде онтологической модели, которая в зависимости от типа значения реализуется в виде тезауруса или сети ситуационных фреймов.

Ключевые слова: *многофункциональный интернет-сервис, тюркские языки, языковые единицы, технологический инструментарий.*

ON THE DEVELOPMENT OF A MULTIFUNCTIONAL MULTI-LANGUAGE INTERNET SERVICE ON THE BASIS OF THE TURKIC MORPHEMA

A.B. Almenova[1], A.R. Gatiatullin[2], A.M.Bashirov[2]
[1]Kyzylorda State University named after Korkyt Ata, Kyzylorda, Kazakhstan;

[2]Academy of Sciences of Tatarstan, Kazan, Russia
pr.almenova@mail.ru, ayrat.gatiatullin@gmail.com, a.basheerov@gmail.com

The article describes a multifunctional, multilingual linguistic Internet service based on a structural-parametric functional model of the Turkic morpheme. The main tasks of which helps to form a resource base for software products aimed to perform computer processing of Turkic languages, such as machine translation systems, information retrieval systems, electronic corpora markup and data extraction systems, etc.

The service can be used as an information and reference system, which contains almost complete information on the Turkic language units, namely morphemes, and as a toolbox for turkology researchers. This toolkit can be used for comparative analysis of the Turkic language units and their proximity on the corresponding language levels of the Turkic languages.

Key words: *Multifunctional Internet service, Turkic languages, language units, technological tools.*

ВВЕДЕНИЕ

Сохранение и развитие малоресурсных языков, включая тюркские языки, посредством внедрения их в инфо-коммуникационные технологии, является одной из важнейших междисциплинарных задач, которая решается совместными усилиями специалистов в области информатики, математики и лингвистики.

Создание интернет-ресурсов, интегрирующих опыт исследований и разработок лингвистической активности в компьютеризованном информационном пространстве особо ценно и актуально для близкородственных языков, к которым относятся и языки тюркской группы.

Такая интеграция создаст условия для совместных разработок исследователей и практиков в этой группе, что позволяет направить усилия специалистов на нерешенные проблемы, достичь общего прорыва в области создания технологий для обработки тюркских языков и даже создавать новые технологии обработки информации на основе их лексико-грамматических особенностей.

Многофункциональный лингвистический и ннтернет-сервис представляет собой прагматически-ориентированное структурно-функциональное описание элементов морфологии (Сулейманов, Гатиатуллин, 2003) и 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. Важным свойством данной архитектуры структурно-параметрической функциональной модели тюркской морфемы, является ее открытость, т.е. технология ее реализации позволяет авторизованному пользователю вносить изменения в структуру модели как по горизонтали – удаляя или добавляя подмодели (новый язык),

так и по вертикали – удаляя или добавляя параметры соответствующего уровня в подмодели.

На рис. 1 представлена архитектура структурно-параметрической функциональной модели тюркской морфемы.

Структурно-параметрическая функциональная модель тюркской морфемы является ядром многофункциоанльного, многоязычного интернет-сервиса, архитектура которого представлена на рисунке 2.

Данный многофункциональный интернет-сервис на базе тюркской морфемы представляет собой, с одной стороны, каталог с описанием программных модулей для компьютерной обработки тюркских языков, а с другой стороны, веб-сайт, который предоставляет возможность работать с этим интернет-сервисом (Альменова, 2018).

Реализованы основные программные модули: модуль администрирования, модуль заполнения базы данных, модуль лингвостатистического сравнения близости языков и модуль морфологического анализа, а также структура базы данных на основе структурно-параметрической функциональной модели тюркской морфемы.

В настоящее время продолжается реализация семантических моделей и семантического блока.

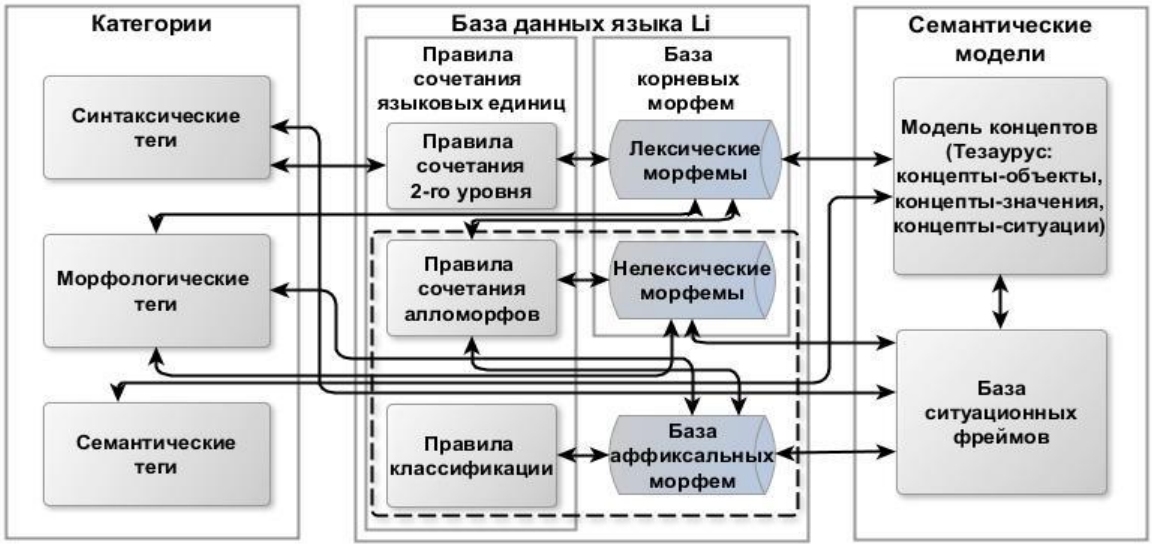


Рис. 1. Архитектура структурно-параметрической функциональной модели тюркской морфемы

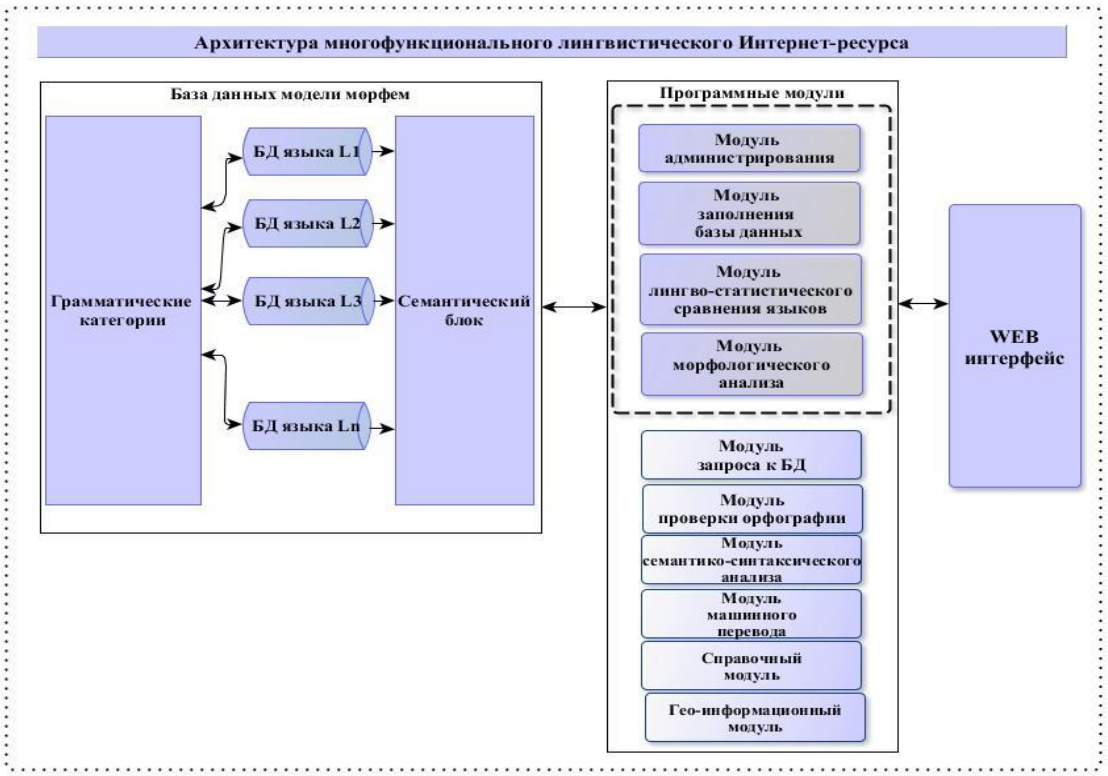


Рис. 2. Архитектура многофункционального лингвистического интернет-сервиса

На рис.1 показано, что связующим элементом моделей описания языковых единиц отдельных языков Li являются семантические модели. В нашей модели для представления семантики используются онтологические модели двух типов. Первый тип — это тезаурусы, а второй тип это реляционно-ситуационные фреймы. Тезаурусы используются для описания значений корневых морфем, выражающих именные сущности. Все концепты взаимосвязаны между собой отношениями гипоним-гипероним и часть-целое. Структура части тезауруса, которая используется для представления именных концептов аналогична структуре тезауруса WordNet. Все именные концепты

Корневые морфемы каждого языка, представленного в модели тюркской морфемы, связываются с концептами тезауруса.

На рис.3 представлен фрагмент связи концептов с описаниями корневых морфем каждого из языков. Связь осуществляется с помощью уникальных идентификаторов концептов (**Concept.ID**). Корневые морфемы также обладают своим уникальным идентификатором, который образуется из 2-х чисел: первое это номер языка и второе номер морфемы в этом языке.

Например:

ID: 01.29137 – морфема татарского языка,

ID: 02.1034 – морфема казахского языка.

Из рис.3, что корневая морфема татарского языка связана с двумя концептами и является многозначной. Корневые морфемы казахского и узбекского языков ссылаются только на один концепт.

Для работы с базой концептов также разработан инструментарий, который позволяет производить поиск по тезаурусу, добавлять новые концепты, редактировать и удалять концепты, имеющиеся в базе данных.

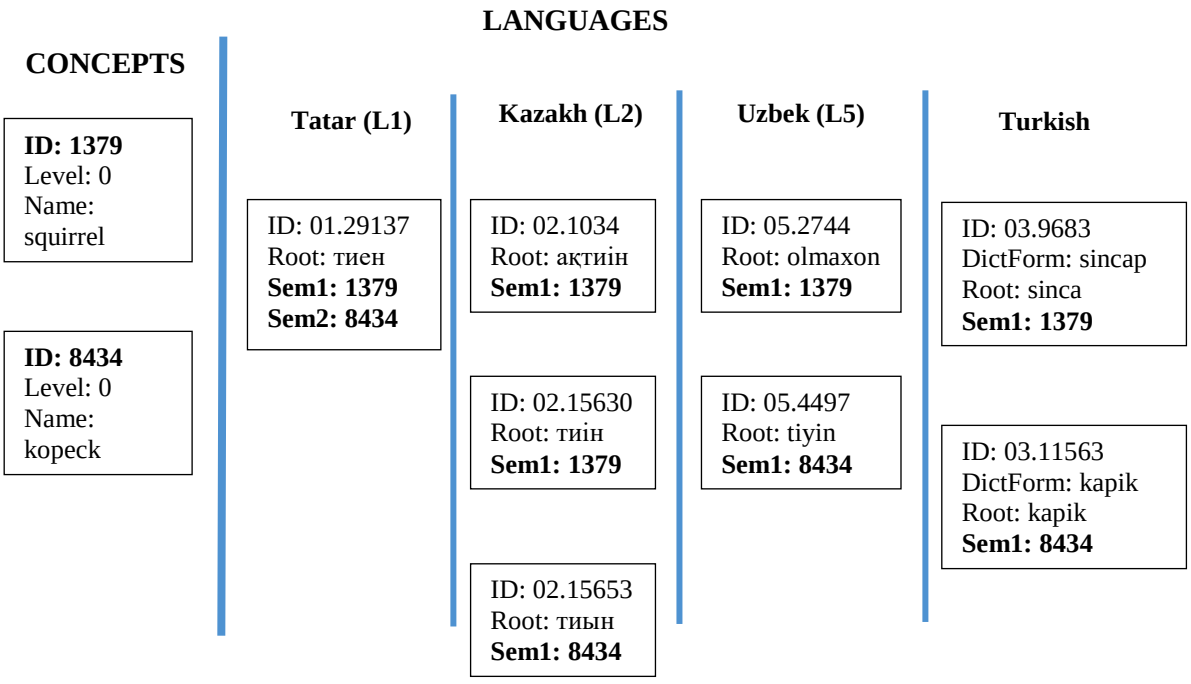


Рис.3. Фрагмент связи концептов с корневыми морфемами

Для работы с именными и ситуационными сущностями используются интерфейсы с разными структурами. На рис.4 представлено окно для редактирования концептов тезауруса, описывающего именные концепты. Для более удобной работы с концептами пользователь, работающий с программой, может видеть все корневые морфемы, которые уже прикреплены к данному концепту.

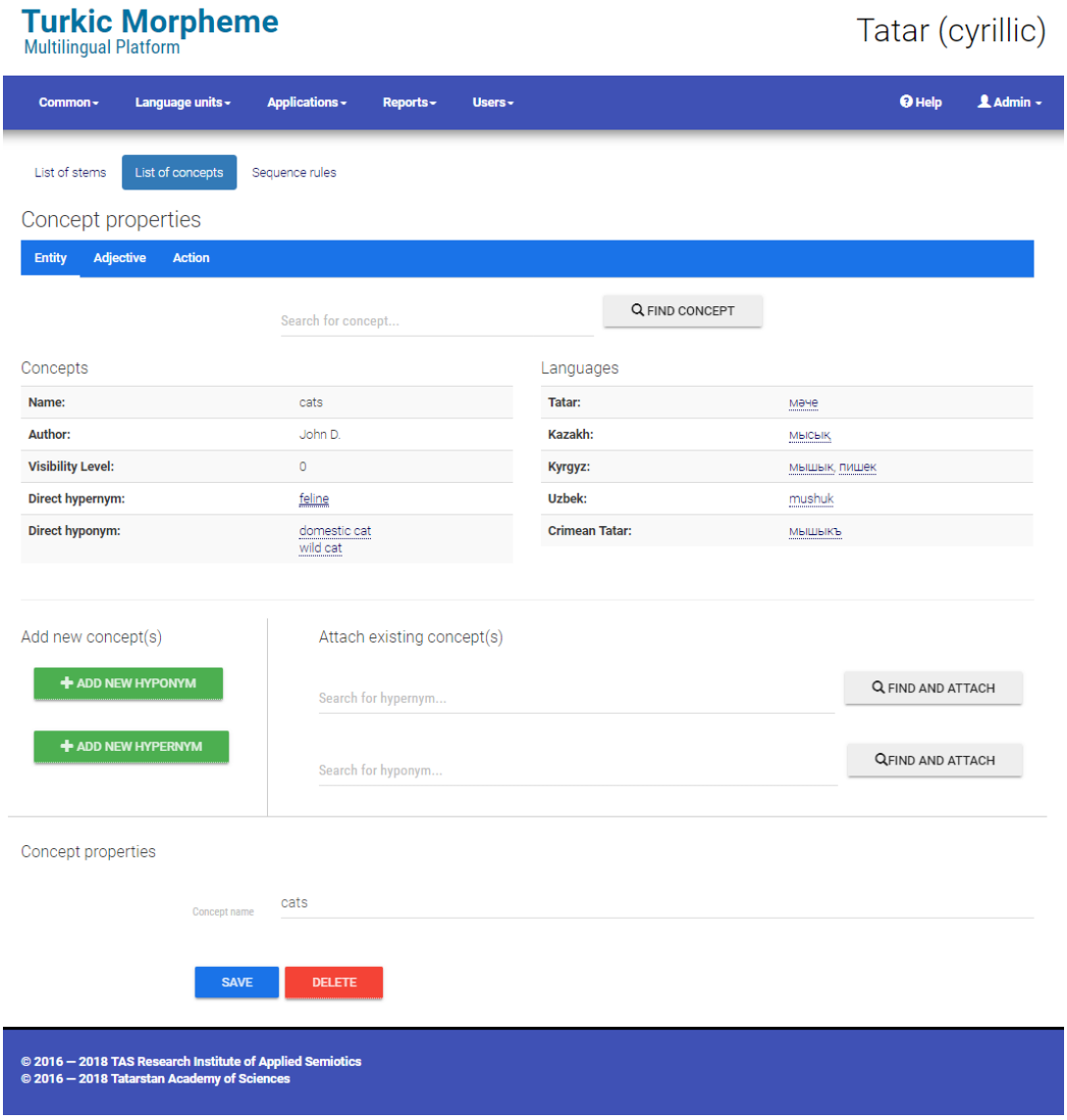


Рис. 4. Окно для заполнения базы концептов

Информация, представленная в тезаурусе, используется в работе расширенного морфологического анализатора (Рис.5.), который является одним из предоставляемых сервисов. Алгоритм работы морфологического анализатора оптимизирован под структуру базы данных многофункционального интерфейса и не требует дополнительных преобразований. Анализатор выделяет не только корневую морфему, но и семантический тип этого концепта. Для определения семантического типа выполняется поиск по тезаурусу. На рис.5 представлен пример анализа, когда на вход анализатора поступает словоформа *тиеннәрне*, а на выходе анализатора информация представляется в следующем виде:

N: mammal:animal (тиен) + PL (-ЛАр) + ACC (-н[ЫІ]).

здесь для основы *тиен ‘белка’* указывается, что это именная сущность (N), обозначающее ***животное*** **(animal):** ***млекопитающее*** **(mammal)**.

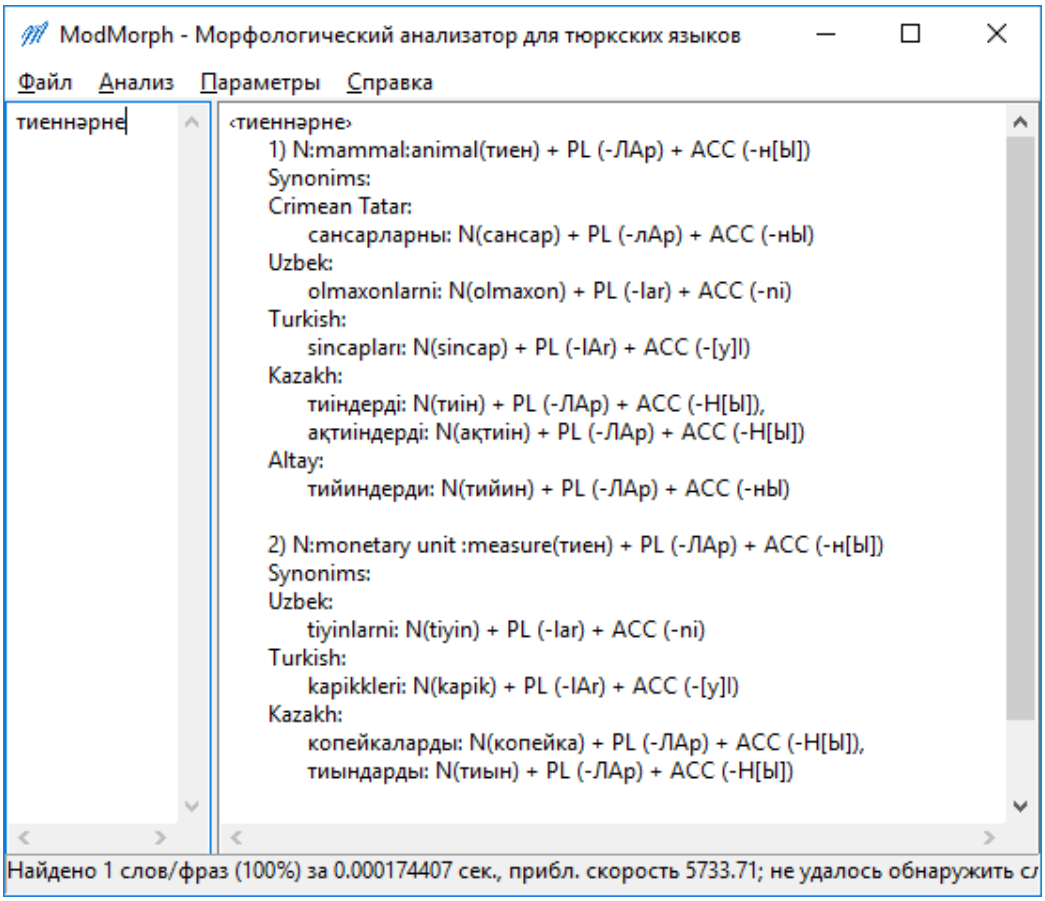


Рис. 5. Морфологический анализатор с выводом семантической информации

Также на рис.5 показано, что вторым концептом, с которым связана корневая морфема тиен ‘копейка’, является концепт типа **N:monetary unit:measure.**

ЛИТЕРАТУРА:

1. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань: Фэн, 2003. – 220с.
2. Альменова, А.Б. Модели и сервисы многофункционального лингвистического интернет-ресурса на базе структурно-параметрической модели тюркской морфемы: диссертация … кандидата технических наук: 05.13.11 / Альменова А.Б. – Казань, 2018. – 191 с.

SO‘Z BIRIKMASI TAHLILIDA MODELLARDAN FOYDALANISH

S. Zarmasov, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti, shakhbozzarmasov1993@mail.ru

Mazkur maqolada so‘z birikmasi tarkibi hamda mazkur tarkibning birikishiga ko‘ra usullarining modellari va modellashtirilish jihatlari yoritib berilgan. Shuningdek, qoliplashtirish jarayonida birikma qismlarining morfologik holatiga ham alohida e’tibor qaratilgan.

Tayanch so‘zlar: *model, modellashtirish, qolip, qoliplashtirish, birikma, so‘z birikmasi, tobe so‘z, hokim so‘z, moslashuv, bitishuv.*

ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ ВАНАЛИЗЕ СЛОВОСОЧЕТАНИЙ

Ш.Зармасов, Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, Ташкент, Узбекистан, shakhbozzarmasov1993@mail.ru

В статье анализируются структурные модели словосочетаний, а также модели анализа словосочетаний. Акцент делается на морфологическом состоянии компонентов словосочетания при моделировании.

Ключевые слова: *модель, моделирование, снятие, формование, сочетание, словосочетание, зависимое слово, главное слово, согласование, примыкание.*

USAGE OF MODELS IN THE ANALYSIS FOR WORD COMBINATIONS

Sh.Zarmasov, [2]Tashkent state university of Uzbek language and literature named after Alisher Navoi, Tashkent, Uzbekistan, shakhbozzarmasov1993@mail.ru

In the article the structure of the word combination, the model of method of their combining and modeling were analyzed. Moreover, it is focused morphological state of combination components during modeling.

Key words: *model, modeling, mold, molding, combination, combination of words, slave, dominant word, adaptation, conciliation.*

Globallashuv davrida murakkablikdan soddalikka, ixchamlikka, tushunarlilik va aniqlikka intilar ekanmiz, ijtimoiy-iqtisodiy, ma’naviy-madaniy, siyosiy va ma’rifiy hayotimizda ham modellashtirishning elementlariga duch kelamiz. Barcha yangilanishlarning negizi bo‘lgan innovatsiyaning ahamiyati ortib borishi bilan modellashtirish ham innovatsion faoliyatning bir turi sifatida hayotimizning har bir burchagiga kirib keldi. Jamiyat taraqqiyotida yetakchilikni o‘z qo‘liga olgan bu jarayon fan-texnika sohasi, xususan tilshunoslikni ham chetlab o‘tmadi. Modellashtirish tushunchasi haqida fikr yuritish, tilshunoslikka modellashtirish metodini tatbiq qilishdan oldin model tushunchasiga izoh keltirib o‘tish lozim.

Model lotincha «modelus» so‘zidan olingan bo‘lib, biror obyekt yoki obyektlar tizimining obrazi yoki namunasidir. Boshqacha aytganda, model obyektlarning o‘xshashi, taqlidiy tashqi ko‘rinishidir. Ko‘pincha biz ularni modellashtirilgan obyektlarning qolipi, andozasi deb ataymiz. Masalan, yerning modeli — globus, osmon va undagi yulduzlar modeli — planetariy ekrani, pasportdagi suratni shu pasport egasining modeli deyish mumkin. Yanada soddaroq qilib aytadigan bo‘lsak, dizayner yaratgan kiyimning qog‘ozdagi chizmasi, yaratilayotgan kiyimning modeli bo‘la oladi.

Har qanday ilmiy tadqiqotni amalga oshirish bosqichma-bosqich amalga oshirilganligi kabi lingvistik birliklarni modellashtirish ham ma’lum bir bosqichlar, ketma-ketlikda amalga oshiradigan qonun — qoidalar tizimidan iboratdir.

Birinchi bosqichda modellashtirilayotgan birliklar haqidagi muayyan bilimlarni to‘plash va ularni tahlil qila olish ko‘nikmalariga ega bo‘lish hamda modellashtirilmoqchi bo‘lgan birliklar o‘rtasidagi munosabatlarni aniqlay bilish lozim. Ikkinchi bosqich esa muammoni modellashtirish bosqichi bo‘lib, u biroz soddalashtiruvchi, ammo bir vaqtning o‘zida uning nisbatan muhim unsurlari, tuzilishi, ichki va tashqi shakllarining yaxlit qabul qilinishini nazarda tutishga mo‘ljallangan tasavvurlarni ishlab chiqishga qaratilgan. Boshqacha aytganda, o‘rganilgan, tahlil qilingan birliklarning modellarini yaratish bosqichidir.

Ma`lumki, har bir tilda so‘zlarning o‘zaro birikishi o‘zining alohida spetsifik xususiyatiga ega. Shunga mos har bir tilda so‘z birikmalarining tuzilish uslublari, strukturasi har xildir. O‘zbek tilining grammatik tizimida so‘z birikmasining bir necha turlari mavjud. Ular so‘z birikmasi komponentlarining o‘zaro sintaktik aloqasiga: ma’lum grammatik vositalar yoki grammatik ko‘rsatkichlarsiz birikishiga, hokim komponentining qaysi so‘z turkumi bilan ifodalanganligiga, funksional xarakteri va tarkibiga ko‘ra bir-biridan tubdan farq qiladi. Birikma modellari ham ana shunday xususiyatlari orqali shakliy jihatdan turli ko‘rinishda bo‘ladi. Bu esa birliklarni bir-biridan farqlay olish imkoniyatini yaratadi.

Ilmiy adabiyotlarda, darslik, o‘quv qo‘llanmalari, monografiya va maqolalarda «Ikki va undan ortiq so‘zlarning grammatik va ma’no jihatdan birikuviga so‘z birikmasi deyiladi. So‘z birikmasi tobe va hokim qismlardan tuziladi», deb ta’rif berilgan. Tobe qism birikmaning grammatik aloqasini ko‘rsatib turuvchi va hokim so‘zda ifodalangan semantik mazmunning konkretlashtiruvchisi sifatida yuzaga chiquvchi blokdir. Turkiy tillar, xususan, o‘zbek tilida tobe va hokim bo‘lak o‘rni qat’iydir. Hokim qism pozitiv, tobe qism esa perepozitiv holatda turadi. So‘roq hokim so‘zga qarab beriladi, so‘roqqa javob bo‘luvchi so‘z esa tobe qism sanaladi. Struktur jihatdan o‘zgarmas tartibga ega ekanlik, birikma qolipini hosil qilishda uning soddaroq ko‘rinishga ega bo‘lishini ta’minlaydi.

So‘z birikmasini qoliplashtirishda dastlab birikma komponentlari; tobe va hokim qismlarni aks ettiruvchi modellar tanlanadi va keyingi bosqichda mazkur qismlarning grammatik jihatdan birikish usullari: bitishuv, boshqaruv, moslashuv aloqalari hamda ular asosida hosil qilingan birliklar, sintaktik aloqani ta’minlovchi grammatik vositalarning turlariga ko‘ra kelishikli va ko‘makchili boshqaruv munosabatli birikmalar qoliplashtiriladi. Shuningdek, so‘z birikmasi strukturasida hokim qismning qaysi turkum bilan ifodalanishiga ko‘ra fe’lli va otli birikma kabi turlari ham mavjud bo‘lib, ularni farqlay olish, hokim qism modelining strukturasiga bog‘liq.

So‘z birikmasi komponentlarining qoliplarini yaratishda ularning birinchi harfini asos sifatida qabul qilamiz. Ya’ni tobe so‘z uchun katta **[T],** hokim so‘z uchun esa katta **[H]** harflari modeldir. Qismlarning o‘zaro bir-biriga tobelik aloqasida bog‘lanishini anglatish uchun tobelanishning modeli sifatida yo‘nalishni ko‘rsatuvchi **[⇒]** belgisi qabul qilinadi. Ushbu modelning tobe so‘zdan hokim so‘zga qarab yo‘nalganligi, hokim komponentning yetakchilik, tobe komponentning esa bo‘ysunuvchilik xususiyati bilan bog‘liq.

Aytilgan fikrlarning yanada tushunarliroq bo‘lishi uchun so‘z birikmasining modellarini jadvallarda berib o‘tamiz:

№	So‘z birikmasi qismlari	Modellar	Joylashish o‘rniga ko‘ra so‘z birikmasi	Joylashish o‘rniga ko‘ra modellar
1	Tobe so‘z	**T**	Tobe ⇒ Hokim	**T ⇒ H**
-	Tobelanish	⇒		
2	Hokim so‘z	**H**		

So‘z birikmasi tahlilida modellashtirish metodini tatbiq qilishda birikmaning sintaktik turlarini ham qoliplash muhimdir. Bu jarayonda an’anaviy usulga tayangan holda model sifatida o‘sha birliklarning birinchi harflari olinadi. Masalan, moslashuv munosabatli so‘z birikmasi **[** M_b **];** bitishuv munosabatli so‘z birikmasi **[** B_b **];** boshqaruv munosabatli so‘z birikmasi esa grammatik vositalarning turiga ko‘ra ikkiga ajraladi va kelishik qo‘shimchalari bilan bog‘langan birikmalar; kelishikli boshqaruv munosabatli so‘z birikmasi **[** KeB_b **]**; ko‘makchilar bilan bog‘langan so‘z birikmasi; ko‘makchili boshquv munosabatli so‘z birikmasi esa **[** $Ko'B_b$ **]** belgilari bilan qoliplanadi. Boshqaruv munosabatli birikma modellarini ichki guruhlar bo‘yicha belgilash, uni bitishuv munosabatli birikma qolipidan farqlash imkonini beradi. Koiffetsentdagi $_b$ esa birikma so‘zining qolipidir.

	SO‘Z BIRIKMASI TURLARI	**MODELLAR**
1	Moslashuv munosabatli so‘z birikmasi	M_b
2	Bitishuv munosabatli so‘z birikmasi	B_b
3.1	Kelishikli boshqaruv munosabatli so‘z birikmasi	KeB_b
3.2	Ko‘makchili boshqaruv munosabatli so‘z birikmasi	$Ko'B_b$

Keyingi jarayon so‘z birikmasi sintaktik turlarining grammatik qoidalari asosida tuzilgan birikmalarning qoliplarini hosil qilishdir. Barcha ilmiy adabiyotlarda, o‘quv qo‘llanmalari, darslik va monografiyalarda so‘z birikmasi tarkibidagi qismlarning birikishiga ko‘ra uch turga: moslashuv, bitishuv va boshqaruv munosabatli so‘z birikmalariga bo‘linishi, bu birliklarning har biri uchun alohida-alohida grammatik qoidalar hamda qismlarni bog‘lovchi grammatik vositalar qayd etilgan. Endilikda, so‘z birikmasi qoliplari ham grammatik qoidalar va grammatik ko‘rsatkichlarning turiga muvofiq bir-biridan farqli ko‘rinishlarga ega bo‘ladi.

Tobe so‘zning hokim so‘zga qaratqich kelishigi **[-ning]**, hokim so‘zning tobe so‘zga egalik qo‘shimchasi **[-i]; [-si]; [-lari]** shakllari orqali bog‘lanishi moslashuvli so‘z birikmalari deyiladi. Masalan, *qushning qanoti, qushning bolasi, bolaning onasi, do‘stimning maktubi, so‘zning sehri, qo‘lning harakati, olmaning guli, uyning tomi* kabi.

Moslashuvli so‘z birikmasida ba’zan tobe qism tarkibida qaratqich kelishigi qo‘shimchasi, ba’zan esa hokim so‘z tarkibida egalik qo‘shimchasi qatnashmasligi, ba’zi hollarda har ikkala komponentda ham grammatik kategoriya ishlatilmasligi ham mumkin. Shunisi ahamiyatliki, mazkur grammatik vositalar birikma tarkibida qatnashmayotgan bo‘lsa-da, nutq vaziyatiga ularni tiklash imkoniyati mavjud bo‘ladi. Grammatik vositalarning belgili yoki belgisiz shaklda kelishi, so‘z birikmasi modellari sonining ortishiga hamda tarkibida

grammatik ko‘rsatkichlar belgili shaklda kelgan birliklar alohida, belgisiz shaklda kelgan birliklar esa alohida qoliplanishiga olib keladi. Masalan, *futbol ustasi, otalar so‘zi, buloq suvi, o‘rik sharbati, navro‘z bayrami, Mashrab g‘azali, Navoiy bayti, raqqosa raqsi* kabi birliklar uchun:

$$Tot_{\emptyset} \Rightarrow Hot_{eq} \Leftrightarrow M_b;$$

o‘rik danak kabi birikmalar uchun esa:

$$Tot_{\emptyset} \Rightarrow Hot_{\emptyset} \Leftrightarrow M_b$$

qoliplari ramz sifatida belgilanadi.

Tobe so‘zning hokim so‘zga hech qanday qo‘shimchalarsiz faqat ohang yordamida bog‘lanishi ilmiy adabiyotlarda bitishuv munosabatli so‘z birikmalari deb ataladi. Tarkibda grammatik kategoriyalarning mavjud emasligi, modellarining birmuncha soddaroq bo‘lishiga zamin yaratadi.

Boshqaruv munosabatli so‘z birikmasini modellashtirish birmuncha murakkablikni talab qiladi. Bu holat tarkibdagi grammatik vositalarning ko‘pligi va xilma-xilligiga bog‘liq. Shu nuqtai nazardan so‘z birikmasining bir turi hisoblangan boshqaruvli so‘z birikmasi: kelishikli va ko‘makchili boshqaruv munosabatli so‘z birikmasi modellarini hamda modellashtirish asoslarini alohida bayon qilamiz.

Yana shuni ham alohida ta’kidlab o‘tish kerakki, so‘z birikmasi qismlari turli so‘z turkumlariga mansub bo‘lishi mumkin. Komponentlarning turli so‘z turkumlari bilan ifodalanishi ham uning murakkabligini yuzaga keltiradi. Chunki, birliklarning har bir shakli uchun alohida model ya’ni qolip tuzish zarur.

Grammatik qoidalar asosida tuzilgan va turli so‘z turkumlari bilan ifodalangan moslashuv hamda bitishuv munosabatli so‘z birikmalarning modellarining tushunarliligiga erishish, modelning qaysi birlik uchun tegishli ekanligini, tarkibidagi qismlarning morfologik shaklini hamda grammatik ko‘rsatkichlarini farqlay olish uchun jadval asosida aks ettiramiz:

BITISHUVLI MUNOSABATLI SO‘Z BIRIKMASI MODELLARI

№	Tobe qism tarkibi	Hokim qism tarkibi	Qiymatsiz model	Qiymat hosil qilingan model	Misollar
1	Ot bilan ifodalangan	Ot bilan ifodalangan	$Tot \Rightarrow Hot$	$Tot \Rightarrow Hot \Leftrightarrow B_b$	Tilla soat Temir sandiq
2	Sifat bilan ifodalangan	Ot bilan ifodalangan	$Tsf \Rightarrow Hot$	$Tsf \Rightarrow Hot \Leftrightarrow B_b$	Beqasam to‘n Hashamdor ziyofat

3	Olmosh bilan ifodalangan	Ot bilan ifodalangan	$Tol \Rightarrow Hot$	$Tol \Rightarrow Hot \Leftrightarrow B_b$	O‘sha bog‘ Anavi kishi
4	Ravish bilan ifodalangan	Ot bilan ifodalangan	$Tr \Rightarrow Hot$	$Tr \Rightarrow Hot \Leftrightarrow B_b$	Do‘stona munosabat Erkakcha qadam
5	Son bilan ifodalangan	Ot bilan ifodalangan	$Tsn \Rightarrow Hot$	$Tsn \Rightarrow Hot \Leftrightarrow B_b$	To‘rtinchi fasl Beshinchi oy
4	Sifatdosh bilan ifodalangan	Ot bilan ifodalangan	$Tsd \Rightarrow Hot$	$Tsd \Rightarrow Hot \Leftrightarrow B_b$	Pishgan gilos So‘ngan umid
5	Ravish bilan ifodalangan	Harakat nomi bilan ifodalangan	$Tr \Rightarrow Hhn$	$Tr \Rightarrow Hhn \Leftrightarrow B_b$	Sekin gapirish Tez yurish
6	Ravishdosh bilan ifodalangan	Harakat nomi bilan ifodalangan	$Trd \Rightarrow Hhn$	$Trd \Rightarrow Hhn \Leftrightarrow B_b$	O‘ylab gapirish O‘ynab o‘tirish
7	Ravish bilan ifodalangan	Fe’l bilan ifodalangan	$Tr \Rightarrow Hf$	$Tr \Rightarrow Hf \Leftrightarrow B_b$	Sekin turmoq Tez yugurmoq
8	Sifat bilan ifodalangan	Ravishdosh bilan ifodalangan	$Tsf \Rightarrow Hrd$	$Tsf \Rightarrow Hrd \Leftrightarrow B_b$	Ohista o‘tirib Chuqur o‘ylab
9	Sifat bilan ifodalangan	Sifatdosh bilan ifodalangan	$Tsf \Rightarrow Hsd$	$Tsf \Rightarrow Hsd \Leftrightarrow B_b$	Qiyg‘os pishgan Yaxshi o‘qigan
10	Ravish bilan ifodalangan	Sifat bilan ifodalangan	$Tr \Rightarrow Hsf$	$Tsf \Rightarrow Hsf \Leftrightarrow B_b$	Onadek aziz Qushdek yengil

MOSLASHUV MUNOSABATLI SO‘Z BIRIKMASI MODELLARI

№	Tobe qism tarkibi	Hokim qism tarkibi	Qiymatsiz model	Qiymat hosil qilingan model	Misollar

1	Ot bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Ot bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tot_{qk} \Rightarrow Hot_{eq}$	$Tot_{qk} \Rightarrow Hot_{eq} \Leftrightarrow M_b$	Olmaning guli Qushning ini
2	Ot bilan ifodalangan, qaratqich kelishigi belgisiz qo‘llangan	Ot bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tot_{\emptyset} \Rightarrow Hot_{eq}$	$Tot_{\emptyset} \Rightarrow Hot_{eq} \Leftrightarrow M_b$	Buloq suvi Olma sharbati
3	Ot bilan ifodalangan, qaratqich kelishigi belgisiz qo‘llangan	Ot bilan ifodalangan, egalik qo‘shimchasi belgisiz qo‘llangan	$Tot_{\emptyset} \Rightarrow Hot_{\emptyset}$	$Tot_{\emptyset} \Rightarrow Hot_{\emptyset} \Leftrightarrow M_b$	O‘rik danak Olma sharbat
4	Olmosh bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Ot bilan ifodalangan egalik qo‘shimchasi belgili qo‘llangan	$Tol_{qk} \Rightarrow Hot_{eq}$	$Tol_{qk} \Rightarrow Hot_{eq} \Leftrightarrow M_b$	Hammaning vazifasi Uning ishi
5	Olmosh bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Ot bilan ifodalangan egalik qo‘shimchasi belgili qo‘llangan	$Tol_{qk} \Rightarrow Hot_{\emptyset}$	$Tol_{qk} \Rightarrow Hot_{\emptyset} \Leftrightarrow M_b$	Bizning uy Sizning qishloq
6	Ot bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Son bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tol_{qk} \Rightarrow Hsn_{eq}$	$Tol_{qk} \Rightarrow Hsn_{eq} \Leftrightarrow M_b$	Ularning biri Sizlarning biringiz
7	Ot bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Ravish bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tot_{qk} \Rightarrow Hr_{eq}$	$Tot_{qk} \Rightarrow Hr_{eq} \Leftrightarrow M_b$	Ko‘prikning usti Hovlining o‘rtasi
8	Ot bilan ifodalangan, qaratqich kelishigi belgisiz	Ravish bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tot_{\emptyset} \Rightarrow Hr_{eq}$	$Tot_{\emptyset} \Rightarrow Hr_{eq} \Leftrightarrow M_b$	Ko‘prik usti Hovli o‘rtasi

	qo‘llangan				
9	Ot bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Harakat nomi bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tot_{qk} \Rightarrow Hhn_{eq}$	$Tot_{qk} \Rightarrow Hhn_{eq} \Leftrightarrow M_b$	Gulning ochilishi Talabalarning ishlashi
10	Ot bilan ifodalangan, qaratqich kelishigi belgisiz qo‘llangan	Harakat nomi bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tot_{\emptyset} \Rightarrow Hhn_{eq}$	$Tot_{\emptyset} \Rightarrow Hhn_{eq} \Leftrightarrow M_b$	Gul ochilishi Talabalar ishlashi
11	Harakat nomi bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Ot bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Thn_{qk} \Rightarrow Hot_{eq}$	$Thn_{qk} \Rightarrow Hot_{eq} \Leftrightarrow M_b$	O‘qishning mashaqqati So‘zlashning qoidasi
11	Harakat nomi bilan ifodalangan, qaratqich kelishigi belgisiz qo‘llangan	Ot bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Thn_{\emptyset} \Rightarrow Hot_{eq}$	$Thn_{\emptyset} \Rightarrow Hot_{eq} \Leftrightarrow M_b$	O‘qish mashaqqati So‘zlash qoidasi
12	Ot bilan ifodalangan, qaratqich kelishigi belgili qo‘llangan	Son bilan ifodalangan, egalik qo‘shimchasi belgili qo‘llangan	$Tot_{qk} \Rightarrow Hsn_{eq}$	$Tot_{qk} \Rightarrow Hsn_{eq} \Leftrightarrow M_b$	Kitoblarning biri Olmalarning biri

Jadvallarda berilgan modellar uchun ba’zi hollarda so‘zning yoki o‘sha birlikning birinchi harfi bilan birgalikda bir necha harflar asos qilib olingan. Bunga sabab bir tovush bilan farqlanuvchi birliklarning modellarini alohida shakllarda ko‘rsatib berishdir. Jumladan, so‘z turkumlaridan sifat, son hamda fe’lning sifatdosh shakllari bir xil **[s]** tovushi bilan boshlangan. Shakliy o‘xshashlikni modellarda farqlash maqsadida sifat — **[sf],** son — **[sn]** va sifatdosh **[sd]** harflari qabul qilingan. Shuningdek, bu farqlash usullarini ravish va ravishdosh, ot va olmosh turkumlarida ham uchratamiz, ya’ni ravish **[r],** ravishdosh **[rd]**; ot — **[ot],** olmosh esa **[ol]** harflari bilan qoliplanadi.

Bitishuv munosabatli so‘z birikmalaridan farqli o‘laroq, moslashuv munosabatli so‘z birikmalarining modellarida koeffitsent tarzida grammatik kategoriyalarning modeli ham berilgan: tobe so‘z tarkibidagi qaratqich kelishigi

qo‘shimchasi — qk, tushum kelishigi — tk , jo‘nalish kelishigi — jk, o‘rin-payt kelishigi — $o'rk$, chiqish kelishigi qo‘shimchasi — chk shakllari, hokim so‘z tarkibida keladigan egalik qo‘shimchasi — eq shakllari bilan qoliplanadi. Birikmalar tarkibida grammatik kategoriyalardan qaratqich va tushum kelishigi ko‘rsatkichlari belgisiz shaklda kelishi, qaratqich kelishigining modeli qk va egalik kategoriyalarining modeli eq shakllarining o‘rnida $\emptyset$ shartli belgisi, kelishini taqazo etadi. Belgisizlikni aks ettiruvchi mazkur qolip qo‘shimchaning umuman mavjud emasligini emas, balki mavjud bo‘lib, tushirib qoldirilganligini hamda uni tiklash imkoniyati mavjudligini ko‘rsatadi.

So‘z birikmalarining tuzilish jihatdan ham modellashtirish dolzarbdir. Bu jarayonlar hosil qilingan modellarning kengayishi natijasida yuzaga chiqadi. Masalan, besh qavatli uy birikmasida tobe qism besh qavatli son va sifatdan iborat bo‘lgan so‘zlardir. Mazkur birliklar turkum doirasida, ya’ni besh soni — **[sn],** qavatli so‘zi — **[sf]** shaklida qoliplanadi. Besh qavatli uy so‘z birikmasining modeli esa $Tsn.sf \Rightarrow Hot$ yoki $Tsn.sf \Rightarrow Hot \Leftrightarrow B_b$ ko‘rinishida bo‘ladi.

Aytish mumkinki, modellashtirish metodidan tilshunoslik sohasida, xususan, grammatik tahlil jarayonida foydalanish bir qaraganda ko‘pchilikning tasavvurida soddalikdan murakkablikka o‘tishdek ko‘rinsa-da, yoritilayotgan mavzuning tushunarlilik va aniqlik darajasini oshirishga xizmat qiladi.

THE NUMERAL MODELING OF SEPARATING UZBEK WORDS INTO SYLLABLES

A.M.Norov, Karshi State University, Karshi, Uzbekistan, nam_71@mail.ru

In this article is said about the numeral modeling of separating the words into syllables belong to assimilated layer from Uzbek and eastern languages in Uzbek language.

Key words and phrases: *syllable, the chain of symbols, the length of word, the phonematic length of word, non-letter orthographic symbol, the chain of phoneme, the syllable boundary, numeral model.*

МОДЕЛИРОВАНИЕ РАЗДЕЛЕНИЯ НА СЛОГИ СЛОВ УЗБЕКСКОГО ЯЗЫКА

А.М. Норов, Каршинский государственный университет. Карши, Узбекистан, nam_71@mail.ru

В статье рассматривается моделирование разделения на слоги слов узбекского языка, которые ассимилированы из языков близкородственных узбекскому и других восточных языков.

Ключевые слова: *слог, цепочка символов, длина слова, фонематическая длина слова, орфографический знак, цепочка фонем, граница слога, численная модель.*

The problem of syllable is one of the most actual and complex issue series of linguistics. Because the importance of the syllable in language and speech is reflected in the followings [1]:

1. The syllable serves as the «building material» and template tasks in formation of the word, in particular the phonetic word.
2. Teaching and writing with separating syllable give effective result in order to teach first-class pupils in the correct reading and writing skills, as well as the formation of correct pronunciation and spelling skills on them.
3. One part of the orthographic rules is based on separating into the syllable.
4. The importance of the syllable poetry is also great in poetry as the means of providing the reams, consonance and musicality in the poetry, namely, the syllable is one of the main elements created the rhythm in poetic couplets, and serves as the measure of the rhythm at that moment.

In addition to it, it can be said that the syllable can serve as an algorithmic base for separating automatically the words to morphemic units.

The problem of separation automatic syllables of words which is one part of it in the sphere of computer-aided language teaching, has been solved positively in many foreign countries. This idea can be argued through several online sites (http://www.youryoga.org, https://www.wordsmyth.net, https://rifma.poncy.ru, http://soft-arhiv.com, http: /slogi.su, http://perenosslov.ru, http://enjoy-eng.ru,). However, the scientific research has yet been made for the words in Uzbek language.

When we say about the syllable, first of all, the word will be researched by phonetic aspects.

If we consider that the word is some kind of **the chain of symbols**, from a mathematical linguistic point of view (here, we do not give attention to its meaning character), these chain of symbols, naturally form a sequence of different combinations of vowel and consonant sounds. For example, the word *chiroq* (lamp) has 6 symbols and 5 sound.

Therefore, **the word length** is usually understood as the number of symbols which comprised it, but the principle of «one sound-one letter» is used in **the mathematical or numerical modeling of** separating into the syllable phenomenon. According to this context, we add the term «phonematic length of the word» using the term «the word length». So, **the phonematic length of the word** means the total number of vowel and consonant sounds comprised it, and the non-letter orthographic symbols.

For example, if we consider n is the phonematic length of the word which has p vowel sound, q consonant sound and r non-letter orthographic symbol, then the following expression can be written: $n = p + q + r$.

Only one apostrophe (’) is used instead of r non-letter orphographic symbol in words of Uzbek language, and usually it participates only once in the composition of any word. Therefore, r dimensions accept only one of two values as the variable: 0 or 1. Namely, apostrophe (’) is taken $r = 1$ for the case which participated in the word, $r = 0$ for the case which it is not participated in the word.

For example, the word *sur’at* contains two vowel sounds, three consonants, and one non-letter orphographic symbol, here $p = 2$, $q = 3$ and $r = 1$ are taken, so $n = 6$.

Thus, in general, the following expression is written for a word which its phonematic length is equal to n:

$$l(W) = l(W(V^p, C^q, X^r)) = n \qquad (1)$$

Here: $l(W)$ - the phonematic length of given word; p — the number of vowel (V) sounds in the word; q — the number of consonants (C) in the word; r — the number of non-letter orphographic symbols (X) in the word.

When we act the linguistic criterions mentioned directly by M.Mirtojiev and Sh.Imyaminova for building the mathematical (numerical) model of separating the syllables of words, so we consider to give the opinions of these authors about it one by one in this case:

«Only first syllable in the words belong the lexicon of the Uzbek language may not have the beginning of syllables.

... All syllables except it have beginning of syllables, it begins with consonant.

… In separating syllable of word, beginning of syllables of next syllables is given as one consonant. This is a strict law.

... The end of the syllable in the syllable of word belongs to its lexicon has been never seen in the case of exceeds two consonants» [2].

If we take into account to the syllable composition of the root words which are equal to a single syllable analyzed by M.Mirtojiev, then the syllabus model for such monosyllabic words is generally in the form CVCC, it appears in several specific cases in real term (Table 1).

Table 1.The specific cases specific to CVCC template

№	C	V	C	C	Exemplary independent words (In Latin)
1	0	1	0	0	*A?* (interrogative word)
2	0	1	0	1	***a'l*** part of the word *La'l*
3	0	1	1	0	*Oq, osh, …*
4	0	1	1	1	*Adr, abr, …*
5	1	1	0	0	*Ma, va, …*
6	1	1	0	1	*Sha'n, sa'y, …*
7	1	1	1	0	*Non, tun, …*
8	1	1	1	1	*Qadr, sabr, …*

In this table, having the value equal to 1 or 0 of C consonant indicates whether its participation in general template is existed or is not existed. V vowel sound always accepts a value equal to1 as the condition of creating the syllable and it means that it must be continuously existed in the general template.

Also, Sh.Imyaminova gives the syllable boundary rules for the words in the Uzbek language [3]:

1. Only the self-contained layers words of the Uzbek language can begin with the vowel sounds. For example: *ol-din, o-dam, in-son*.

2. When one syllable comes between two vowel sounds in multisyllable (polysyllabic) words in Uzbek language, the next syllable will be noted from the consonant. For example: *ke-ta-di*.

3. When adjacent consonants come between two vowel sounds in multi-syllable words, the syllable will be separated between two consonants. For example: *un-dosh-lar.*

4. If germination phenomenon (double consonants) happens in multi-syllable words, the syllable passes through the middle of the geminate. For example: *kat-ta.*

5. When two consonants between two vowels in multi-syllable words and *я, ю, е, ё* graphemes are given before the second consonant, the syllable will be separated between the first consonant and the second vowel. This phenomenon often occurs in composite words. For example: *музёрар – муз-ё-рар (muzyorar — muz-yo-rar).*

6. When more than two consonants come between two vowels in two-syllable words, the syllable will be separated before the last consonant of them. For example: *йирт-моқ, сурт-моқ* (*yirt-moq*, *surt-moq*).

The mathematical model of separating the syllable of word is based on the principle of determining the boundaries of each next syllable from the second syllable of the word, and it will be created on the basis of the following sequence:

1. We can reform $W(\alpha_1\alpha_2...\alpha_n...\alpha_s)$ word given on the basis of particular chain of symbols on the basis of the chain of sounds $W(\varphi_1\varphi_2...\varphi_n)$.

$$W(\alpha_1\alpha_2...\alpha_n...\alpha_s) \to W(\varphi_1\varphi_2...\varphi_n),\ n \le s. \qquad (2)$$

2. We define the order numbers of the vowel sounds in the word:

$$\underbrace{j, k, ..., l}_{m}, \text{ here } m \le n. \qquad (3)$$

3. We draw $A = (j, k, ..., l)$ linear matrix with dimension $1 \times m$ on the basis of the previously defined number of the vowel sounds in words in order to determine the syllable boundaries (for instance, the initial of each next syllable from the second).

Then we enter the marks as $j = a_1, k = a_2, ..., l = a_m$ for the elements of A linear matrix, and then we will create a new $A^* = (a_1, a_2, ..., a_m)$ matrix.

We also mark $a_i + a_{i+1} = \sigma_i$ and $a_{i+1} - a_i = \delta_i$ for simplicity, and we do the following calculations:

$$b_{i+1}=\begin{cases}\dfrac{\sigma_i}{2}+\left\{\dfrac{\sigma_i}{2}\right\}+\dfrac{2\delta_i-5}{4}+\dfrac{1}{4}\cdot(-1)^{\delta_i},\ if\ \ \delta_i>3,\\ \dfrac{\sigma_i}{2}+\left\{\dfrac{\sigma_i}{2}\right\},\ if\ \ 0\le\delta_i\le 3,\quad i=\overline{1,\ m-1}.\end{cases} \qquad (4)$$

Here b_2, b_3 and b_m (b_{i+1}, $i=\overline{1,\ m-1}$) respectively, represent the initial of the 2nd, 3rd and final *m*-syllables.

4. And, finally, φ_{b_2}, φ_{b_3}, … φ_{b_m} sounds (symbols) in b_2, b_3, ..., b_m - places are replaced accordingly by «$-\varphi_{b_2}$«, «$-\varphi_{b_3}$«, …, «$-\varphi_{b_m}$« sounds (symbols).

If ’ (apostrophe) comes after vowel or consonant in some kind of word (for example, *sur’at* or *ra’no*), then this mark will automatically be regarded as the end of the previous syllable, and the rest of the word will be checked for the syllable.

It should be noted that the apostrophe which came after the vowels in word, serves to determine the syllable boundary as the task of consonant sound, if it comes before the vowels and after consonants, it will serve as the task of two consonant sounds.

It is necessary to preserve their original features for separating into syllable the words which assimilated from European languages through Russian language in the lexicon of the Uzbek language. For example, the word «samolyot» («airplane») is separated into syllables in the form «sa-mol-yot» in its original position, so it is true that if this word is separated into the syllable not in the form «sa-mol-yot», but in the form «sa-mo-lyot».

In such cases, given word is translated to its original form in the Cyrillic script, and then it is separated into syllables in this form, it is expedient to retranslate it into Latin in the form «sa-mo-lyot».

It is possible to conclude from that, it is necessary to use the transliteration, which fully embodies the spelling rules in Cyrillic and Latin scripts, in order to make unmistakably to separate all words in the lexicon of the Uzbek language into syllables.

Thus, the software belongs to separating the Uzbek words into syllables on the basis of this lingo-mathematical model and moving them to new line, is created, «The main spelling rules of Uzbek language» accepted in 1995 and the linguistic model given by M.Mirtojiev (Mirtojiev, 2013) are the base for them.

The software has a very simple interface and its overall look is as follows (Figure 1):

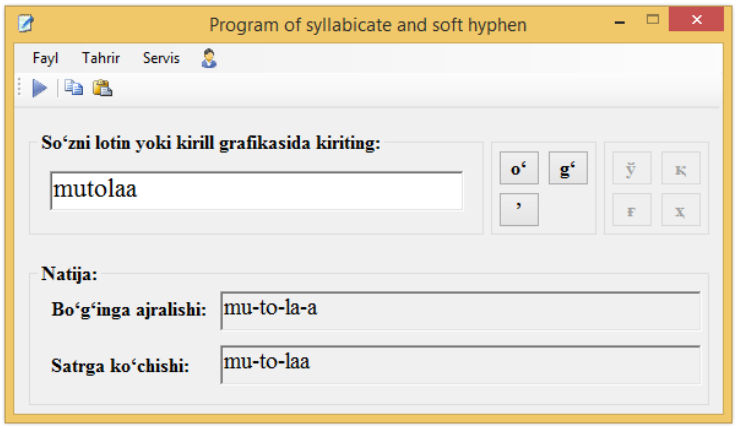


Figure 1. The software separating the Uzbek words into syllable and moving to new line

This Windows application has 1.86 MB (in the disk: 1 953 792 bytes) file size and it works on all Windows 7, Windows 8/8.1, and Windows 10 operating systems.

LITERATURES:

1. Жамолхонов Ҳ. Ўзбек тилининг назарий фонетикаси. – Т.: Фан, 2009. – 224 б.
2. Миртожиев М.М. Ўзбек тили фонетикаси. – Т.: Фан, 2013. – 424 б.
3. Ш.Имяминова. Немис ва ўзбек тилларида бўғин ҳосил бўлиши. – Т.: 2010. – 88 б.
4. Пиотровский Р.Г и др. Математическая лингвистика. – М.: Высшая школа, 1977. – 383 стр.

О ПАРНЫХ СОЧЕТАНИЯХ ИМЕННЫХ ГРАММАТИЧЕСКИХ АФФИКСОВ БАШКИРСКОГО ЯЗЫКА

З. А. Сиразитдинов, Институт истории, языка и литературы УНЦ РАН, г. Уфа, Россия, sazin11@mail.ru

В статье рассматриваются парные сочетания именных грамматических аффиксов башкирского языка. Показывается, что описание любой именной словоформы башкирского языка через парные сочетания морфем может служить основой при разработке систем автоматической переработки текстов.

***Ключевые слова**: башкирский язык, прикладная лингвистика, грамматика, морфология.*

ABOUT THE COMBINATION OF NOMINAL GRAMMATICAL AFFIXES OF THE BASHKIR LANGUAGE

Z. A. Sirazitdinov, Institute of History, Language and Literature of UC RAS, Ufa, Russia, sazin11@mail.ru

The article discusses paired combinations of nominal grammatical affixes of the Bashkir language. It is shown that the description of any nominal word form of the Bashkir language through paired combinations of morphemes can serve as a basis for the development of automatic text processing systems.

***Keyword:** Bashkir language, applied linguistics, grammar, morphology.*

Формальное описание словоформы является является актуальной задачей прикладной лингвистики, служит основой при разработке систем автоматической переработки текстов. Выявление и описание возможных парных сочетаний аффиксов может сыграть роль в этом направлении.

Башкирские языковеды, следуя общей лингвистической традиции, выделяют словообразовательные и грамматические аффиксы [1: 18]. Последние подразделяются на словоизменительные и формообразовательные. Однако среди лингвистов до сих пор нет единого мнения о принадлежности некоторых аффиксов к той или иной категории. Так ряд языковедов к грамматической категории относят аффиксы категорий множественности, отрицания, залога, наклонения и времени, неличных форм глагола, степени и качества, падежа, принадлежности, сказуемости, субъективной оценки [2;3;4].

М.Х.Ахтямов, перечисляя словоизменительные и формообразовательные аффиксы, не включает в их разряд аффиксы субъективной оценки [1:с.138].

В академической грамматике, в вузовских и школьных учебниках аффиксы вопросительности, усиления и неопределенности рассматриваются в составе частиц [5]. Видимо, здесь сыграл роль авторитет Н.К.Дмитриева, утверждавшего, что эти частицы не выражают чисто морфолого-синтаксическую функцию, а скорее синтаксическую [6: 110.]. Однако отдельные лингвисты считают целесообразным рассматривать их в составе формообразующих аффиксов, так как в формальном отношении они являются не отдельными словами, а частями слов [2: 25].

Несмотря на отдельные разногласия, в башкирском языкознании понятие словообразовательной категории в целом имеет сформированные границы функционирования. К данной категории относятся аффиксы, которые, присоединяясь к корню или основе слова, изменяют его лексическое значение и образуют новое слово. Эти аффиксы в языке представлены тремя видами: а) аффиксы, преобразующие в процессе словообразования данное слово в другую часть речи; б) аффиксы, образующие слова в той же части речи, к которой относилась производящая основа; в) многофункциональные аффиксы. [2: 20].

Учитывая высказывания большинства башкирских языковедов, нами под грамматическими аффиксами понимается все регулярные аффиксы словоизменения, формообразования и частицы, которые не создают словоформы с новыми значениями.

В академической грамматике башкирского языка для именных форм выделяются 15 следующих категорий словоизменения и формообразования [6]:

1 – категория множественного числа (Pl),

2 – категория падежного склонения (Case),

3 – категория сказуемости (Pred),

4 – категория принадлежности (Poss),

5 – категория вопросительности (Q),

6 – категория неопределенности (Indf: частица неопределенности -дыр/-дер),

7 – категория усиления (Int: усилительно-утвердительная частица -сы/-се,),

8 – категория притяжательности (PssAtr: аффикс -дыҡы/-деке),

9 – категория уменьшительно-ласкательности (Dmn: аффикс -ҡай/-кәй),

10 – категория уподобления (Comp1: аффикс -дай/-дәй, Comp2: аффикс -са/-сә),

11 – категория атрибутивного локатива (LocAtr: аффикс -тағы/-тәге)

12 – категория обладательности (Poss: аффикс -лы/-ле),

13 – категория лишительности (Abs: аффикс -һыҙ/-һеҙ),

14 – категория предельности (Term: аффикс -ғаса/-гәсә),

15 – категория сравнительной степени ((DgCom: аффикс -раҡ/-рәк).

Аффиксы рассмотренных категорий могут присоединяться непосредственно к основе, далее к этим аффиксам могут добавляться другие. Для именных частей речи башкирского языка существует определенный порядок следования и сочетания грамматических морфем между собой.

1. Аффикс множественности может сочетаться со всеми аффиксами, при этом данный аффикс стоит в подавляющем случае перед всеми остальными аффиксами, реализует парные сочетания:

1.1 Pl+Case: балаларҙың (Pl+Gen), балаларға (Pl+Dat), балаларҙы (Pl+Acc), балаларҙан (Pl+Abl), балаларҙа (Pl+Loc).

1.2. Pl+Pred: балаларбыҙ (Pl+Pred: pl p1), балаларһығыҙ (PL+Pred: pl p2).

1.3. Pl+Poss: балаларым (Pl+Poss: sg p1), балаларың (Pl+Poss: sg p2), балалары (Pl+Poss: sg p3), балаларыбыҙ (Pl+Poss: pl p1), балаларығыҙ (Pl+Poss: pl p2), балалары (Pl+Poss:pl p3).

1.4. Pl+Q: балалармы.

1.5. Pl+Indf: балаларҙыр.

1.6. Pl+Int: балаларсы.

1.7. Pl+PssAtr: балаларҙыҡы.

1.8. Pl+LocAtr: балаларҙағы.

1.9. Pl+Abs: балаларһыҙ.

1.10. Pl+Term: балаларғаса.

В трех случаях данный аффикс сочетается в обратном порядке:

1.11. Comp+Pl: тауҙайҙар

1.12. Dmn+Pl: балаҡайҙар.

1.13. Comp+Pl: (для имен прилагательных): ҡыҙылыраҡтар

2. Аффиксы падежного склонения могут сочетаться со всеми аффиксами и частицами. При этом они употребляются перед частицами, но после аффиксов, за исключением аффикса сказуемости.

2.1. Case+Pred: баламын (Case_Nom+Pred), ауылдамын (Case_Loc+Pred), ауылданмын (Case _bl+Pred), ауылғамын (Case_Dat+Pred).

2.2. Case+Q: баланыңмы (Case_Gen+Q), балағамы (Case_Dat+Q), баланымы (Case_Acc+Q), балананмы (Case_Abl+Q), балаламы (Case_Loc+Q).

2.3. Case+Indf: баланыңдыр (Case_Gen+Indf), баланылыр (Case_Acc+Indf), балағалыр (Case_Dat+Indf), баланандыр (Case_Abl+Indf), балалалыр (Case_Loc+ Indf).

2.4. Case+Int: баланыңсы (Case_Gen+Int), баланысы (Case_Acc+ Int), балағасы (Case_Dat+Int), баланансы (Case_Abl+Int), балаласы (Case_Loc+Int).

2.5. Case+Com: ситкәрәк (Case_Dat+Com), ситтәнерәк (Case_Abl+Com), ситтәрәк (Case_Loc+ Com).

Из-за требования к объему статьи, остальные модели словоизменения и формообразования здесь не приводятся. На основе анализа порядка следования аффиксов словизменения по материалам существующих корпусов башкирского языка нами составлены парные сочетания именных аффиксов.

Результат приведен в таблице 1.

Таблица 1.

ПАРНЫЕ СОЧЕТАНИЯ ИМЕННЫХ СЛОВОИЗМЕНИТЕЛЬНЫХ АФФИКСОВ

Pl	Case		Poss	Term		Comp	Q		Poss	Com
Pl	Pred		Poss	DgCom		Comp	Indf		Abs	Pl
Pl	Poss		Q	Indf		Comp	Int		Abs	Case
Pl	Q		PssAtr	Pl		Comp	PssAtr		Abs	Pred
Pl	Indf		PossAtr	Case		Comp	LocAtr		Abs	Poss
Pl	Int		PossAtr	Pred		Loc	Pl		Abs	Q
Pl	PssAtr		PossAtr	Q		Loce	Case		Abs	Indf
Pl	LocAtr		PossAtr	Indf		LocAtr	Poss		Abs	Int
Pl	Abs		PossAtr	Int		LocAtr	Q		Abs	PssAtr
Pl	Term		PossAtr	Comp		LocAtr	Indf		Abs	Dmn
Case	Pred		Dmhp	Pl		LocAtr	Int		Abs	Comp
Case	Q		Dmn	Case		LocAtr	PssAtr		Abs	LocAtr
Case	Indf		Dmn	Pred		LocAtr	Comp		Abs	DgCom
Case	Int		Dmn	Poss		LocAtr	DgCom		Term	Q
Case	DgCom		Dmn	Q		LocAtr	PredQ		Term	Indf
Pred	Q		Dmn	Indf		Poss	Pl		Term	Int
Pred	Indf		Dmn	Int		Poss	Case		Com	Pl
Pred	Int		Dmn	PssAtr		Poss	Pred		Com	Case
Poss	Case		Dmn	LocAtr		Poss	Poss		Com	Pred
Poss	Pred		Dmn	Poss		Poss	Q		Com	Poss

Poss	Q		Dmn	Abs		Poss	Indf		Com	Q
Poss	Indf		Dmn	Term		Poss	Int		Com	Indf
Poss	Int		Comp	Pl		Poss	PossAtr		Com	Int
Poss	PossAtr		Comp	Case		Poss	Dmn		Com	PssAtr
Poss	Comp		Comp	Pred		Poss	Comp		Com	Comp
Poss	LocAtr		Comp	Poss		Poss	LocAtr		Com	Abst

Как видно из таблицы 1, парные сочетания образуют определенный порядок и могут быть использованы как модели образования словоформ из основ в целях автоматического морфологического описания любой словоформы языка. Проиллюстрируем это на следующих примерах:

1. Словоформа кешеләребеҙҙе формально на основе списка грамматических аффиксов и словаря основ может быть разбита на морфы как кеш +е+ләр+е+беҙ+ҙе и как кеше+ләр+ебеҙ+ҙе. Однако парные сочетания исключают членение в виде сочетания аффиксов Poss+Pl+Poss+Pred+Case и допускают членение в виде сочетания аффиксов Pl+Poss+Case.

2. Словоформа балыҡсыға может быть формально разбита на морфы балыҡсы + ға (Case) и балыҡ+сы+ға (Int+Case). Модель парных сочетаний аффиксов именных частей речи исключает второй вариант.

Таким образом моделирование башкирской словоформы парными сочетаниями может найти применение в разработке новых вариантов морфоанализатора и спеллера национального языка.

ЛИТЕРАТУРА:

1. Әхтәмов М. Х. Хәҙерге башҡорт теле: һүҙъяһалыш. Өфө, 2000. 153 б.
2. Ишбаев К.Г. Башҡорт теленең һүҙъяһалышы. Өфө: УНЦ РАН, 1994. 284 с.
3. Абдуллина Г.Р. Словоизменение башкирского языка. Уфа: Гилем, 2008. 124 с.
4. Абдуллина Г.Р. Формообразование башкирского языка. -Уфа: Гилем, 2008. 236 с.
5. Грамматика современного башкирского литературного языка (под. ред. А.А.Юлдашева). -М.: Наука, 1981. – 795 с.
6. Дмитриев Н.К. Грамматика башкирского языка. М.:Наука, 1948. 276 с.

O‘ZBEKCHA MATNLARNI OVOZLASHTIRISH DASTURINING LINGVISTIK TA’MINOTINI ISHLAB CHIQISHDA AYRIM MASALALAR TADQIQI

N. Abduraxmanova,[11]NamDU huzuridagi xalq ta’limi xodimlarini qayta tayyorlash va ularning malakasini oshirish hududiy markazi, Toshkent, O‘zbekiston, anazokat@inbox.ru

Maqolada o‘zbekcha matnlarni ovozlashtirish dasturining lingvistik ta’minotini ishlab chiqishda tadqiq etish kerak bo‘lgan ayrim masalalar tahlili keltirilgan. Logografemalar, tinish belgilarini ovozlashtirish jarayonida yuzaga keladigan muammolar atroflicha o‘rganilgan va ularning yechimi ko‘rsatilgan.

Tayanch so‘zlar: *o‘zbekcha matnlarni ovozlashtirish dasturi, lingvistik ta’minot, tahlil, logografemalar, tinish belgilari.*

ИССЛЕДОВАНИЕ НЕКОТОРЫХ ПРОБЛЕМ СОВЕРШЕНСТВОВАНИЯ ЛИНГВИСТИЧЕСКОГО ОБЕСПЕЧЕНИЯ ПРОГРАММЫ ОЗВУЧИВАНИЯ УЗБЕКСКИХ ТЕКСТОВ

Н. Абдурахмонова, [11]Региональный Центр переподготовки и повышения квалификации сотрудников Народного образования при Наманганском государственном университете, Наманган, Узбекистан, anazokat@inbox.ru

В статье рассматриваются особенности создания электронных программ озвучивания текстов на узбекском языке и некоторые проблемы, которые необходимо учитывать при их анализе. Также в статье описываются логографемы и пунктационные знаки.

Ключевые слова: *программа озвучивания узбекских текстов, лингвистическое обеспечение, анализ, логографемы, пунктуационные знаки.*

SOME ISSUES IN THE DEVELOPMENT OF THE LINGUISTIC DATABASE OF THE TEXT TO SPEECH IN UZBEK

N. Abdurakhmonova,[11]Regional center of training and retraining public education staff under Namangan State university, Namangan, Uzbekistan, anazokat@inbox.ru

In the article it is written about the peculiarities of creating the linguistic base of the program Text to speech in Uzbek and discussed some problems which is necessary to take into consideration to develop it. The problem which has come into the process of text to sound such as logo graphemes, punctuation marks are analyzed and their solutions are also given.

Key words: *text to speech program in Uzbek, linguistic base, analyses, logo graphemes, punctuation marks.*

Kompyuter dasturlari bilan ishlash bugungi kunda insonlarning birlamchi ehtiyojiga aylanib bormoqda. Bu jarayonda ularning o‘zbek tilidagi variantini ishlab chiqish shu tilda so‘zlashuvchilarning dasturdan foydalanish imkoniyatini oshiradi, desak mubolag‘a bo‘lmaydi. Dasturlarning o‘zbek tilida yaratilishi uchun esa, eng avvalo, texnik ta’minotdan tashqari lingvistik ta’minotni ham mukammal ishlab chiqish talab etiladi. O‘zbekcha matnlarni ovozlashtirish dasturini ishlab chiqishda unga lingvistik ta’minot tayyorlash jarayonida tadqiq etish zarur bo‘lgan ayrim masalalar ham borki, ularni quyida tahlil etamiz.

Logografemalar. O‘zbek tilida son so‘z turkumiga oid so‘zlar yozuvda ma’lum belgilar bilan ifodalanadi. Bunday belgilar logografemalardir. Logografemalarning bu yozuv tipi tushunchalarni yoki tushuncha nomi bo‘lgan so‘zlarni yozuvda ifodalashga asoslanadi. Bunday belgilar aslida ideografik (semasiografik) yozuv tipiga mansub birliklar sanaladi, ammo ulardan fonografik yozuv tarkibida ham foydalaniladi. Yozuv tizimining murakkab optik-grafik sistema ekanligi ham shundan[1].

Tilshunoslikka oid adabiyotlarda logografemalar 2 turga bo‘linadi (bunda odatda raqam grafemalar nazarda tutiladi):

1. Arab raqamlari: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 100, 10 000, 100 000, 1000 000 000 kabi.
2. Rim raqamlari: I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI, XVII, XVIII, XIX, XX, XXX, XL, L, LX, LXX, LXXX, XC, C, CD, D, DC, CM, M, MC, MD kabi.

Bu belgilar I (bir), V (besh), X (o‘n), L (ellik), C (yuz), D (besh yuz), M (ming) kabi bir necha belgilarning kombinatsiyasiga asoslangan.

O‘zbekcha matnlarni ovozlashtirish dasturi logografemalarni o‘qishi uchun yuqorida keltirilgan arab hamda rim raqamlarini dasturning lingvistik ta’minotiga ham yozma, ham ovozli tarzda kiritish lozim bo‘ladi. Shuningdek, bu belgilar son so‘z turkumi doirasida o‘rganilgani uchun ham sonning grammatik ma’no turlarini o‘rganib chiqish va ularning qay tarzda dastur ta’minotiga kiritilishini ko‘rib chiqishimizga to‘g‘ri keladi. Son grammatik ma’nosiga ko‘ra miqdor va tartib sonlarga tasniflanadi. Miqdor son ham o‘z ichida sanoq, dona, chama, jamlovchi, taqsim sonlarga bo‘linadi. O‘zbekcha matnlarni ovozlashtirish dasturi sanoq, dona, chama, jamlovchi sonlarni o‘qish imkoniyatiga ega bo‘lishi mumkin. Lekin taqsim sonning turlari bo‘lgan kasr hamda aralash sonlarni o‘qishda muammo yuzaga keladi: $\frac{2}{5}$, 2/5 (beshdan ikki yoki ikki taqsim besh), $3\frac{2}{10}$; $3{,}2$ (uch butun o‘ndan ikki). Dastur kasr va aralash sonlarni o‘qishda **taqsim, butun** kabi so‘zlarni avtomatik tarzda kasr chizig‘i yoki butunni anglatuvchi ifoda uchun qo‘llay olishi maqsadga muvofiq bo‘ladi. Buning uchun esa kasr chizig‘i kelgan o‘rinlarda **taqsim** so‘zini va aralash son kelgan o‘rinlarda **butun** va **o‘ndan, yuzdan, mingdan, o‘n mingdan** va h.k. so‘zlarni qo‘llash uchun

[1] Jamolxonov H. Hozirgi o‘zbek adabiy tili. T.: Talqin, 2005. B – 98.

ularni lingvistik ta’minotga kiritish lozim bo‘ladi.

Sonlar bilan yuzaga keladigan yana bir masalalardan biri tartib sonlarning talaffuzi bilan bog‘liq. Ma’lumki, tartib sonlar yozuvda son bilan ifodalanganda bunday sonlardan so‘ng chiziqcha qo‘yiladi va shu chiziqcha talaffuzda –(i)nchi tarzida aytiladi: 5-sinf, 6-qator, 17-uy (beshinchi sinf, oltinchi qator, o‘n yettinchi uy).

O‘zbekcha matnlarni ovozlashtirish dasturi tartib sonlardagi chiziqchani juft sonlar, juft so‘zlardagi defisdan farqli o‘qishi uchun bu dasturga lingvistik ta’minot ishlab chiqish mobaynida bunday masalalarga e’tibor qaratish va muayyan qoidalarni kiritish muhim deb o‘ylaymiz.

Nutqimizda sanoq sonlar hisob so‘zlari bilan birga qo‘llaniladi. Ayrim hisob so‘z turlari yozuvda qisqargan shaklda qo‘llaniladi.

1) Uzunlik o‘lchovlari: metr (m), kilometr (km), detsimetr (dm), santimetr (sm), millimetr (mm);

2) Yuza o‘lchov birliklari: kvadrat kilometr (kv.km), kvadrat metr (kv.m), kvadrat detsimetr (kv.dm), kvadrat santimetr (kv.sm), gektar (ga), ar (a);

3) Og‘irlik o‘lchov birliklari: tonna (t), sentner (s), kilogramm (kg), gramm (g), milligramm (mg);

4) Hajm o‘lchov biriklari: kub metr (kub.m), kub detsimetr (kub.dm), kub santimetr (kub.sm), litr (l), gektolitr (gl);

5) Vaqt o‘lchov birliklari: soat (s), minut (min.), sekund (sek.);

6) Elektron texnika xotira sig‘imining o‘lchov birliklari: bayt (b), megabayt (mb), gegabayt (gb).

O‘zbekcha matnlarni ovozlashtirish dasturi qisqargan shaklda matnda qo‘llanilgan hisob so‘zlarini aslidek berilgan son bilan birga o‘qib berishi uchun dastur ta’minotiga ularning audio formatda tayyorlangan ro‘yxati kiritilishi talab etiladi.

Tinish belgilari. Matnda eng asosiy qo‘llaniladigan belgilar – bu tinish belgilari. Boshqacha aytganda **prosodemografema**lardir. «Grafemalarning bu guruhi tovush tilining ritmik-intonatsion vositalarni (urg‘u, ohang, melodika, pauza kabilarni) yozuvda ifodalash uchun xizmat qiladi. Tinish belgilari gap va nutqning mazmun-mundarijasini shakllantirishda muhim rol o‘ynaydigan fonetik-fonologik vositalarni (ko‘tariluvchi ohang, pasayuvchi ohang, to‘lqinli ohang, sanash ohangi, pauza kabi supersegment birliklarni) yozuvda ifodalash uchun qo‘llanadi.

Bunday belgilar quyidagi vazifalarni bajarish uchun ishlatiladi:

a) Ijtimoiy aloqani (fikr almashuvi jarayonini) yozuvda to‘g‘ri ifodalash uchun;

b) Maqsad, mazmun yoki ma’noni, ularning o‘ziga xos «rang» va «tus»larini yozuvda aniq ifodalash uchun;

c) Gapning tarkibini hamda shu tarkib elementlari (komponentlari) o‘rtasidagi grammatik-semantik munosabatlarni ifodalash uchun;

d) Yozma nutqning ixcham va ravonligini ta’minlash uchun;
e) Yozma nutqdagi murakkab fikriy munosabatlarni ifodalash uchun»[2].

O‘zbek tilshunosligida tinish belgilarining quyidagi 10 ta turi grafikaning markaziy sistemasiga kiritiladi:

Nuqta (.)	Vergul (,)
So‘roq belgisi (?)	Ikki nuqta (:)
Undov belgisi (!)	Tire (–)
Nuqtali vergul (;)	Qavs (), []
Ko‘p nuqta (…)	Qo‘shtirnoq («»)

Bu tinish belgilari matnda so‘zlarning, gaplarning ohangini belgilaydi. O‘zbekcha matnlarni ovozlashtirish dasturi gaplarni, gap tarkibidagi so‘zlarni, birikmalarni tegishli ohangda o‘qishi uchun dastur lingvistik ta’minotini ishlab chiqishda bu masalalarga e’tibor qaratish zarur. Dastur tabiiy talaffuz intonatsiyasini hosil qilib bera olmasligi mumkin, lekin kerakli o‘rinlarda qo‘llanilgan tinish belgilarini nomi bilan o‘qib berishi masalaning yechimi bo‘la oladi.

Yuqoridagi fikrlarimizni xulosalab aytishimiz mumkinki, o‘zbekcha matnlarni ovozlashtirish dasturining har qanday o‘zbek tilidagi matnlarni hech qiyinchiliksiz o‘qib berishida uning lingvistik ta’minotining qay darajada muakammal ishlab chiqilgani katta ahamiyatga egadir. Xususan, hozirda biz amalda lotin hamda kirill yozuvidan foydalanamiz va dasturning har ikkala yozuvda bitilgan matnlarni o‘qib berishi bizga o‘zbek tilida mavjud boy merosdan unumli foydalanishimizni ta’minlaydi. Shuningdek, o‘zbek tiliga boshqa tillardan, asosan, rus tili va u orqali boshqa tillardan o‘zlashgan ruscha-internatsional so‘zlarni tadqiq etish va bunday so‘zlarni dastur lingvistik ta’minotiga kiritish masalalarini o‘rganish vazifasi ham oldimizda ko‘ndalang turibdi. O‘zlashma so‘zlarning talaffuzi o‘zbek tili so‘zlari talaffuzidan farq qilgani bois ham ularning audio formatdagi ham yozma shaklini lingvistik ta’minotga kiritish dasturning bunday so‘zlarni xatosiz o‘qishiga imkon yaratadi.

Matn turli sohalarga oid bo‘lishi mumkin, o‘zbekcha matnlarni ovozlashtirish dasturidan barcha soha vakillari o‘z sohalaridagi adabiyotlarni o‘qish uchun foydalanishlari tabiiy hol. Mana shu jihatlarni inobatga olgan holda daturning mukammal ishlashi, barcha sohalardagi matn turlarini o‘qib berishi uchun qisqartmalar, logografemalar, chiziqcha bilan yozilgan so‘zlar, tinish belgilari, lotin tili so‘zlari, matematik amallar, belgi, simvollar, shuningdek, sheva so‘zlarini ham o‘qish imkoniyatiga ega bo‘lishi talab etiladi. Qolaversa, qadimgi turkiy tilga oid so‘zlar, eski turkiy til manbalaridagi so‘zlar ham bundan mustasno emas.

[2] Назаров К. Ўзбек тили пунктуацияси. Т.: Ўқитувчи, 1976. – Б. 7-8.

МОДЕЛИРОВАНИЕ ЯЗЫКОВ (ЕСТЕСТВЕННЫХ И ИСКУССТВЕННЫХ) В КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

З. Ашуров, Самаркандский государственный институт иностранных языков, Самарканд, Узбекистан, zaki_uz@mail.ru

Компьютерное моделирование языка и речевой деятельности нуждается в солидной теоретической базе, и фундаментальная наука должна вплотную заняться соответствующими актуальными проблемами. Моделирование языков (естественных и искусственных) вписывается в проблему моделирования способностей человека.

Ключевые слова: *компьютерное моделирование языка, textanalyst, язык моделирования, виртуальная реальность, язык унифицированного моделирования, интеллектуальная автоматическая обработка, лингвистика, вычислительная лингвистика.*

MODELING (NATURAL AND ARTIFICAL) LANGUAGES IN COMPUTATIONAL LINGUISTICS

Z.Ashurov, Samarkand state institute of foreign languages, Samarkand, Uzbekistan, zaki_uz@mail.ru

Computer modeling of language and speech activity requires a solid theoretical base, and fundamental science must come to grips with relevant topical issues. Modeling languages (natural and artificial) fits into the problem of modeling human abilities.

Key words: *computer modeling of language, textanalyst, modeling language, virtual reality, language of unified modeling, intelligent automatic processing, linguistics, computational linguistics.*

Язык моделирования — это любой графический или текстовый компьютерный язык, который обеспечивает проектирование и построение структур и моделей, следуя систематическому набору правил и основ.
Язык моделирования является частью и аналогичным искусственному языку.

Язык моделирования в основном используется в области информатики и техники для проектирования моделей нового программного обеспечения, систем, устройств и оборудования. Контекст языка моделирования в основном текстовый и графический, но на основе требований и конкретного используемого домена языки моделирования подразделяются на следующие четыре категории:

- Язык моделирования системы;

- Языки моделирования объектов;
- Язык моделирования виртуальной реальности;
- Язык моделирования данных.

Язык унифицированного моделирования (UML) — популярный язык моделирования, который используется для графического построения системных и объектных моделей.

РОЛЬ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

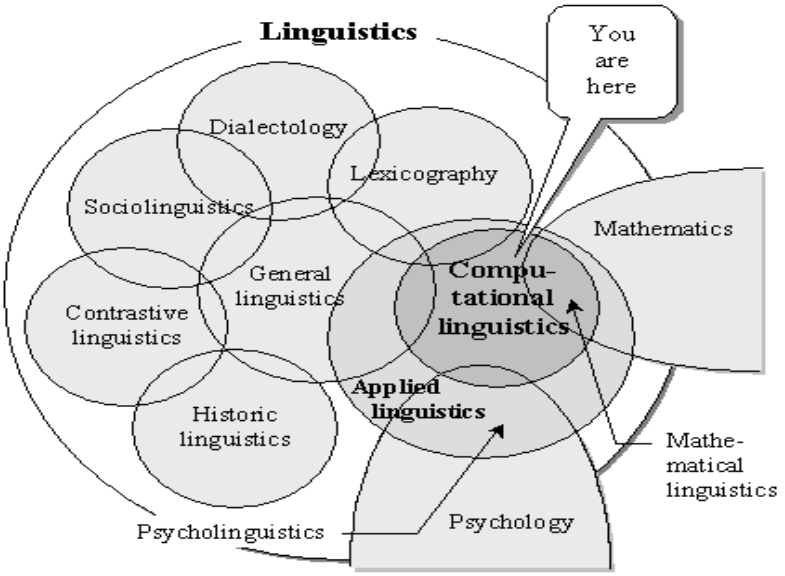
Мы живем в эпоху информации. Он льется нам со страниц газет и журналов, радио-громкоговорителей, телевизоров и экранов компьютеров. Основная часть этой информации имеет форму текстов на естественном языке. Даже в области компьютеров большая часть информации, которую они манипулируют в настоящее время, имеет форму текста. Похоже, что персональный компьютер в основном превратился в инструмент для создания, коррекции, хранения, управления и поиска текстовых документов.

Наши предки изобрели естественный язык много тысяч лет назад для нужд развивающегося человеческого общества. Современные естественные языки развиваются в соответствии со своими законами, в каждую эпоху они являются адекватным инструментом для общения людей, для выражения человеческих чувств, мыслей и действий. Структура и использование естественного языка основывается на предположении, что участники беседы имеют очень похожий опыт и знания, а также способ чувствовать, рассуждать и действовать. Большой проблемой проблемы интеллектуальной автоматической обработки текста является использование неограниченного естественного языка для обмена информацией с существом совершенно другого характера: компьютером.

В течение последних двух столетий человечество успешно справлялось с автоматизацией многих задач с использованием механических и электрических устройств, и эти устройства добросовестно служат людям в их повседневной жизни. Во второй половине двадцатого века человеческое внимание обратилось к автоматизации обработки естественного языка. Теперь люди нуждаются в помощи не только в механических, но и в интеллектуальных усилиях. Они хотели бы, чтобы машина читала неподготовленный текст, чтобы проверить его на правильность, выполнить инструкции, содержащиеся в тексте, или даже понять его достаточно хорошо, чтобы дать разумный ответ на основе его значения. Люди хотят сохранить для себя только окончательные решения.

Лингвистика — это наука о естественных языках. Точнее, он охватывает целый ряд различных смежных наук (см. Рис. I.1).

Рис. I.1



Вычислительная лингвистика может рассматриваться как синоним автоматической обработки естественного языка, поскольку основная задача вычислительной лингвистики — это просто разработка компьютерных программ для обработки слов и текстов на естественном языке.

Обработка естественного языка должна рассматриваться здесь в очень широком смысле, который будет рассмотрен ниже.

Фактически, этот курс немного «более лингвистичен, чем вычислительный», по следующим причинам:

- Мы в основном заинтересованы в формальном описании языка, относящегося к автоматической обработке языков, а не в чисто алгоритмических вопросах. Алгоритмы, соответствующие программы и технологии программирования могут различаться, а основные лингвистические принципы и методы их описания намного более стабильны.
- В дополнение к некоторым чисто вычислительным вопросам мы также косвенно затрагиваем вопросы, связанные с информатикой, косвенным образом. Ниже описывается более широкий набор понятий и моделей общей лингвистики и математической лингвистики.

Для компьютерной системы или ее части, которая должна считаться лингвистической, она должна использовать некоторые данные или процедуры, зависающие от языка, т.е. переход от одного естественного языка к другому.

Таким образом, не каждая программа, касающаяся текстов естественного языка, связана с лингвистикой. Хотя такие текстовые процессоры, как Windows Notebook, имеют дело с обработкой текстов на естественном языке, мы не рассматриваем их как лингвистическое

программное обеспечение, поскольку они недостаточно зависят от языка: их можно использовать одинаково для обработки испанского, английского или русского языков тексты, после некоторых алфавитных настроек. Таким образом, они зависят от языка. Однако они не полагаются на достаточно большие лингвистические ресурсы. Поэтому простые программы переноса переносятся только на программное обеспечение, которое можно считать лингвистическим. Что касается инструментов проверки орфографии, которые используют большой список слов и сложные морфологические таблицы, это всего лишь лингвистические программы.

Прикладная лингвистика развивает методы использования идей и понятий общей лингвистики в широкой практике человека. До середины двадцатого века приложения лингвистики были ограничены разработкой и совершенствованием грамматик и словарей в печатной форме, ориентированной на их более широкое использование неспециалистами, а также на рациональные методы преподавания естественных языков, их орфографии и стилистики, Это был единственный чисто практический продукт лингвистики. Другая система TextAnalyst для определения основных тем документа и отношений между словами в документе была разработана MicroSystems. Эта система не основана на словах, хотя в ней есть небольшой словарь стоп-слов (это предлоги, статьи и т. Д., И они не должны обрабатываться как содержательные слова).

TextAnalyst это работа с текстовой информацией. TextAnalyst формирует **семантическую сеть** — интегральное представление смысла текста, служащее основой для всех видов дальнейшего анализа.

Семантическая сеть — это множество понятий текста — слов и словосочетаний, связанных между собой по смыслу. В семантическую сеть включены не все слова текста, а лишь наиболее значимые, несущие основную смысловую нагрузку. При этом в сеть не входят обще употребимые слова, а также слова, очень редко встречавшиеся в тексте (этот параметр — частоту встречаемости, вы сможете настраивать по своему желанию). В TextAnalyst список важных слов используется для следующих задач:

- Сжатие текста путем исключения предложений или абзацев, содержащих минимальное количество важных слов, до тех пор, пока размер текста не достигнет порога, выбранного пользователем,
- Создание гипертекста путем построения взаимных ссылок между самыми важными словами и важными словами для других, к которым они якобы связаны.

Технология TextAnalyst основана на специальном типе динамического алгоритма нейронной сети. Поскольку программа Clasitex

основана на большом словаре, это программа, основанная на знаниях, а TextAnalyst — нет.

Важно иметь в виду, что язык — это не столько «форма выражения» готовых мыслей, сколько способ содержательной организации и представления знаний. Этот способ первичен, универсален, возникает с самим зарождением человеческого интеллекта и служит надежным инструментом его развития.

Компьютерное моделирование языка и речевой деятельности нуждается в солидной теоретической базе, и фундаментальная наука должна вплотную заняться соответствующими актуальными проблемами. Моделирование языков (естественных и искусственных) вписывается в проблему моделирования способностей человека. Языковая способность — это способность, делающая человека человеком, возникающая и развивающаяся под воздействием практических потребностей.

ЛИТЕРАТУРА:

1. И. Николаев, О. Митренина,Т. Ландо Прикладная и компьютерная лингвистика.. 2016 год. Издательство «Ленанд». 316 стр.
2. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. А.Анисимов 1991 год. Издательство: Киев: Наукова думка. 202 стр.
3. Автоматическое понимание текстов: системы, модели, ресурсы. Н. Леонтьева 2006 год. Издательство Академия. — 153 стр.
4. Информационные технологии и лингвистика XXI века.А. Гуслякова 2016 год. Правообладатель: МПГУ (Московский педагогический государственный университет). 130 стр.

Интернет ресурсы:

5. http://www.krugosvet.ru/enc/lingvistika/kompyuternaya-lingvistika
6. https://www.techopedia.com/definition/20810/modeling-language
7. http://genhis.philol.msu.ru/osnovy-kompyuternoj-lingvistiki/

ТУРКИЙ ТИЛЛАРНИНГ ФОРМАЛ МОДЕЛЛАРИДА СЎЗ ЯСАЛИШИ МАСАЛАСИ (ЎЗБЕК ТИЛИ МИСОЛИДА)

Турсунов Акмал[1], [1]НавДКИ, Тошкент, Ўзбекистон, akmal_10@mail.ru

Ушбу мақолада туркий тиллар, жумладан, ўзбек тилида сўз ясалиш масаласи ҳақида сўз боради. Ўзбек тилидаги формал моделлар, қолиплашган шакллар, янги бирликларнинг ҳосил бўлиши туркий тилларга хос хусусият ҳисобланиб келинганлиги асослаб берилган.

Таянч сўзлар: *морфемалар, дистрибутив, асос, сўз ясовчи, тур, препозитив, аффиксоид, боғловчилар, формант.*

ВОПРОСЫ СЛОВООБРАЗОВАНИЯ В ФОРМАЛЬНЫХ МОДЕЛЯХ ТЮРКСКИХ ЯЗЫКОВ (НА ПРИМЕРЕ УЗБЕКСКОГО ЯЗЫКА)

А. Турсунов, НавГМИ, akmal_10@mail.ru

В данной статье обсуждается вопрос словообразования в тюркских языках, в частности, узбекском языке. Формальные модели словообразования в узбекском языке являются отличительной чертой тюркских языков.

Ключевые слова: *морфемы, дистрибутив, основа, словообразование, вид, препозитив, аффиксоид, союзы, формант.*

ISSUES OF DERIVATION IN THE FORMAL MODELS OF TURKIC LANGUAGES (AS EXAMPLE OF UZBEK LANGUAGE)

A.Tursunov[1], NavSMI[1] akmal_10@mail.ru

This article discusses the issue of word formation in the Turkic languages, in particular, the Uzbek language. It is justified that the formal models, the formed signs and the formation of new forms in the Uzbek language is a special feature of the Turkic languages.

Key words: *Article, Turkish languages, uzbek language, official, formal, models.*

Туркий тиллар, жумладан, ўзбек тилида сўз ясалиш масаласи ҳамиша долзарб бўлиб келган. Ўзбек тилидаги формал моделлар, қолиплашган шакллар, янги бирликларнинг ҳосил бўлиши туркий тилларга хос хусусият ҳисобланиб, барча даврларда ўрганилишга ҳаракат қилинган.

Туркий тиллар морфемикаси (фонетик ва морфологик) реконструкциясини бериш, унинг ҳозирги ва қадимги даврлардаги тараққиётини кўрсатишни мақсад қилиб қўйган ва 1988 йилда нашр этилган «Туркий тилларнинг қиёсий–тарихий грамматикаси (морфология)»да барча морфемалар дастлаб икки турга ажратилади: тўлиқ маъноли (полнозначные) морфемалар ҳамда ёрдамчи (служебные) морфемалар.

Биринчи турга ўзак (асос)лар тааллуқли бўлса, иккинчи тур морфемалар формал морфемалардир. Формал ёки ёрдамчи морфемалар фақат аффикслардан иборат бўлмай, отлардаги йўналиш маъносини конкретлаштирувчилар, феъллардаги тус тарзни тасвирловчилар, боғловчилар, кўмакчи ва юкламалар, ундов ва мимемаларни ҳам қамраб олади. Ёрдамчи морфемалар тарқалиш (дистрибутив) жиҳатдан турлича белги хоссага эга бўлиб, шу асосда уларни уч турга ажратиш мумкин: 1) эркин морфемалар – мустақил қўллана оладиган морфемалар. Бу турга боғловчи, кўмакчи юкламалар киради; 2) боғлиқ морфемалар фақат айрим гуруҳ ва ўзак морфемалар билан қўллана оладиган морфемалар.

Буларга сон, эгалик, келишик, шахс, замон, майл, от ва феъл деривацияси (сўз ясалиши)да қатнашувчи бирлик (аффикс)лар тааллуқлидир; 3) нисбий боғлиқ морфемалар айрим тип (тур) тузилмаларда боғлиқ, айрим тип (тур) тузилмаларда эса тобе бўлмаган мавқеида қўлланиладиган морфемалар, булар от ва феълдаги кўмакчи ва ёрдамчи бирликлардир.

Феълларда ёт–, ўтир–, юр–, тур– бирликларида грамматикализациялашиш (мавҳумлашиш) жараёни юқори даражада бўлса, қолган феъллар мустақил феъллардан ёрдамчи (кўмакчи) феълларга ўтаётган, айланаётган оралиқ шакл кўринишидадир. Агар аффикс учун анъанавий тушунча сақланиб қолинадиган бўлса, от ва феъллардаги ёрдамчи бирликларни нисбий аффикслар ёки ярим аффикслар деб номлаш мумкин [1; 3-7].

Китобднинг «Кириш» қисмини ёзган Э.Р.Тенишевниннг қайд этишича, туркий тилларда префиксация мавжуд эмас деб ҳисобланади. (Н.К.Дмитриев, Ж.Дену, Э.Сепир каби олимлар), шунга қарамай туркий тилларда:

- келиб чиқиш манбаига кўра араб, форс ҳамда европа тилларидан кириб келган ва туркий асосларга қўшилиб, сўз ясалишида иштирок этадиган препозитив элементлар учрайди [1; 7];
- келиб чиқишига кўра туркий профикслар ҳамда профиксга айланиш босқичида турган сўзлар (ярим префикслар ёки префиксоидлар):

Олиб бор – обор, олиб чиқ – опчиқ каби тузилмалардаги о-, оп- каби элементлар; ярим префиксларга сифат ва равишдаги энг (энг катта) ҳамда ноаник артикль вазифасида қўлланилувчи бир сўзи тааллуқлидир. [2]

Демак, юқоридаги фикрлардан маълум бўладики, дастлаб ўзлашма сўзлар таркибида келиб, кейинчалик туркий асосларга қўшилиб, янги сўз ясаш учун ишлатиладиган препозитив бирликларни ажратиш имкони бўлса, айрим ҳолатларда ярим аффикслар ёки префиксоидларни ҳам белгилаш имкони мавжуд. Акад. А.П.Ҳожиев «Ўзбек тили сўз ясалиши» (Тошкент, 1989) китобида от ва сифат ясалишида қуйидаги аффиксоидлар борлигини кўрсатади: ҳам-: ҳамшаҳар, ҳамкурс (39-бет); -хона: қабулхона, дарсхона, элчихона; -нома: арзнома, мурожаатнома, таомнома (51-бет); хуш-: хуштабиат, хушбичим, хушҳаво; -аро (асли ора) русча между (меж) қатнашадиган сўзларни ўзлаштириш жараёнида қўллана бошлаши: хўжаликлараро, тармоқлараро; -бай: донабай, ишбай, кунбай (76-77 бет.)

Ушбу китобда асосдан олдин қўшиладиган ҳамда асосдан кейин қўшиладиган бирликларга нисбатан умумлаштирувчи атама-аффиксоид атамаси ишлатилган, ўз-ўзидан, префиксоид атамаси қўлланилмаган (бунга эҳтиёж бўлмаган). Проф. Ш.У. Раҳматуллаев рус-европа тилшунослигидаги айрим қарашларга эргашган ҳолда морфемика объектига фақат грамматик маъно ифодаловчи ва уларга эквивалент бирликларни киритади. Бир ўринда аффиксларни асосга қўшилиш ўрнига кўра префикс ҳамда суффикс деб номласа, бошқа ўринда «Ўзбек тили морфемалари тизимига префикс хос эмас», [3] деб ёзади.

Ш.У.Раҳматуллаев «морфемаларнинг табиатига кўра таснифи» бўлимида уларни уч турга ажратади: аффикс, аффиксоид, лексик табиатли морфема. Аффиксоидларга ошхона, чойхона каби мисоллар келтирилган, мавжуд адабиётлардаги фикрлардан фарқланувчи бирор фикр айтилмаган (маълум фикрлар такрорланган). Лексик табиатли морфемалар асосга зич қўшилмаслиги, алоҳида келиши билан лексемаларга ўхшайди. Мазмун жиҳатини ҳисобга олиб, уларни лексик табиатли морфемалар деб номлаш мумкин. Бундай морфемалар билан, учун каби кўмакчилар, сифат ва равишдаги энг, жуда каби бирликлар, феъл туркуми доирасида –бер, –ол–, бўл- каби кўмакчилардир.

Кўринадики, аффиксоидлар грамматик маъно ифодалаш талаби билан (сўз ясаш эҳтиёжи туфайли) пайдо бўлади. Улар ўзи келиб чиққан мустақил сўз (корреляти) билан алоқасини тамомила узмаган бирликлардир. Шу асосда уларни лексик табиатли морфемалар дейиш объектив бўлади, лексик мустақил сўз сифатидаги манбаси билан алоқаси узилган, унитилган билан, учун, энг, жуда каби бирликларни «лексик табиатли» деб номлаш қай даражада ўринли бўлиши мумкин? Ҳолбуки, юқорида келтирилган бирликлар лексик маъносини йўқотиб, грамматик бирликларга айланган,

шу сабабдан, уларни «лексик табиатли» деб бўлмайди. Фақат эди, экан, эмиш каби тўлиқсиз феъллардан ташқари, барча кўмакчилар аффиксоидларга ўхшаш вазифада кела олади.

Шу асосда юқоридаги тасниф тил бирликлари-морфемалар табиатини объектив акс эттирмайди, ноаниқ ва чалкаш фикрлар уйғонишига сабаб бўлади.

Акад. А.П. Ҳожиевнинг «Ўзбек тили сўз ясалиши тизими» китобида «сўз ясовчи» термини мазмунидаги умумий (асосий) хусусият ҳақида фикр билдириб, сўз ясалиш бирлиги ва материал бирлик тушунчаларини фарқлайди: «сўз ясовчи» термини ҳам ясама сўзнинг иккинчи таркибий қисмини умумий (асосий) хусусиятига кўра атайди ва бу номда унинг материал жиҳатдан қандай бирлик экани акс этмайди. Масалан, рангдор, қабулхона, камчиқим сўзлардаги -дор, -хона, кам- сўз ясалиш бирлиги сифатида – сўз ясовчи. Ҳар учала материал-аффикс, аффиксоид, ёрдамчи сўз учун «формант» («сўз ясовчи формант») терминини умумлаштирувчи термин сифатида қўллаш тавсия этилади [4]. Акад. А.П.Ҳожиев сўз ясалиши тизимида аффикс, аффиксоид ва ёрдамчи сўзнинг вазифа жиҳатидан ўхшашлиги, муштараклигини алоҳида таъкидлайди: аффиксоид ва ёрдамчи сўзлар «сўз ясовчи сифатида, сўз ясовчи аффикслардан фарқли бирон белгига эга эмас. Аксинча, сўз ясовчи аффиксларга хос умумий хусусиятларга бу сўз ясовчилар ҳам эга. Ҳатто кам- ва ҳам- сўз ясовчилари сўз ясовчи сер- ва -дош аффиксларига антоним, маънодош: камҳосил – серҳосил, ҳамсуҳбат – суҳбатдош. Шунинг учун буларни сўз ясовчи аффикслардан ажратиб (алоҳида олиб) таҳлил этишнинг ҳожати йўқ» [5] Проф. Э.Р.Тенишев эса бу хил бирликларни «нисбий аффикслар», «ярим аффикслар» каби тавсифлаб, от ва феъл сўз туркумидаги ёрдамчи (служебные) сўзларнинг бир тури эканлигини қайд этади [6].

Ўзбек тилининг ўзига оид сўз ясовчи ёрдамчи сўз сифатида қил (айла, эт), бўл элементларини ёрдамчи сўз номи билан бир гуруҳга бирлаштиради.

Проф. Т.Мирзақулов қўшимчасимон бирликлар сўз ясалишидагина эмас, балки шакл ясалишида ҳам учрашини кўрсатади.»Ташқи томондан сўзга ўхшаб кўринадиган, шу билан бирга, мустақил сўз ҳолатини, алоҳида сўзлик белгисини ҳам сақлаб қўлланилаётган бирликлар (билан, учун, сари; энг, эди, экан, эмиш каби бирликларнинг мустақил сўзга тенг келган, шу белгига эга бўлган таянч манбаи дастлабки маъноси сакланиб қолмаган, шунинг учун улар аффиксоидлар эмас) феъл сўз туркумида ҳам кенг тарқалган ҳодисалардир»[7]. Олимнинг қайд этишича, ҳозирги ўзбек адабий тилида 30 га яқин феъл маълум модель «етакчи феъл+ кўмакчи феъл» моделида (кулиб қўйди, айтиб берди, гапириб юрибди) қўлланади.

Ҳаракатнинг турли тавсифини (ҳаракатнинг тўсатдан бўлиши, узоқ давом этиши каби) ифодаловчи кўмакчи феъллар аффикс мақомига эга (аслида ушбу моделдан –(и)б куй-, -(и)б бер-, -(и)б юр- кўринишидаги

аналитик формантлар белгиланиши керак). Кўмакчи феъллар маъносининг грамматик мавҳумлашиши (абстрактлашиши) уларнинг аффикслар томон силжишини кучайтиради: -(и)б юбор-: кетиб юборган – кетворган. Бу товуш ўзгаришлари аффиксоидларнинг аффиксларга айланишини тезлаштиради: бора турғон – борадигон – борадиган: бор-а-диган; бориб ётибди – бораётибди – бораяпти.

Шундай савол туғилади: аффиксоид ва у боғланган таянч манба мустақил сўз ўзаро қандай алоқага эга, уларни омоним морфемалар деб тавсифлаш мумкинми? Масалан, ошхона, чойхона, кулгихона, гримхона, элчихона ясалмаларидаги -хона аффиксоиди ва мустақил сўз (бу уйда тўрт хона бор) хона ўзаро омоним бўла оладими?

-Хона аффиксоид маъносидаги мавҳумлашиши ўрин маъносини билдирадиган вазифа билан аффиксал тизим ихтиёрига тушиб кетиши, шу тизим таркибидаги алоқаларнинг муҳим бирлигига айланганлиги энди уларни ўзаро омоним морфемалар деб ҳисоблашга имкон беради. Аффиксоидлар таянч манбадаги айрим маъно компонентларини сўз ва шакл ясалишидаги янги маъноларнинг шаклланишига олиб келадиган функционал бирликлардир. Луғавий маъноси ва бирикувчи қисмларнинг хусусиятига қараб такрор сўзлар жамлик, умумлаштириш, экспрессив-услубий маъноларни билдиради. Бу хил такрорлар (ота-она, ака-ука, от-пот, овқат-повқат; ади-бади; бола-чақа) оғзаки сўзлашув, бадиий адабиёт, оммавий ахборот воситалари, илмий-сиёсий манбалар, таржима адабиётларида, нутқнинг деярли барча услубларида кенг қўлланади.

Проф. Ш.У.Раҳматуллаев шу ўринда «формула», «қолип» тушунчаларидан фойдаланади: «Мустақил сўздан, шуни фонетик сояси, формуласи асосида жуфтлаш жуда актив йўл бўлиб, буларни деярли ҳар бир от асосида тузиш мумкин; шунга кўра бу тур жуфтликни тил бирлигидан кўра нутқ бирлиги деб қилиш тўғрироқ: сетка-петка, дафтар-сафтар каби. [8]

Тилда бундай фонетик сояларни ҳосил этиш қолиплари бор, ана шу қолиплар билан истаган от асосида жуфтлик тузаверамиз. [9]

Бу хил жуфтликлар маълум формула, модель асосида тузилар экан, юзага келган конкрет тузилма (такрор)лар нутқ бирлиги, тил эгалари онгида мавжуд бўлган моделларнинг маҳсули, ҳосиласи деб баҳоланиши зарур. Демак, такрорларнинг модели тил бирлиги бўлса, шу моделлар асосида пайдо бўлувчи конкрет бирликлар нутқ бирлигидир.

АДАБИЁТЛАР:

1. Сравнительно- историческая грамматика тюркских языков. Морфология. М., 1988, стр. 3-7
2. Кубрякова Е.С Части реги в ономасиологическом осещении. М., 1978. –С. 84.

3. Ш.Раҳматуллаев Ҳозирги ўзбек адабий тили. Тошкент, 2000 –Б.151-153
4. Ш.Раҳматуллаев Ҳозирги адабий ўзбек тили. Тошкент, 2006 –Б.114-122
5. Ҳожиев А., Ўзбек тилида қўшма, жуфт ва такрорий сўзлар. -Тошкент, 1963. –Б. 23
6. Кубрякова Е.С. Типы языковых значений. Семантика производного слова. -М., 1981, –Б. 88
7. Мирзакулов Т. Грамматика ўқитишнинг лингвистик асослари. -Тошкент,1994, –Б.33
8. Ш.Раҳматуллаев Ҳозирги ўзбек адабий тили Тошкент, 2000 йил. –Б. 151-153
9. Ўзбек тили грамматикаси, –Б.268.

KOMPYUTER LINGVISTIKASI MASALALARINI «ELASTIC SEARCH» TEXNOLOGIYASI YORDAMIDA YECHISH

M. Sh. Norpulotova, A. Q. Nematov,
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti,
Toshkent, O‘zbekiston, norpulatova@navoiy-uni.uz

O‘zbek tilini kompyuterga olib kirish masalalaridan biri tarjima lug‘atlar yaratish masalasidir. Bu masalani Elastic Search texnologiyasi yordamida bajarish ancha samaraliroqdir. SQL texnologiyasidan foydalanishda qo‘yiladigan ish maydonlari har xil kasbdagilar uchun har xil bo‘lishi mumkin, bu esa aniq strukturaga ega SQL texnologiyasida noqulayliklarni keltirib chiqaradi. NoSQL texnologiyasi esa bundan mustasno, ya’ni unda ishlash davomida berilganlar bazasida qo‘shimcha xususiyatga ega bo‘lgan maydonlar kiritish mumkin. NoSQL texnologiyalaridan biri Elastic Search bo‘lib, unda axborotni saqlovchi asosiy element hujjat hisoblanadi.

Tayanch so‘zlar: *NoSQL, SQL, Elastic Search, JSON, tarjima lug‘at, replicas, type, node, cluster, shards*

SOLVING ISSUES OF COMPUTATIONAL LINGUISTICS BY ELASTIC SEARCH TECHNOLOGY

Norpulotova M.Sh., Nematov A.Q.,Tashkent state university of Uzbek language and literature, Tashkent, Uzbekistan

The main problem of introducing the Uzbek language to the computer is that creating translation dictionaries. Doing that by Elastic Search is more profitable. Working screen which is given for using SQL technology is probably diverse for varies professions. SQL is strictly structured technology, that’s why, this difference may be cause a few defects. But NoSQL is different from SQL technology the meaning that, while working with NoSQL you may add additional fields to given another base. One of the technologies of NoSQL is Elastic Search and it is basic element which keeping the information’s documents.

Key words*: NoSQL, SQL, Elastic Search, JSON, translation dictionary, replicas, type, node, cluster, shards.*

РЕШЕНИЕ ЗАДАЧИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ ЭЛАСТИЧЕСКИМИ ПОИСКОВЫМИ ТЕХНОЛОГИЯМИ

М. Норпулатова, А. Нематов, Ташкентский государственный университет узбекского языка и литературы им. Алишера Навои, Ташкент, Узбекистан, norpulatova@navoiy-uni.uz

Одной из основных проблем введения узбекского языка в компьютерные технологии является проблема создания переводных словарей. Эффективнее решать этот вопрос с помощью программы Elastic Search. При использовании SQL технологий объем рабочей области программы различается для различных профессий. SQL сильно структурированная программа, что создает некоторые неудобства в использовании этой технологии. В отличие от SQL технология NoSQL позволяет добавлять в базу данных дополнительные поля. Elastic search является одной из технологий NoSQL, базовые элементы которого содержатся в информационных документах.

Ключевые слова: *NoSQL, SQL, Elastic Search, JSON, переводной словарь, реплики, тип, узел, кластер, осколки.*

O‘zbek tilini kompyuterga olib kirish, ya’ni til bilan bog‘liq masalalar — o‘zbek tiliga o‘qitish, bilimlarni baholash (test), matnlarni o‘zbekcha ovozlashtirish, axborotni ovoz orqali kompyuterga kiritish, matnlarni tahrirlash, tarjima lug‘atlar yaratish, tarjima qilish, kompyuterda bajarish kabi imkoniyatlarni yaratadi.

Shu masalalardan tarjima lug‘atlar yaratish masalasini Elastic Search texnologiyasi yordamida bajarish ancha samaraliroq. «Elastic Search» texnologiyasi bu nima va u qanday amalga oshiriladi? Quyida savollarga javob beramiz.

SQL texnologiyasidan foydalanishda qo‘yiladigan ish maydonlari har xil so‘zlar uchun har xil bo‘lishi mumkin, bu esa aniq strukturaga ega SQL texnologiyasida noqulayliklarni keltirib chiqaradi. NoSQL texnologiyasi esa bundan mustasno, ya’ni unda ishlash davomida berilganlar bazasida qo‘shimcha xususiyatga ega bo‘lgan maydonlar kiritish mumkin.

NoSQL texnologiyalaridan biri Java tilida yozilgan Lucent qidiruv bibliotekasi asosida yaratilgan Elastic Search bo‘lib, unda axborotni saqlovchi asosiy element hujjat hisoblanadi. Hujjatlar JSON (JavaScript Object Notation) ko‘rinishida saqlanadi.

Elastic Searchning asosiy maqsadi hujjatlar bazasida murrakab va samarali qidiruv tizimini qo‘llashdir. Uning asosiy xarakteristikalari quyidagilardan iborat:

- gorizontal kengayuvchanlikka asoslangan, berilganlar bir emas, bir qancha tugunlarda saqlanadi;
- berilganlar bazasida turli ko‘rinishdagi ma’lumotlarni saqlash mumkin;
- berilganlar bazasining strukturasini ishlash davomida o‘zgartirish mumkin; ko‘p protsessorlik va taqsimlangan tizimlarda unumli foydalanish imkoniyati ko‘zda tutilgan;
- tizim yaratish davrini qisqartirish va har bir foydalanuvchi uchun kerakli interfeys yaratishga mo‘ljallangan;
- qoniqarli ishlash tezligi va unumdorligi ta’minlanadi.

Elasticsearchda asosiy qo‘llaniluvchi elementlar quyidagilardir:

document (hujjat) — axborotni saqlovchi asosiy element. Hujjatlar JSON (JavaScript Object Notation) ko‘rinishida saqlanadi.

index (indeks) — hujjatlarni ma’lum xarakteristikalar asosida guruhlash uchun ishlatiladi.

type (tip) — indeksning ichida bir yoki bir necha tipni aniqlash mumkin. U indeksning mantiqiy qismi yoki kategoriyasi bo‘lib, foydalanuvchinig ixtiyoriga qarab tanlanishi mumkin.

node (tugun) — alohida ajratilgan server bo‘lib, berilganlar bazasini saqlash, indekslash va izlash uchun ishlatiladi.

cluster (klaster) — bir yoki bir necha tugundan (node) iborat to‘plam bo‘lib, berilganlarni birgalikda saqlash, indekslash va izlash uchun ishlatiladi. Klaster foydalanuvchilar uchun ma’lumotlar bilan ishlashda shaffoflikni ta’minlaydi.

shards (shard) — indekslarni bo‘lib saqlash imkonini beradi. Ba’zan indekslar juda katta bo‘lib ketishi natijasida bir tugundagi magnit saqlovchiga sig‘masligi mumkin. Shuning uchun indeks yaratayotga vaqtda shardslarning soni berilsa, Elastic search indekslarni shuncha tizim tugunlariga bo‘lib taqsimlash imkoniga ega bo‘ladi.

replicas — tarmoqda va taqsimlangan tizimlarda ishlaganda foydalanuv-chanlikni ta’minlash uchun ishlatiladi. Tizimning biror tuguni ishdan chiqqanda ro‘y berishi mumkin bo‘lgan holatning oldini olish uchun, shardlarning nusxasini saqlash va ular bilan ishlash imkonini beradi.

Shuni aytib o‘tish zarurki, indekslarni bo‘laklarga ajratish (Shards) va ularning nusxalarini saqlash (Replicas), shuningdek ular bilan ishlash to‘liq Elastic Search zimmasida bo‘lib, u foydalanuvchi uchun shaffoflikni ta’minlaydi. Shunday qilib, har bir indeks bir qancha shardlarga bo‘linishi mumkin. Shardlar o‘z navbatida 0 yoki bir nechta replikalarga (nusxa) ega bo‘lishi mumkin. Shardlarning va replikalarning soni indeksni yaratish vaqtida berilishi mumkin. Ishlash davrida foydalanuvchi replikalarning sonini dinamik o‘zgartirishi mumkin, lekin shardlarning sonini o‘zgartirib bo‘lmaydi. Agar alohida ko‘rsatilmasa, har bir indeks uchun 5 ta birlamchi shard va har bir shard uchun 1

tadan replika ajratiladi. Elastic Searchda taqsimlangan tranzaksiya mexanizmi tom ma’noda qo‘llanilmaydi, shu sababli berilganlar bilan ishlash va izlash deyarli real vaqt ichida amalga oshiriladi. Odatda, berilganlarni indekslash, yangilash, o‘chirish va ularni izlash natijalarida akslanishi orasidagi vaqt bir sekund ichida amalga oshiriladi. Odatdagi SQL tizimlarida esa barcha o‘zgarishlar faqat tranzaksiya tugagandan keyinggina ko‘rinadi.

Serverga barcha hujjatlarni olish so‘rovi JSON ko‘rinishiga quyidagcha amalga oshiriladi:

```
$ curl -XGET 'http:://localhost:9200/free/JOB//_search?pretty=true' -d '
{
        «query» : {
                «match_all» : {}     //Barcha hujjatlarni chop qil;
        }
}
```

«pretty=true» parametri javobni aniq chop qilishga yordam beradi.

«c++» termini qatnashgan hujjatlarni olish so‘rovi esa quyidagicha amalga oshiriladi:

```
$ curl -XGET 'http:://localhost:9200/free/JOB//_search?pretty=true' -d '
{
        «query» : {
                «query_string» : {
                        «query» : «c++»    // «c++» qatnashgan hujjatlarni chop qil;
                }
        }
}
```

ADABIYOTLAR:

1. http://elaticsearch.docwiki.ru
2. www.elastic.co
3. A.Q.Po‘latov, «Kompyuter lingvistikasi», Toshkent, «Akademiknashr» 2011.

ИЛМИЙ МАТННИ ҚАЙТА ИШЛАШДА ПРАГМАТИК ЁНДОШУВ

А. Шерматов, Самарқанд давлат чет тиллар институти, Самарқанд, Ўзбекистон, akramshermatov76@gmail.com

Мазкур мақолада илмий матнни қайта ишлаш, таҳрир жараёнида унинг таркибий ўзгариши, бундан ташқари, илмий матнни қайта ишлашда прагматик ёндашув ҳақида сўз боради.

Таянч сўзлар: *матн, илмий матн, матнни қайта ишлаш, контент, прагматик ёндашув, матнни таҳлил ва таҳрир этиш, матн белгилари, автоматлаштирилган таҳрир жараёни.*

ПРАГМАТИЧЕСКИЙ ПОДХОД К ОБРАБОТКЕ НАУЧНЫХ ТЕКСТОВ

А. Шерматов (PhD), Самаркандский государственный институт иностранных языков (СамГИИЯ), Самарканд, Узбекистан, akramshermatov76@gmail.com

В этой статье описывается структура научного текста, его композиционная структура, изменение содержания в процессе переработки, изменение текста или сущность автоматизации создания текста. Также рассмотрены прагматично-коммуникативные аспекты научного текста.

Ключевые слова: *текст, научный текст, прагматический, обработка текста, контент, информация, шифтер, код, ссылка, дейксис, дейктические знаки, анализ текста, текстовые знаки, автоматическое редактирование.*

PRAGMATIC APPROACHING PROCCESING OF SCIENTIFIC TEXTS

A. Shermatov, Samarkand state institute of foreign languages, Samarkand, Uzbekistan, akramshermatov76@gmail.com

This article describes the structure of the scientific text, its compositional structure, content change in the process of modification, changing the text, or the essence of an automated mechanism in creating text. In addition, the pragmatic-communicative aspect of scientific text has been studied there.

Key words: *text, scientific text, pragmatic, text processing, content, information, shifter, code, footnote, deixis, deictic tag, text analysis, text tags, automatic editing.*

И.Р.Гальпериннинг таърифлашича, матн – бу аниқ бир мақсадга йўналтирилган ва прагматик кўрсатмага эга бўлган, лексик, грамматик, мантиқий, стилистик алоқалар воситасида таркиб топган, адабий тил андозасига солинган, ёзма ҳужжат кўринишида мужассамлаштирилган ва якуний хусусиятга эга бўлган нутқий-ижодий асардир [2; 18]. Демак, матннинг таркиб топишида унинг бўлаклари ўртасидаги ўзаро маъно муносабатлари муҳим роль ўйнайди. Зеро, «мазмуний валентлик мазмуний семаларнинг боғланиш имконияти»ни яратади. Бундан ташқари, матнни тушунишда «мазмунан тугалланган, юқори даражадаги коммуникатив бирлик» [5; 219] сифатида таърифланиши ҳам эътирофга лойиқдир. Бироқ, шу йўсинда компьютер тизимида матнни қайта ишлаш деганда электрон матнни ўзгартириш ёки яратишнинг автоматлашган механизми тушунилади. Компьютер буйруқлари одатда матнни қайта ишлашда жалб этилади. Бунда янги контентни яратишга кўмак бериш ёки контентга ўзгаришларни олиб кириш, контентни жойини алмаштириш ёки тадқиқ қилиш, контентни форматлаш ёки контентнинг нозик қисмларини умумлаштириш англашилади.

Илмий-техник матн ўз мақсади, структураси, мазмуни ҳамда синтактик тузилиши ҳамда ўзига хос стилистик хусусиятлари билан бадиий матндан фарқ қилади [6, 8, 10]. Масалан, илмий матннинг мақсади интеллектуал маълумотларни бериш, китобхоннинг диққат-эътиборини муҳокама қилинаётган илмий муаммога тортиш ва уни кейинги маълумотларнинг моҳиятига қараб йўналтиришдан иборатдир. Илмий-техник матнларда маълумотни узатишда аниқлик ва расмийликка риоя қилиш асосий талабдир. Бу эса, ўз навбатида, мантиқийликни, лўндаликни, аниқликни, таъсирчанликни, объективликни, андозавийликни талаб қилади. Аммо илмий баён руҳий ҳиссиётлардан тўлиғича холи бўлишини тасаввур қилиш ҳам қийин, чунки илмий матн муаллифи ўзгалар фикрини изоҳлашда бу фикрга нисбатан шахсий муносабатини билдириб, турли эмоционал ҳолатларни намоён қилиши мумкин [10; 16].

Илмий матнлар бадиий адабиётдан қатор мазмуний ва синтактик хусусиятли стилистик белгилари билан фарқ қилади. Илмий асар кенг китобхонлар оммаси учун эмас, балки маълум доирадаги билимлар мажмуига, аниқ соҳа бўйича тажрибага эга бўлган, айнан шу соҳа мутахассислари ҳисобланган шахслар учун мўлжалланади. Китобхондан нафақат матнни тушуниш қобилияти, балки унда тасвирланган бутун борлиқни, воқеликни билиш ҳам талаб қилинади. Умуман олганда, илмий мақола матни дейксис асосида қурилган: маълум соҳа бўйича билимга эга бўлмаган шахсга бу матн ноаниқ бўлади, чунки унда доимо фаннинг ўша соҳаси бўйича далиллар, асосларга мурожаат қилиш ҳолатлари мавжуддир [4; 53]. Илмий ишда берилаётган маълумотларни идрок қилишда кўпроқ тасаввур қилиш қобилиятини фаоллаштиришга тўғри келади, чунки маълумот берилаётган

объектни тўғридан-тўғри кўриш имконияти бўлмаган ҳолда инсоннинг тасаввур қилиш имконияти амалга ошади. Илмий матн учун хусусиятли ҳолат тасаввур қилинаётган объектга «ишора қилиш» ва анафорик дейктик белгилар воситасида воқелик узлуксизлигини ёритиб беришдир. Масалан, матнда далиллар, ҳаволаларнинг борлигининг ўзи китобхонга ўша соҳада илгари бажарилган тадқиқотларни кўрсатиш дейктик вазифасини ўтайди. Айни пайтда, далил, ҳавола саҳифаси рақами катафорик кўрсатма вазифасини бажариб, матн охирида худди шу каби тадқиқот иши ким томонидан ва қачон бажарилганлиги ҳақида хабар беради: In accordance with theory [12-13] developed for complexes with intermolecular H-bonds and applied in Refs [2,4,5,] for intramolecularly hydrogen – bonded systems. It has been observed in many cases [1-6] that the NMR spectra of some molecular organic crystals of two components in a certain temperature range.

Илмий матнлар моҳиятига кўра қуйидаги композицион-нутқий шаклларга эга бўлади [1; 137]: ахборотни баён қилиш, тавсиф бериш, мулоҳаза юритиш. Бу шаклларнинг ички-мантиқий моҳияти баён қилиш услубини белгилайди. Баён қилиш тури бўйича мақолалар илмий-назарий, илмий-экспериментал ва шарҳловчи турларга бўлинади. Илмий-назарий мақолаларнинг мантиқий моҳияти асосида мулоҳаза юритиш шаклий тузилмаси ётади.

Матнга тегишли тадқиқ қилинган муаммоларнинг кўпчилиги Р.О.Якобсоннинг шифтерлар ва феъл категорияларига бағишланган мақоласида [11; 95-113] ҳам ўз аксини топган. Муаллифнинг диққат-эътиборида грамматик категориялар тизимида нутқ фаолиятининг роли ва унинг асосий таркибий қисмлари ётади. Р.О.Якобсон «шифтер» тушунчасини батафсил ёритиб, билдирилаётган ахборотни нутқ фаолиятига бевосита алоқадорлигига ишора қилувчи ҳар қандай элементни (шунингдек, феъл ёки олмошни) назарда тутган. Шифтерни тасвирлаш учун Р.О.Якобсон информатика фанининг атамаларидан фойдаланди. Масалан, «код» атамаси воситасида Р.О. Якобсон ахборотни узатишни маъқул келган ҳолда шакллантирувчи услубни атайди.

Р.О. Якобсон дейктик элементларни «шифтерлар» деб атади ва бу туркумни маълум ахборотни кодлаштириш учун хизмат қилувчи воситалар сифатида талқин қилди. Шахс олмошлари шифтерларнинг типик туридир. Р.О. Якобсон шифтерларга замон, шахс ва майл категорияларини ҳам киритди.

Дейксис муаммосини тадқиқ қилишда Р.О.Якобсон билдирган фикрларнинг моҳияти шундаки, у тил тизимидаги эгоцентрик сўзларнинг ўзига хос хусусияти ва қўлланилиши борасидаги тадқиқотларнинг муҳимлигини яна бир бора қайд этди. Олимнинг шифтерлар ҳақидаги тадқиқоти, матн таҳлилининг грамматик асосини яратиш учун хизмат қилади. Матнни қайта ишлаш энг юқори компьютерлаштириш

даражасидаги матн белгиларига йўналтирилган. Бошқача қилиб айтганда, матнни қайта ишлаш ахборотни автоматик узатиш билан боғлиқ ҳолатда кечади [12; https://en.wikipedia.org/wiki/Text_processing].

Матнни қайта ишлаш жараёнини сўзни қайта ишлаш билан аралаштирмаслик керак. Матнни қайта ишлаш жараёнида тасодифий содир бўладиган кириш ўрнига ёндошувга асосланган ва тўғридан-тўғри тақдимот сатҳида ва билвосита дастур қатламида ишлайди. Сўзни қайта ишлашдан фарқли ўлароқ, матнни қайта ишлаш тўлиғича пишиқ бўлмаган материаллар сатҳида ишлайди ва мустақил ҳолатда фаолияти юқори. Матнни қайта ишлаш жараёни техник буйруқлар ёки матн муҳаррири ёрдамида амалга оширилади. Компьютерлаштириш тизимида матнни қайта ишлаш жараёни асосан янги мақолалар, китоблар ва журналларни яратишда қўл келади. Матнни қайта ишлаш, шунингдек, манба ҳужжатларини муайян процессор форматида сақламайди ва translators ва parsers каби янги восита ва функцияларга мурожаат қилишда ёрдам беради [13; https://www.techopedia.com/definition/22541/text-processing].

Хулоса қилиб айтганда, илмий матннинг таҳрир жараёнини автоматлаштириш учун биринчи навбатда унинг лингвистик таъминотини амалга ошириш керак. Илмий матнларни автоматик таҳрир қилувчи дастурнинг лингвистик таъминотини яратиш учун матндаги таркибий қисмларни семантик майдон сифатида ажратиш, уларнинг маъно тузилишини ва грамматик хусусиятларини моделлаштириш зарур бўлади. Илмий матн таркиби ўрганилганда, баён йўналиши сўзловчи (муаллиф)дан бошқа шахсга мурожаатида эмас, балки баён предмети, яъни тадқиқ қилинган объект, унинг натижаларига кўчишини кўрамиз. Муаллиф томонидан узатилаётган ахборотнинг диққат марказида унинг ўзи эмас, балки бажарилган иш-тадқиқоти ва ушбу тадқиқот объекти ҳақида фикр билдириш мақсади туради.

АДАБИЁТЛАР:

1. Акентьева В.Н. Типологические характеристики научных статей и некоторые трудности их понимания // Текст как важнейшая единица коммуникации (в диахронии и синхронии). – Киев, 1984. – С.137-142.
2. Гальперин И.Р. Текст как объект лингвистического исследования. –М.: Наука, 1981. – 139 с.
3. Йўлдошев Б. Компьютер лингвистикаси. Услубий қўлланма. – Самарқанд: СамДУ нашриёти, 2011. – 112.
4. Кожевникова К. Об аспектах связности текста // Синтаксис текста. –М.: Наука, 1979. – С.49-67
5. Марчук Ю.Н. Компьютерная лингвистика. – М.: АТС: Восток и Запад, 2007. – 317 с.
6. Мукаррамов М. Ҳозирги ўзбек адабий тилининг илмий стили. – Т.: Фан, 1984. – 160 б.

7. Муҳамедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъмин яратиш (методик қўлланма). – Тошкент: ТошДПУ нашри, 2006. –Б. 14-29.
8. Разинкина Н.М. Развитие языка английской научной литературы: лингвостилистическое исследование. – М.: Наука, 1978. – 211 с
9. Холмонова З.Т. Компьютер лингвистикаси асослари модули бўйича ўқув-услубий мажмуа. – Тошкент: ТошДЎТАУ, 2016. – 62 б.
10. Ҳакимов М.Х. Ўзбек илмий матнининг синтагматик ва прагматик хусусиятлари: Филол. фан. ном. дис…. автореф. – Т.: Ўз ФА ТИ, 1993. – 27 б.
11. Якобсон Р.О. Шифтеры, глагольные категории и русский глагол // Принципы типологического анализа языков различного строя. – М.: Наука, 1972. – С.95-113.
Интернетдан олинган манбалар:
12.https://en.wikipedia.org/wiki/Text_processing
13.https://www.techopedia.com/definition/22541/text-processing

НЕРАСПРОСТРАНЁННЫЕ ПРЕДЛОЖЕНИЯ В УЗБЕКСКОМ И АНГЛИЙСКОМ ЯЗЫКАХ

Н.А. Садуллаева, Национальный университет Узбекистана, Ташкент, Узбекистан, nilufar_sadullaeva@mail.ru

В этой статье рассматриваются нераспространённые предложения в узбекском и английском языках. В узбекском языке распространённое и нераспространенное предложения изучены и описаны подробно в учебниках, созданных до получения независимости. Рассмотренные материалы доказывают необходимость тщательного сравнительного анализа этих двух языковых единиц.

Ключевые слова: *распространённое предложение, нераспространенное предложение, грамматическое строение, синтаксис, узбекское языкознание, английское языкознание, семантическая категория, формальное языкознание.*

SIMPLE SENTENCES IN UZBEK AND ENGLISH LANGUAGES

N. Sadullayeva, National university of Uzbekistan, Tashkent, Uzbekistan, nilufar_sadullaeva@mail.ru

This article is devoted to the analysis of unextended sentences in English and Uzbek languages. Extended and unextended sentences have been analyzed more in Uzbek language than in other languages. Analyzed materials prove the need to study the similarity and difference of these two units.

Key words: *extended sentence, unextended sentence, grammatical structure, syntax, linguistic categories, Uzbek linguistics, English linguistics, semantic category, formal linguistics.*

Синтаксис является одним из сложных и интереснейших областей языкознания, так как мысль человека выражает предложение – одна из основных единиц синтаксиса. Учёные доказали, что ещё не родившийся ребёнок уже в утробе матери, закладывается способность соединять слова, а по истечению времени он соединяет их и постепенно учится говорить и образует простые и краткие предложения. Это чудо природы в будущем помогает сформировать способность мыслить и её выражать, а также её развитию.[1] Общеизвестно, что самая краткая мысль, как правило, выражается ясно, чётко и лаконично в нераспространённом предложении. В настоящей статье мы рассмотрим некоторые особенности данного вида простого предложения.

Нераспространённые предложения в узбекском языке были исследованы намного лучше, по сравнению с некоторыми другими языками. С достижением нашей республики Независимости, во всех социально-экономических сферах жизни, в том числе и в научной, появились возможности более глубокого изучения многих явлений, основываясь при этом на их собственную сущность. Основываясь на данных позициях, создавались учебные программы, учебные пособия и учебники. В опубликованных учебниках приводятся следующие определения нераспространённому предложения: «Предложение, состоящее только из подлежащего и сказуемого, считается нераспространенным».[2] Аналогичные определения приводятся и в других языках, например, в грамматиках русского, английского и немецкого языков. И это – естественно, так как грамматика узбекского языка развивалась не только под влиянием русского языка, но и некоторых европейских языков. И, как было отмечено выше, только после обретения Независимости появилась возможность исследовать лингвистические явления, основываясь на сущность своего родного языка. Таким образом, за основу простого предложения была взята модель простого предложения [WPm]. Данная теория сформировалась, была апробирована и одобрена, а также разъяснена в учебных пособиях и учебниках нового поколения.[3] Работа была проведена на основе узбекского субстанциального языкознания.

Для более ясного выражения мысли, следует подробнее остановиться на истории этого вопроса. Как и все категории, категория нераспространённости в целом или нераспространённые простые предложения, в частности, имеют свою историю. Считаем, что целесообразным является остановиться на более современном этапе развитияданной категории. В книге «Ўзбек тили грамматикаси. II қисм. Синтаксис», опубликованной в середине XX в., нераспространённым и распространённым предложениям даётся следующее определение: «Биргина фикрни англатиб, фақат бош бўлаклардан тузилган гап йиғик гап» дейилади.[4] Более того, приводится и такое определение:»...биргина фикрни англатиб, иккинчи даражали бўлакка эга бўлган гап ёйиқ гап» дейилади.[4] В грамматиках русского языка, равно как и в других грамматиках зарубежных языков этого периода приведены подобные определения. Наши предшественники, а затем и последующее поколение языковедов, взяли за основу их теорию и начали развивать теорию уже своего родного языка. Очивиден тот факт, что в русских грамматиках именно в то время, то есть в 70-80гг. XX в., много внимания уделялось семантике. В связи с этим, нераспространённые предложения изучались как «элементарные семантические категории». [5] Вместе с тем, предложения, где главные члены согласованы между собой (например: Мама – преподаватель) делятся на две группы, а именно: 1) предложения с

субъектом: Свобода – это осознанная необходимость; Жестокость – та же трусость; 2) предложения, где есть признаки субъекта, его связи и сравнения, например: Человек – это стиль; Каждое дерево – пуд мёда. Данные группы, в свою очерердь, делятся ещё на подгруппы. Предложения таких типов называются самыми малыми семантическими категориями и в них входят как неполные предложения, так и односоставные предложения. Кроме этого, приводятся мысли и о нераспространённых предложениях. Нераспространённые предложения – это такие предложения, которые состоят из двух слов, образуют одну синтагму и даже каждая из них также может образовать синтагму. [5]

В узбекское языкознание термин «нераспространённое предложение» ввёл основатель формального узбекского языкознания Айюб Гулямович Гулямов. И он даёт совершенное определение этой категории. Учёный отмечает, что в основе синтаксиса находится учение о предложении. [2] А в другом выпуске этого учебника А.Г.Гулямов отмечает, что предикативные отношения слов (предикат-сказуемое), обычно, образует предложение. Предикативная связь – отношения подлежащего и сказуемого является ядром предложения. [3] Как известно, в предложении участвуют как главные, так и второстепенные члены. Они создают подчиняемость членов предложения, чем и создают предложения. Эту мысль можно выразить по-иному: «Наименование подлежащего и сказуемого главными членами, а дополнения, обстоятельства и определения второстепенными членами не исходит из того, что вторые – зависимые члены от первых, а такая зависимость является их грамматическим свойством. Оно воспроизводится из речевых функций членов предложения».

Как известно, главные члены могут образовать самостоятельные предложения без второстепенных элементов. Подобные предложения являются нераспространёнными. К примеру: Дилдор ўйланиб қолди.(Саид Аҳмад) — Дилдор задумалась. Йигит ҳам шунақа бўладими?(П.Қодиров) — Разве парень может быть таким? Сиз Тошкентликмисиз? (Ойбек) –Вы из Ташкента? Или вы ташкентец? [6] Оставленные нам в наследство сведения показывают, что предложение, образованное только из подлежащего и сказуемого, является нераспространённым предложением. И оно имеет модель [S→P]. Естественно, данная теория была введена в узбекское языкознание посредством русского языка из западных языков. Очевидно, что в то время не было обращено внимания на свойственные только узбекскому языку особенности. Мы понимаем, что этому есть объективные объяснения, то есть, в тот период развития, как мы не раз ещё повторим, подобные трактовки были необходимы.

В формальном языкознании нераспространённое предложение состоит из двух членов – подлежащего и сказуемого. Это предложение, обычно, соответствует интенции (замыслу) автора и его желанию. Оно может

употребляться в составе сложного предложения. Если даже члены этого предложения будут менятся местами, то всё равно они будут считаться нераспространёнными: Қиш келди. — Келди қиш. (Зима пришла. –Пришла зима; Булбуллар сайради. — Сайради булбуллар.(Соловьи запели. – Запели соловьи). Употребление подобных нераспространённых предложений зависит от автора, который выстраивает их намеренно, с определенной целью, что, естественно, зависит не только от эмоционально-экспрессивных характеристик текста, авторской модальности или прагматики текста, но и эмфазой.

При сравнительном рассмотрении мнений учёных-специалистов английского языка, можно заметить следующую особенность. В грамматике английского языка нераспространённое предложение, каким оно является и рассмотрено в главе «Структурные схемы предложения», именуется как «элементарное предложение». За структурную схему этого предложения взята такая структура, состав которой состоит из минимальной грамматической конструкции и простого содержания. Приведём следующую цитату: «Предложение, состоящее из одного подлежащего и сказуемого, называется простым нераспространённым предложением». [7] Из вышеприведённого следует, что в английском языке нераспространённые предложения называются «элементарными». Обратимся к некоторым примерам: The rain has stopped. It is cold. Основываясь на примерах, можно сказать, что данные простые предложения являются нераспространёнными. Рассмотрим нижеследующие примеры на английском языке:

• A cat is sleeping (Кот спит)

• A cat is sleeping on the chair (Кот спит на стуле)

• The sun was shining, the birds were singing and the girl was happy (Светило солнце, пели птицы, и девочка была счастлива)

В первом предложении присутствуют только подлежащее и сказуемое. Предложение – нераспространенное. Во втором примере присутствует грамматическая основа и дополнение. Предложение – распространенное. А третий пример является сложным предложением. Это – сложносочинённое предложение, состоящее из трёх частей и после каждого из его частей можно поставить точку. Более того, части предложения равноправны.

Местоположение подлежащего и сказуемого в нераспространенных английских предложениях может выглядеть так:

- Подлежащее+сказуемое: Leaves turned yellow.
- Подлежащее+сказуемое, сказуемое: Everything turned green, blossomed.
- Подлежащее+сказуемое, сказуемое, сказуемое: The wolf cubs played, fought, tumbled.

- Подлежащее, подлежащее, подлежащее+сказуемое: The boy, his sister, and his dog went swimming.
- Подлежащее, подлежащее+сказуемое, сказуемое:. Temur and Sarvar came in and sat down.

Выше мы вкратце остановились на некоторых высказываниях о нераспространённых предложениях в языкознании. Подобный экскурс дал нам краткое, но ясное представление о том, что и в русском, и английском, и в узбекском языках имеются подобные предложения и определения, а классификации, данные им, очень даже похожи.

ЛИТЕРАТУРА:

1) Сайфуллаева Р.Р., Сайфуллаев Ш.Р. Би-трилингвистический эффект имени А.Р.Сайфуллаева. Журнал научных и прикладных исследований. №4, 2013
2) Ғуломов А., Асқарова М. Ҳозирги ўзбек адабий тили. Синтаксис. –Т.: Ўқитувчи, 1987.
3) Ўзбек тили грамматикаси. II том. Синтаксис. –Т.: Фан, 1976.
4) Ғуломов А., Маъруфов З., Шермуҳамедов Т. Ўзбек тили грамматикаси. 2-қисм. Синтаксис. –Т.: Ўқувпеддавнашр, 1948.
5) Русская грамматика. Том III. Синтаксис. -М.: 1980
6) Ўзбек тили грамматикаси. II том. Синтаксис. –Т.: Фан, 1976.
7) I.P.Krylova, E.M.Gordon. A Grammar of Present-day English, -M.: Книжны дом «Университет» 2000. 442 б

СЕКЦИЯ 2.
МАШИННЫЙ ПЕРЕВОД

СИСТЕМА РУССКО-ТАТАРСКОГО НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА

А.Ф.Хусаинов, Д.Ш. Сулейманов, Р.А.Гильмуллин,
Академия наук Республики Татарстан, Казанский федеральный университет,
Казань, Россия, khusainov.aidar@gmail.com, dvdt.slt@gmail.com, rinatgilmullin@gmail.com

В 2016 году системы машинного перевода, основанные на нейросетевом подходе, превзошли по качеству работы системы на основе статистики по фразам и синтаксису (phrase-, syntax-based) (Bojar, 2016). С тех пор многие компании разработали нейронные версии своих переводчиков для самых популярных языковых пар (Bojar, 2017; Wu, 2016). Целью данного исследования является создание системы машинного перевода на основе последних достижений в области машинного обучения, которая могла бы хорошо работать в случае малоресурсной татарско-русской языковой пары. Полученная система включает в себя инструменты, которые позволяют искусственно увеличивать объем обучающих данных, выполняют алгоритмы предварительной и последующей обработки наряду с encoder-decoder-attention алгоритмом перевода.

Нейросетевой подход к построению систем машинного перевода подтвердил качество работы в экспериментах со многими языковыми парами. Большинство аспектов перевода успешно моделируются с помощью нейросетевого подхода. Основным ключом к этому является чистый, репрезентативный и достаточно большой параллельный текстовый корпус. Известно, что системы NMT неэффективны при обучении на ограниченных данных (Fadaee, 2017; Zoph, 2016). Таким образом, решением поставленной задачи построения татарско-российской системы машинного перевода было бы создание достаточно большого параллельного корпуса и построение системы NMT. Ограничением здесь является отсутствие параллельного корпуса и небольшое количество ресурсов, на основе которых он может быть построен.

Идея данной работы заключается в разработке инструментов, которые позволят собрать и использовать максимум параллельных данных, доступных для татарского языка. Одним из основных источников двуязычной информации являются сайты министерств и других государственных ведомств. В Республике Татарстан действует закон, обязывающий организации вести документооборот одновременно на русском и татарском языках. Другим источником являются литературные произведения, в основном печатные книги с доступным переводом. Для загрузки данных из веб-источников мы разработали программу, которая может быть настроена на загрузку информации на основе списка сайтов и конкретных правил, которые помогают определять соответствие между русской и татарской страницами (шаблоны URL, ссылки на страницу перевода и т.д.). Мы также подписали соглашение между Академией наук Татарстана и библиотеками о передаче прав на использование некоторых из их книг; имеющиеся книги, для которых был выполнен перевод, были отсканированы с использованием профессионального сканирующего оборудования.

Таким образом, мы оценили возможность использования нейросетевого подхода при построении системы машинного перевода для татарско-русской языковой пары. Мы использовали современные подходы к увеличению данных (data augmentation), обучению нейронных сетей. Основными результатами работы являются создание первой нейронной татаро-русской системы машинного перевода и улучшение качества перевода в этой языковой паре с точки зрения метрики BLEU с 12 до 39 и с 17 до 45 для обоих направлений перевода (по сравнению с существующей системой перевода).

***Ключевые слова**: нейронный машинный переводчик; татарский язык,; малоресурсный язык; увеличение данных.*

NEURAL TATAR-RUSSIAN MACHINE TRANSLATION SYSTEM

A.F. Khusainov, D.Sh.Suleymanov, R. Gilmullin, Academy of Sciences of the Republic of Tatarstan, Kazan Federal University, Kazan, Russia
khusainov.aidar@gmail.com, dvdt.slt@gmail.com, rinatgilmullin@gmail.com

2016 was the year when machine translation systems built on the neural network approach surpassed the quality of the phrase- and syntax-based systems (Bojar, 2016). Since that time, many companies have developed neural versions of their translators for the most popular language pairs (Bojar, 2017; Wu, 2016).

Motivated by the goal of creating a machine translation system that could work well for the low-resourced Tatar-Russian language pair, we propose such a

technology that would include the latest achievements in machine learning. The resulting system includes tools that augment training data, execute pre- and post-processing algorithms along with the attention-based encoder-decoder translation algorithm.

The neural network approach of constructing machine translation systems has confirmed its success in experiments with many language pairs. There are some important language features that affect the quality of the system, for instance, translation from a gender-neutral language (e.g. Turkish, Tatar) to a non-gender neutral one (e.g. Russian) could lead to some biasing problems (Schiebinger, 2013). But most aspects of translation are successfully modelled by the NN approach. The main key to that is a clean, representative and big enough parallel text corpus, as NMT systems are known to under-perform when trained on limited data (Fadaee, 2017; Zoph, 2016). Thus, the solution to our task of constructing the Tatar-Russian MT system would be to create a large-enough parallel corpus and build the NMT system. The limitation here is the absence of a parallel corpus and a small amount of resources from which it could be built.

The idea of this paper is to develop tools that will allow to collect and use maximum of the parallel data available for Tatar. One of the main sources of bilingual information are websites of ministries and other state departments. In Tatarstan there is a law that oblige organizations to keep document circulation simultaneously in the Russian and the Tatar language. The other source are literary works, mostly printed books with available translation. To download data from web sources we have developed a program that can be configured to download information based on sites' list and specific rules that help to determine the correspondence between the Russian and the Tatar pages (i.e. url patterns, translation links on the source page). We have signed an agreement between the Tatarstan Academy of Sciences and libraries on the transfer of rights to use some of their books; available books for which there was a translation were scanned using professional scanning equipment.

To summarize the idea of this paper, we assess the possibility of using the neural network approach to the construction of the machine translation system for the Tatar-Russian language pair. We incorporated modern approaches of data augmentation, neural networks training. The main results of the work are the creation of the first neural Tatar-Russian translation system and the improvement of the translation quality in this language pair in terms of BLEU scores from 12 to 39 and from 17 to 45 for both translation directions (comparing to the existing translation system).

Key words: *neural machine translation; Tatar language; low-resourced language; data augmentation.*

ВВЕДЕНИЕ

2016 год стал годом, когда системы машинного перевода, основанные на нейронных сетях, превзошли по качеству работы систем, основанные на статистике (phrase- and syntax-based) (Bojar, 2016). С тех пор многие компании разработали нейронные версии своих переводчиков для самых популярных языковых пар (Bojar, 2017; Wu, 2016). Более того, большое количество исследований было посвящено улучшению качества перевода за счет использования лингвистически мотивированных или лингвистически информированных моделей, что привело, например, к использованию многофакторных моделей и морфем в качестве базовых единиц перевода (subword units).

Однако, как и в других областях искусственного интеллекта, например, при построении систем распознавания речи или диалоговых систем, использование современных методов машинного обучения для класса малоресурсных языков ограничено отсутствием необходимого объема обучающих данных. Даже компании с относительно неограниченным доступом к данным (например, Google, Yandex) используют различные методы, чтобы обойти это ограничение: объединение различных подходов к переводу и выбор лучшего результата (One model is better than two, 2017), использование промежуточных языков при переводе (например, английского или родственного языка с большим количеством доступных данных).

С целью создания системы машинного перевода, способной демонстрировать хорошее качество работы для татарско-русской малоресурсной языковой пары, мы использовали последние достижения в области машинного обучения. Результирующая система включает в себя инструменты, которые позволяют дополнить обучающие данные (data augmentation), выполняют алгоритмы предварительной обработки текстов совместно с алгоритмом перевода на основе внимания (attention-based encoder-decoder).

В разделе 2 данной статьи представлен обзор процесса сбора параллельного корпуса, в разделе 3 описаны основные характеристики нейросетевой модели, в разделе 4 приводятся результаты проведенного эксперимента, раздел 5 содержит заключение.

Подготовка данных

Нейросетевой подход к построению систем машинного перевода подтвердил качество своей работы в экспериментах для многих языковых пар. Большинство аспектов перевода успешно моделируются нейросетями, при этом ключевое значение имеет репрезентативный параллельный текстовый корпус, поскольку системы нейросетевого машинного перевода,

как известно, недостаточно эффективны при обучении на ограниченном наборе данных (Fadaee, 2017; Wu, 2016). Таким образом, для решения задачи построения системы татарско-русского перевода необходимо создание достаточно большого параллельного корпуса. Ограничением здесь является небольшой объем источников, на основе которых он может быть построен.

Идея данной статьи заключается в разработке инструментов, которые позволят использовать максимум параллельных данных, доступных для татарского языка. Одним из основных источников двуязычной информации являются веб-сайты министерств и других государственных ведомств. В Республике Татарстан действует закон, обязывающий организации вести документооборот одновременно на русском и татарском языках. Другим источником данных являются литературные произведения: печатные книги с доступным переводом. Чтобы загрузить данные из веб-источников, мы разработали программу, которая может быть настроена для загрузки информации на основе списка сайтов и конкретных правил, которые помогают определить соответствие между русскими и татарскими страницами (например, шаблоны url, ссылки на исходной странице). Доступные книги, для которых был доступен перевод, были отсканированы с использованием профессионального сканирующего оборудования.

Собранные данные были отфильтрованы в соответствии со следующими критериями: как исходное, так и переведенное предложения должны содержать от 1 до 80 слов; повторяющиеся предложения были удалены; все собранные тексты были выровнены с помощью инструмента ABBYY Aligner 2.0 (ABBYY Aligner 2.0, 2017).

Мы также провели ручную корректировку результатов процедуры автоматического выравнивания (литературный перевод книг, например, приводил к наличию пар предложений, которые сильно отличаются друг от друга). Данная работа была завершена примерно в течение 1 месяца.

По результатам проведенной работы было собрано 328 тысяч параллельных татарско-русских предложений. Данный объем параллельного корпуса использовался для обучения первой версии нейросетевого переводчика с русского на татарский язык.

Команда переводчиков осуществляла перевод текстов новостной тематики с русского на татарский язык. Процесс был организован с использованием инструмента ABBYY SmartCAT для профессиональных переводчиков (ABBYY SmartCAT tool for professional translators, 2017). Ручной перевод 35 тысяч предложений занял около 700 человеко-часов или 1 месяц работы команды. В течение данного месяца было завершено обучение промежуточных нейронных моделей для русско-татарского направления перевода. С помощью промежуточной системы готовились автоматические переводы всех новых текстов, которые поступали

переводчикам для ручной корректировки. Это позволило существенно ускорить процесс перевода: суммарно за 2 месяца работы общее количество ручных переводов составило 189 689 предложений. Итоговый параллельный корпус объемом 517 тысяч предложений был использован для построения финальной версии русско-татарских моделей.

При обучении моделей для обратного, татарско-русского, направления перевода, был применен back-translation подход для дополнительного увеличения объема обучающих данных. В его основе лежит предположение, что большое количество (до размера исходного корпуса) автоматически переведенных предложений способно повысить качество работы системы (Sennrich, 2016). Таким образом, с помощью созданного нами русско-татарского переводчика были подготовлены дополнительно 409 тысяч пар параллельных предложений, которые вошли в состав обучающего корпуса.

Основные этапы подготовки параллельного корпуса:

- Формирование базового корпуса на основе Интернет ресурсов и литературных произведений;
- Фильтрация, автоматическое выравнивание и ручная корректировка результатов (328 тысяч пар предложений);
- Построение промежуточных систем русско-татарского переводчика;
- Ручной и полуавтоматический перевод новостных текстов на русском языке (189 тысяч пар предложений);
- Построение финальной версии русско-татарского переводчика;
- Расширение обучающего корпуса за счет back-translated алгоритма: с помощью русско-татарского переводчика было автоматически подготовлено дополнительно 409 тысяч пар параллельных предложений;
- Построение финальной версии татарско-русского переводчика.

Описание архитектуры системы

Для обучения системы нейросетевого машинного перевода мы использовали инструментарий Nematus (Open-source neural machine translation in Theano, 2017) с усовершенствованиями, предложенными в (Sennrich, 2017). В основе подхода лежит архитектура сети encoder-decoder-attention, каждая часть которой представляет собой одно- (для случая декодера с механизмом внимания) или двунаправленную (для энкодера) рекуррентную нейросеть. Размерность слоя векторного представления слов была выбрана равной 512, размерность скрытых слоёв – 1000, значение параметра dropout – 0.2. Размер словаря установлен равным 15000, размера группировки batch – 60 для этапа обучения и 5 для этапа валидации. Обучение нейросети проходило с помощью оптимизатора Adam (значение параметра learning rate – 0.0001).

Татарский язык является агглютинативным языком с богатой морфологией, что требует решения проблемы большого количества внесловарных слов (OOV problem) из-за ограниченного размера словаря и данных для обучения. Чтобы преодолеть эту проблему, мы использовали базовые единицы, построенные на основе алгоритма BPE (byte-pair encoding) (Sennrich, 2015). Модель разбиения слов на составляющие части была применена к объединенному русско-татарском корпусу.

Эксперимент

Мы использовали метрику BLEU (Papineni, 2002) для сравнения качества систем машинного перевода. Несмотря на то, что было показано, что значения BLEU не всегда коррелируют с ручной оценкой качества перевода (Baisa, 2009), данный критерий по-прежнему широко используется, поскольку ручное тестирование является дорогостоящим и требует большого количества времени.

Мы провели сравнение качества работы построенной системы машинного перевода и Яндекс.Переводчика (Yandex translate, 2017), поскольку они являются единственными доступными инструментами для русско-татарской языковой пары. Результаты оценки качества перевода представлены в таблице

Для проведения тестирования мы случайным образом отобрали 1000 русских предложений, которые были вручную переведены; данные предложения не вошли в состав обучающей и валидацинной выборок.

Таблица 1. Результаты сравнения качества работы систем машинного перевода.

Система	Направление перевода	Обучающий корпус	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Яндекс Переводчик	РУ-ТТ	Н/Д	12.6	43.8	16.7	8.0	4.3
Представленная система	РУ-ТТ	0.5М пар	39.6	62.8	44.1	34.6	28.7
Яндекс Переводчик	ТТ-РУ	Н/Д	17.2	47.7	21.6	12.0	7.1
Представленная система	ТТ-РУ	0.9М пар	45.7	65.4	49.2	40.7	33.5

Заключение

В этой статье мы представили первую систему русско-татарского машинного перевода, построенную на основе нейросетевых алгоритмов.

Был подготовлен обучающий корпус параллельных текстов, применены современные методы расширения объема данных и машинного обучения. Полученная система перевода значительно превосходит единственную существующую систему перевода в этой языковой паре от компании Яндекса (в 3 раза по показателю BLEU).

Мы планируем использовать многофакторные модели, так как грамматическая информация может помочь улучшить качество перевода, а также продолжить пополнение параллельного русско-татарского корпуса.

ЛИТЕРАТУРА:

1) Bojar,O., Chatterjee,R.,Federmann,C.,Graham,Y.,Haddow,B.,Huck,M.,JimenoYepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., Zampieri, M.: Findings of the 2016 conference on machine translation. In: Proceedings of the First Conference on Machine Translation. pp.131–198.AssociationforComputationalLinguistics,Berlin,Germany (August 2016), http://www.aclweb.org/anthology/W/W16/W16-2301

2) Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., Turchi, M.: Findings of the 2017 conference on machine translation (wmt17). In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. pp.169–214. Association for Computational Linguistics, Copenhagen, Denmark (September 2017), http://www.aclweb.org/anthology/W17-4717

3) Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv e-prints (Sep 2016)

4) Schiebinger, L., Klinge, I.: Gendered Innovations: How Gender Analysis Contributes to Research. Luxembourg: Publications Office of the European Union (2013)

5) Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: n Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL17). pp. 567–573 (01 2017)

6) Zoph, B., Yuret, D., May, J., Knight, K.: Transfer Learning for Low-Resource Neural Machine Translation. ArXiv e-prints (Apr 2016)

7) One model is better than two. Yandex.Translate launches a hybrid machine translation system. https://goo.gl/PddtYn (2017), [Online]

8) ABBYY Aligner 2.0. https://www.abbyy.com/ru-ru/aligner/ (2017), [Online]

9) Sennrich, R., Haddow, B., Burch, A.: Improving neural machine translation models with monolingual data. In: In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 86–96. Berlin (2016)
10) Open-source neural machine translation in Theano. https://github.com/rsennrich/nematus (2017), [Online]
11) 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 forw mt17. In: In Proceedings of the Second Conference on Machine Translation. vol. 2: Shared Task Papers. Stroudsburg, PA, USA (2017)
12) Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Sub-word Units. ArXiv e-prints (Aug 2015)
13) 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)
14) Yandex translate. https://translate.yandex.com/ (2017), [Online]
15) Baisa, V.: Problems of machine translation evaluation. In: In Proceedings of Recent Advances in Slavonic Natural Language Processing. Brno (2009)

MACHINE TRANSLATION FOR KYRGYZ PROVERBS — GOOGLE TRANSLATE VS. YANDEX TRANSLATE- FROM KYRGYZ INTO ENGLISH AND TURKISH

Y. Polat, A. Zakirov, S. Bajak, Z. Mamatzhanova,
Ala-too International University,
Bishkek, Kyrgyzstan

Although Kyrgyz language is old, rich and the bearer of the glorious epic Manas, it has not been well represented in the area of machine translation yet. So far it has shown a comparatively slow progress in Google and Yandex translation services.

This study investigates the accuracy of machine Kyrgyz-to-English translation at lexical, semantic, and syntactic levels. The present study uses Groves and Mundt (2015) Model of error taxonomy to compare Kyrgyz-to-English translations produced by Google and Yandex Translate. We have selected, 50 texts from the domain of proverbs. The proverbs have been translated by Google and Yandex Translate, as well as three human translators and then evaluated with respect to lexical, semantic and grammatical accuracy.

Materials are composed of four groups, they are (a) declaratives, (b) interrogatives, (c) imperatives, and (d) elliptic proverbs. In this study, we have done a descriptive-comparative human analysis of translations based on Groves and Mundt (2015) Model as the criterion for evaluating and scoring the translations made by machine and human translators. The reason for adopting this model is that it allows for detailed analysis and scoring of the translated materials. We have also got benefited from the studies of Saffari, Sajjadi, Mohammadi (2017) and Ghasemi, Hashemian (2015) as the practical models.

Summing up the results, it can be concluded that Google Translate was more accurate than Yandex Translate at lexical, semantic and syntactic levels in translating phrases and sentences from Kyrgyz into English from the domain of proverbs. Error analysis of grammatical items revealed that verb tense, comma, and spelling were the most frequent errors generated by the two machine translation systems.

Key words: *Yandex Translate; Google Translate; machine translation; translation accuracy.*

МАШИННЫЙ ПЕРЕВОД ДЛЯ КЫРГЫЗСКИХ ПОСЛОВИЦ — GOOGLE TRANSLATE VS. ЯНДЕКС ПЕРЕВОД — ИЗ КЫРГЫЗСКОГО В АНГЛИЙСКИЙ И ТУРЕЦКИЙ

Я. Полат, А.Закиров, С. Баджак, З. Мамамтжанова,
Международный университет Ала-Тоо, Бишкек, Кыргызстан

В этой статье рассматриваются проблемы машинного перевода для кыргызских пословиц — Google translate, а также Яндекс перевод — из кыргызского языка в английский и из кыргызского языка на турецкий язык. Рассмотренные материалы доказывают необходимость тщательного сравнительного анализа этих двух переводных систем.

Ключевые слова: *Яндекс перевод, Google translate, машинный перевод.*

1. INTRODUCTION

The use of the automatic machine translation has increased in recent years dramatically in communication between countries (Shankland 2013). Google and Yandex translation services are two powerful translation engines. In a combination of languages such French and English, they are able to give more accurate translations. Peter Newmark said; «a satisfactory translation is not always possible, but a good translator is never satisfied with it. It can usually be improved.» (1988: 8) No matter how Google and Yandex Translate services are improved, there are always some problems to be solved.

Google Translate supports over 100 languages at various levels and as of May 2017, serves over 500 million people daily (Wikipedia), and apparently offers better performance than other machine translation tools available to the public (Seljan, Brkić and Kučiš 2011). Google translation is a widely used translation tool for inexpensive and instant access to general information about the original texts for moderate quality translation (Anazawa et al. 2013). The previous evaluation of Google translation focused on the levels of words, phrases, sentence length, syntactic structure (Seljan, Brkić and Kučiš 2011), intelligibility and usability (Anazawa et al. 2013), and BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2002).

In November 2016, Google announced that Google Translate would switch to a neural machine translation engine — Google Neural Machine Translation (GNMT) — which translates "whole sentences at a time, rather than just piece by piece. It uses this broader context to help it figure out the most relevant translation, which it then rearranges and adjusts to be more like a human speaking with proper grammar." Originally only enabled for a few languages in 2016, GNMT is gradually being used for more languages. (Wikipedia)

An example of statistical machine translation system is Yandex Translate. Yandex Translate translates separate words, complete texts, and as of March

2018, translation is available in 94 languages. In 2015, the Kyrgyz Yandex Translate began (Hees, Kozlowska, Tian, 2015).

The Yandex Translate search engine is comprised of a three-piece model. A translation model, a language model and a decoder. The translation model is simply a list of all known words in a language with their translations into other languages. This means that each language has its own translation model. These translation models are built by cross-referencing texts and works that have been translated into different languages, also called a parallel corpus. First the system looks for correspondent phrases, then to groups of words or single words and uses these to calculate the probability based on previous encounters that the translation is the correct one. It continuously processes new texts to increase the possibility to encounter the word in different contexts, this is also why you need so many sources (Hees, Kozlowska, Tian; 2015).

Yandex machine translation is based on the statistical approach. To learn a language, the system compares hundreds of thousands of parallel texts that translate each other «sentence by sentence». It has two main components: the translation model and the language model. The translation model constructs a graph containing all the possible ways to translate a sentence. The language model selects the best translation in terms of the optimal word combinations in natural language. The translation model learns from extensive bilingual parallel corpora. The language model is built from large single-language corpora and contains all the language's most frequent n-word combinations. N may be from 1 to 7 (usually 5). Yandex uses BLEU (Bilingual Evaluation Understudy) metrics to automatically evaluate the quality of machine translation; it determines the percent of n-grams (n<=4) that match between the machine translation and the standard translation of a sentence. Translations are usually manually rated for two factors, Adequacy and Fluency, using a 5-point scale (tech.yandex.com).

2. KYRGYZ SYNTAX AND MT

Kyrgyz belongs Turkic language family and one of the two official languages of Kyrgyzstan, the other being Russian. Kyrgyz written «кыргыз тили» alternatively written «Kirghiz» or «Kyrgyz» is a Turkic language spoken in Kyrgyzstan, China, Tajikistan, and Uzbekistan. (Washington, Ipasov, Tyers, 2011) Kyrgyz is a Turkic language, like Kazakh, Uzbek, Turkish, Uyghur, and Tatar. Kygyz is one of the major languages of the Kipchak sub-branch of the, Turkic languages, like neighboring Kazakh. Like many other Turkic languages, Kyrgyz has vowel harmony whereby the vowels of suffixes change to fit the other vowels in the stem. Kyrgyz is also an agglutinative language, where each suffix added to the stem indicates only one meaning; these suffixes attach to the word stem one after another in a set order. Generally, Kyrgyz is divided into two distinct dialects, the Northern and Southern. Standard Kyrgyz is mainly based upon the Northern dialect (Indiana.edu). Kyrgyz is spoken by inhabitants of Kyrgyzstan, Xinjiang, Afghanistan, Kazakhstan, Tajikistan, Turkey, Uzbekistan,

Pakistan, and Russia. Due to Soviet policy, the Cyrillic alphabet became the most common alphabet for writing Kyrgyz and has remained so to this day, though some Kyrgyz still use the Arabic alphabet, particularly in the People's Republic of China (Indiana.edu).

Several linguistic features of Kyrgyz as Göksel A. and Kerslake C, (2005) stated can directly affect the performance of a MT system: (i) agglutinative morphology, (ii) vowel harmony and other phoneme alternation phenomena, and (iii) word order. Whereas the first two features are situated at the word level, the third concerns syntax and the global structure of sentences. In this work, our analysis focuses on word and sentence-level preprocessing.

2.1. Agglutination

Agglutination implies that the vocabulary is built by a wide range of basic suffix combinations. A Kyrgyz word can thus correspond to a single English word, up to phrases of various length, or even to a whole sentence as shown in Table 2. Differences in token range can be observed in the IWSLT training parallel corpus. Table 1: Example of Kyrgyz suffixation:

Үй: ‘home’

Үйүм: ‘my home’

Үйүмдө: ‘in my home’

Үйүмдөмүн: ‘I am in my home’

Given this premise, it is easy to imagine how interlanguage alignments and in general any modeling of the language based on the notion of token may suffer from data sparseness. That is why morphological segmentation is needed as preprocessing. (Arianna Bisazza, Marcello Federico: 2009)

2.2. Vowel Harmony

On a phonological level vowel harmony and other phoneme alternation phenomena systematically lead stems and suffixes to have several surface forms – i.e. allomorphy. For example (see Table 3) the possessive suffix -(I)m ‘my’ can have four different surface forms depending on the last vowel of the word it attaches to (ex.1-4), plus one if it is attached to a word ending with vowel (ex.5). Table 2: Different surface forms of possessive suffix -(I)m.

1) чач + (ым): чачым: ‘my hair’

2) эр + (им): эрим: ‘my husband’

3) кол + (ум): колум: ‘my hand’

4) гөз + (үм): гөзүм: ‘my eye’

5) байке + (м): байкем: ‘my brother’

If we envisage treating suffixes as single tokens, we foresee an additional increase of data sparseness due to suffix allomorphy. To cope with this problem, we need to introduce an abstract notation that factorizes different surface forms of the same suffix into one single form.

2.3. Word order

The typical structure of Kyrgyz phrases is head-final. Sentences mostly belong to the subject-object-verb (SOV) kind, but word order is relatively free and discourse-related phrase movements are quite frequent. As a result alignments between Kyrgyz and English are far from being monotonic, as shown by this example taken from the IWSLT09 corpus:

Бардык адамдар өз беделинде жана укуктарында эркин жана тең укуктуу болуп жаралат.

Bardıq adamdar (*All human beings*) öz bedelinde (*in dignity*) jana (*and*) uquqtarında (*rights*) erkin (*free*) jana (*and*) teñ uquqtuu (*equal*) bolup jaralat. (*are*)

All human beings are ***born*** free and equal in dignity and rights.

Although reordering rules seem hard to describe without using any syntactic information, we believe that morphological segmentation is a first necessary step to take in order to enable machine learning of refined alignments and complex word reordering patterns.

Machine translation between English and Kyrgyz is a challenging task, due to the strong differences between languages. In particular, Kyrgyz has the rich agglutinative morphology, and the word order that differs from English and Russian (SOV in Kyrgyz, SVO in English). This study investigates challenges in translation of Kyrgyz language into English in Google and Yandex Translate engines. It shows the difficulties during the translation from Kyrgyz language to English language and vice versa. Being junior in Google and Yandex Translate services, Kyrgyz language is full of challenges in order to give an accurate translation.

2.4. kkWaC: Kyrgyz corpus from the web

The **Kyrgyz Web Corpus** (**kkWaC**) is a Kirghiz corpus made up of texts collected from the Internet. The corpus was prepared according to standards described in the document *A Corpus Factory for Many Languages* (Kilgarriff et al. at LREC 2010). Data were downloaded in January 2012 with the total size 19 million words. Texts were cleaned and deduplicated. The Kyrgyz (Kirghiz) language belongs to the Turkic languages.

Table 1. The overview of Turkic corpora

LANGUAGE	WORDS	DOCUMENTS (in thousands)	DATA UPDATES
AZERBAIJANI	94 million	365 thousand	Jan 2012
KAZAKH	139 million	378 thousand	Jan 2012
KYRGYZ	19 million	67 thousand	Jan 2012

TURKISH	3.38 billion	12 million	Dec 2011, Jan 2012
TURKMEN	2 million	5 thousand	Jan 2012
UZBEK	18 million	57 thousand	Jan 2012

Source data: The source texts were crawled by the SpiderLing web spider in December 2011 and January 2012. The crawling was constrained to the top-level internet domains corresponding to the countries where the selected languages are officially spoken (.az, .kz, .kg, .tr, .tm, .uz), several exceptions were allowed.

3. NEURAL MACHINE TRANSLATION

Neural machine translation (NMT) is an approach to machine translation that uses a large artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model.

Deep neural machine translation is an extension of neural machine translation. Both use a large neural network with the difference that deep neural machine translation processes multiple neural network layers instead of just one ["Deep Neural Machine Translation". Omniscien Technologies. Retrieved 2017-11-08.].

NMT departs from phrase-based statistical approaches that use separately engineered sub components. (Wołk, Krzysztof; Marasek, Krzysztof, 2015)

Neural machine translation (NMT) is not a drastic step beyond what has been traditionally done in statistical machine translation (SMT). Its main departure is the use of vector representations ("embeddings", "continuous space representations") for words and internal states. The structure of the models is simpler than phrase-based models. There is no separate language model, translation model, and reordering model, but just a single sequence model that predicts one word at a time. However, this sequence prediction is conditioned on the entire source sentence and the entire already produced target sequence. [Philipp Koehn (2016-11-30). "The State of Neural Machine Translation (NMT)". Omniscien Technologies. Retrieved 2017-11-08.]

NMT models use deep learning and representation learning.

The word sequence modeling was at first typically done using a recurrent neural network (RNN). A bidirectional recurrent neural network, known as an encoder, is used by the neural network to encode a source sentence for a second RNN, known as a decoder, that is used to predict words in the target language. (Dzmitry Bahdanau; Cho Kyunghyun; Yoshua Bengio, 2014). "Neural Machine Translation by Jointly Learning to Align and Translate."

Yandex Translator announced that they have a hybrid model: its translation provides a statistical model and a neural network. After this, the algorithm CatBoost, which is based on machine learning, will select the best of the obtained results.

4. Related work

Machine translation has recently entered the realm of many students and researchers, and today, great efforts have been devoted to this new technology. Many investigators have evaluated the output quality of machine translation. Aiken and Balan (2011) evaluated, for the first time, the quality of Google Translate using 50 different languages. Based on their results, machine translations' accuracy in European languages was better than Asian languages. Anna Paananen Dhakar, made a Comparative Analysis of Yandex and Google Search Engines (2012). She concluded that: *The analysis shows that machine learning algorithms retrieve better results for local services and informational searches, while the importance of the page, based on a link analysis leads to better results for commercial searches. For both response time and precision, Yandex proved to be a better performer than Google*.

A master work in Kyrgyz above several diploma works on MT called «Software for text translation from Turkish to Kyrgyz» belongs to Bahoriddin Duşabayev from Manas University in Kyrgyzstan. done in 2010. Second paper called «A finite-state morphological transducer for Kyrgyz» by Jonathan North Washington, Mirlan Ipasov, Francis M. Tyers, Presented in LREC conference in 2011. The paper described the development of a free/open-source finite-state morphological transducer for Kyrgyz. The transducer has been developed for morphological generation for use within a prototype Turkish→Kyrgyz machine translation system, but has also been extensively tested for analysis.

Next paper, «Statistical machine translation implementation and performance tests between Kyrgyz and Turkish Languages.» By Nakılay Tayirova, Mehmet Tekerek and Ulan Brimkulov, 2015. In this work N-GRAM based and Phrase based Statistical Machine Translation methods were applied between Kyrgyz and Turkish languages, using limited training data. For both methods, the translation quality was evaluated with BLEU scoring algorithm. According to test results, both applied methods, provided translations with poor translation quality, 0.1 (one hundredths) on average. In most cases, no translations obtained, or human translation incompatible results were observed. In order to reach higher translation qualities, various suggestions were proposed.

«Kyrgyz Orthography and Morphotactics with Implementation in NUVE» by Züleyha Yiner, Atakan Kurt, Kalmamat Kulamshaev, Harun R. Zafer. In 2016. This new description will enable the implementation of a morphological machine translator between Kyrgyz and Turkish languages.

The first work on an MT system between English and Turkish was in 1981, in a M.Sc. thesis (Sagay, 1981). This work has been developed into an interactive English to Turkish translation system, *Cevirmen*. Turhan describes a transfer-based translation system from English to Turkish (Turhan, 1997), and an interlingua-based approach for translation from English to Turkish is shown by Hakkani et al. (Hakkani et al., 1998). There has also been a work on

implementing a wide-coverage grammar for Turkish: Cetinoglu and Oflazer state the work of developing a Lexical Function Grammar for Turkish (Ozlem Cetinoglu and Oflazer, 2006). Morphological preprocessing of Turkish has been investigated by Oflazer and El-Kahlout. They describe the initial explorations of a Statistical MT system from English to Turkish (Oflazer and İlknur Durgar El-Kahlout, 2007). Arianna Bisazza and Marcello Federico (2009) made an analysis on the complex morphology of Turkish by applying different schemes of morphological word segmentation to the training and test data of a phrase-based statistical machine translation system.

The studies reviewed above are typical examples of research on machine translation in Kyrgyz, Turkish and other languages. Based on the review, however, there is an information gap on the quality of machine translation from English into several languages at lexical, semantic and syntactic levels. Hence, in an attempt to fill in the gap, the current study evaluated the efficiency of Google Translate and Yandex Translator in translating lexical, semantic, and syntactic features from Kyrgyz into English in the domain of Paremiology. To this end, the following research question was formulated:

Is Google Translate lexically, semantically and syntactically more accurate than Yandex Translator, when translating academically oriented phrases and sentences from Kyrgyz into English?

5. METHOD

5.1. Sample selection

The research based on several methods to reach its goal. I have used descriptive, analytical, logical and contrastive methods. A total of 50 Kyrgyz proverbs were selected from different proverb books, currently serving as academic books at different universities in Kyrgyzstan. The domain under study represented different academic fields of Paremiology. The fields are unique in terms of the academic genre common to each academic community under investigation.

The number of words served as the main criteria for the selection of phrases and sentences from the academic domain under study. As such, the selected materials were composed of four groups called (a) *declaratives*, (b) *Interrogatives*, (c) *Imperatives*, and (d) *Elliptic proverbs*.

5.2. Evaluation and scoring procedure

The evaluation of the quality of machine translators was performed by human assessment based on Groves and Mundt (2015) Model, which served as the study's yardstick for evaluating and scoring the translations made by machine and human translators. The reason for adopting this model is that it allows for detailed analysis and scoring of the translated materials. The Model was slightly modified to suit the objectives of the study more effectively. In the original model, there were two categories of 'pronoun incorrect' and 'pronoun reference', but the current study used only 'pronoun incorrect' for the sake of simplicity. Other categories such as

verb tense, apostrophe, sentence structure, run on, and comma splice were excluded from the study because they were not in the scope of the present study, and some of them were not applicable to both languages (e.g., apostrophe). In addition, a new category, named 'unclear', was added to the Model. The addition of this item was because the present study aims to evaluate the semantic accuracy of machine translation. So wherever the semantic concept was not clear, the evaluators used the term 'unclear' to indicate lack of clarity. Besides, because the objective of this study was to assess the accuracy of machine translation lexically, semantically and syntactically, in line with Groves and Mundt (2015) Model, the researcher categorized the errors into three groups: (1) lexis, (2) semantics, and (3) syntax (Table 1).

Table 2. Errors category in the present study (adapted with some modifications from Groves and Mundt (2015) Model

Types of errors	Our category	Explanation
Lexis	Wrong word + word choice + missing word	Incorrect meaning for context + not exactly 'wrong' but could be clearer or more appropriate + missing word(s)
Semantics	Word order + unclear	Incorrect word order + incomprehensible meaning (based on five-point Likert scale)
Syntax	Verb tense	Time is incorrect
	Verb form	Part of speech of noun is incorrect
	Article	Article is missing, unnecessary, or incorrect
	Plural	Noun plural marker is missing, unnecessary, or incorrect
	Agreement	Subject and verb do not agree in number
	Preposition	Wrong preposition
	Comma	Commas missing or unnecessary
	Spelling	Untranslated words or translated with wrong spelling
	Pronoun incorrect	Pronoun used is incorrect for sentence
	Fragment	Incorrect incomplete sentences

5.3. Procedure

All the texts under study (n=240) were first translated into Kyrgyz by two expert translators and then by Google Translate and Yandex Translate individually. The translators, as native speakers of Kyrgyz, with a Ph.D. degree in English language and literature, had full command of translation in both Kyrgyz and English. The Kyrgyz translations done by the two experts served as the benchmark against which machine translations could be judged. All the machine translated materials were given to the evaluators to score based on error category available in Table 5. Error analysis, with a five-point Likert scaling, served as the main criterion for scoring the vocabulary and the structure accuracy of the texts translated by Google Translate and Yandex Translate. As such, the lexical and

syntactic correctness of the machine translations were evaluated and scored based on the number of errors identified in the texts. However, for semantic accuracy, the five-point Likert scaling served as the scoring measure, with 4 indicating the best translation, 3, very good, 2, average, 1, poor, and 0, very poor translation. In this study, all the translations within a score range of 0-2 were considered 'incomprehensible', receiving a negative point. Following the evaluation and scoring results, the data were analyzed and then tabulated.

5.4. Statistical analysis

Chi-square ($\chi 2$) and t-test, from SPSS version 21.0, were used to analyze the data. The tests made a comparison between (a) the machine translations and human translation and (b) between Google Translate and Yandex Translate. The computations were intended to find out if any significant differences in the results would emerge. The two types of computations (i.e. both Chisquare ($\chi 2$) and t-test) yielded similar results. Therefore, to avoid repetition, only Chisquare ($\chi 2$) computations was reported in the Result section. Probability (p) values were considered statistically significant, if less than .05.

6. RESULTS

The study used a quantitative design to investigate the possibility of any significant difference in the quality of translations done by Google Translate and Yandex Translate from English into Kyrgyz considering Groves and Mundt (2015) Model. The translations done by the four human experts just served as the benchmark against which the Google and Yandex translations were judged and scored.

6. The domain of Paremiology

As shown in Table 3, the highest lexical error frequency can be observed in Yandex-translated sentences and phrases. There were total 120 errors done in Google Translate but 167 errors in translation of sentences produced by Yandex Translate.

Table 3. Frequency and percentage of Kyrgyz -to- English lexical errors in Google- and Yandex -translated proverbs

		Declaratives		**Interrogatives**		**Imperatives**		**Elliptic Proverbs**	
	Total of Words	*f*	*%*	*f*	*%*	*f*	*%*	*f*	*%*
		43		70		44		59	
Google		23	56	28	40	31	70	38	65
Yandex		37	86	44	70	41	90	45	76

MT, machine translation; f, error frequency

As Table 4 shows, the highest semantic errors belonged to Interrogatives and Elliptic proverbs translated by both Google translate and Yandex Translate. The percentages of errors in all of these texts were 100%.

Table 4. Frequency and percentage of Kyrgyz to English semantic errors in Google — and Yandex-translated proverbs

		Declaratives		**Interrogatives**		**Imperatives**		**Elliptic Proverbs**	
	Total of sentences	*f*	%	*f*	%	*f*	%	*f*	%
		13		11		11		13	
Google		2	15	11	100	10	91	13	100
Yandex		2	15	11	100	11	100	13	100

MT, machine translation; f, error frequency

Based on Table 5, word form, fragment and agreement in the sentences translated by Google Translate, had the highest frequency, (f = 19, 8 and 13) whereas in the Yandex translated sentences, word form, agreement, spelling and fragment were the highest (f = 22, 13, 24 and 13), respectively.

Table 5. Frequency and percentage of Kyrgyz-to-English syntactic errors in Google- and Yandex translations of the Proverb texts

		Declaratives		**Interrogatives**		**Imperatives**		**Elliptic Proverbs**	
MT	**Error type**	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**
Google	VT	—	—	—	-	4	10	2	3
	WF	6	15	11	16	13	30	19	32
	ART	3	7	5	7	4	10	2	3
	PL	1	2.5	4	6	3	7	3	5
	AGR	4	10	5	7	8	19	8	14
	PREP	4	10	5	7	5	11	4	7
	COM	2	5	4	6	2	5	4	7
	SP	1	2.5	2	3	1	2	4	7
	PRO	1	2.5	2	3	—	-	—	-
	FRAG	7	16	6	9	8	19	13	22

Total of error words		27/43		44/70		48/44		59/59	
Yandex	VT	2	5	1	1.5	1	2	-	-
	WF	32	74	29	41	24	55	20	34
	ART	2	5	3	4	—	-	1	2
	PL	—	—	3	4	2	5	1	2
	AGR	12	28	11	16	11	25	13	22
	PREP	-	-	4	6	4	10	5	8
	COM	—	-	2	3	2	5	2	3
	SP	24	35	15	21	12	30	12	20
	PRO	—	-	—	—	-	-	1	2
	FRAG	12	28	11	16	11	25	13	22
Total of error words		**84/43**		**79/70**		**228/44**		**68/59**	

MT, machine translation; f, error frequency; Boxes show high-frequency errors.

7. Discussion

Based on the computations, Google Translate is more accurate than Yandex Translate. A reason for Google's better performance may be that it has a better corpus than Yandex Translate.

The result of semantic translation of the texts by the two systems revealed not much difference between Google Translate and Yandex Translate. On the contrary the study on semantic translations of proverbs from Turkish into English by Google Translate and Yandex Translate are reasonably more accurate. From the perspective of syntactic translations, the result indicated that Google Translate could outperform Yandex Translate. However, in translation of Declaratives and Interrogatives, the two systems were similar.

Error analysis results of the domain of Paremiology in syntactic translations revealed remarkable results. *Verb form*, *agreement*, and *spelling* were the most common forms of errors in all domains. Both translation systems could not handle the *proverbs* very well. The wrong use of *proverbs* by machine translators might be due to the fact that Kyrgyz language is one of a low resourced Central Asian Languages. Both translation systems indicated some errors in the area of *spelling*. However, in all domains and directions, the wrong use of *spelling* by Yandex

Translate was three times more than Google translate. Most of the *spelling* errors were related to the words that were not translated in source text and left intact in the target text. This problem is more likely due to the word limitation of the dictionary used by Yandex Translate. It was very noticeable that the words that Yandex Translate was not able to translate, and hence were left intact in the target text, were mostly of Kyrgyz origin. *Comma* was another area that the two systems manifested some errors. However, Yandex Translate performed better than Google Translate because Google Translate uses comma exactly the same as the source text without any flexibility, whereas Yandex Translate seems to be more flexible in choosing the comma.

In summary, Google Translate was more accurate than Yandex Translate at lexical, semantic and syntactic levels in translating phrases and sentences from Kyrgyz into English from the domain under investigation. Error analysis of grammatical items revealed that *verb form*, *agreement*, and *spelling* were the most frequent errors generated by the two machine translation systems.

5. Conclusion

The findings of the present study show that machine translation is not yet as accurate as human translation in translating texts from Kyrgyz into English translation. However, so long as machine translation systems use very simple sentence structure, they can render suitable translations in different academic domains, although in some certain domains their translations could be even more appropriate. It is worth indicating that from the viewpoint of a researcher or student, with some knowledge of the language, the translations made by these systems, though occasionally patchy, will be reasonably understandable and accordingly very helpful as they could help the individual to get some knowledge of the text.

There are various machine translation systems with different levels of output. It depends on the researchers and students to decide which one is the best for their purposes and needs. For instance, for translation of mass media texts, Google Translate could be a relatively reliable tool, which can be good news for the researchers and students of journalism with limited command of English. Nonetheless, it is essential for the users to know about the shortcomings of the system as well.

REFERENCES:

1. Aiken, M., & Balan, S. (2011). An analysis of Google Translate accuracy. Translation Journal and the Author, 16(2). Retrieved from http://translationjournal. net/journal/56google.htm.
2. Albat, T. F. (2012) "Systems and Methods for Automatically Estimating a Translation Time." US Patent 0185235.
3. Anazawa, R. &, Ishikawa, H., Park, M. J., and Kiuchi, T. (2013). Online Machine Translation Use with Nursing Literature: Evaluation Method and Usability. Computers Informatics Nursing 31(2): 59-65.

4. Assylbekov, Z. & Nurkas, A.: (2014) Initial explorations in Kazakh to english statistical machine translation. In: The First Italian Conference on Computational Linguistics.
5. Assylbekov, Z., & Washington, J., and others (2016) A free/open-source hybrid morphological disambiguation tool for Kazakh. In: TurCLing 2016. 18–26
6. Bisazza A, & Federico M. (2009) Morphological Pre-Processing for Turkish to English Statistical Machine Translation Proceedings of IWSLT 2009, Tokyo – Japan.
   Statistical Machine Translation
7. Cronin, M. (2013). Translation in the Digital Age. New York: Routledge.
8. Duşabayev, B. (2010) «Software for text translation from Turkish to Kyrgyz» Manas University. Kyrgyzstan.
9. Ghasemi, H., & Hashemian, M. (2016) A comparative study of Google Translate Translations: an error analysis of English-to-Persian and Persian-to-English Translations English Language Teaching. Canadian Center of Science and Education, 9(3). doi: 10.5539/elt.v9n3p13 Saffari.
10. Göksel A. and Kerslake C. (2005) «Turkish. A Comprehensive Grammar», London and New York, Routledge. p. 68–97, 103–108, 195–196, 284–285.
11. Groves M., & Mundt (K. (2015). Friend or foe? Google Translate in language for academic purposes. English for Specific Purposes, 37: 112-121. Retrieved from http://www. sciencedirect.com/science/article/pii/S088949061400060X
12. Hakkani, D. Z., &, G., Oflazer, K., Mitamura, T., and Nyberg, E. (1998). An English-to-Turkish interlingual MT system. In Proceedings of AMTA'98: Conference of the Association for Machine Translation in the Americas, pages 83–94.
13. Hees M. van, & Kozłowska P, Tian N. (2015) Web-based automatic translation: the Yandex.Translate API
14. Koehn, P. (2010). Statistical Machine Translation. New York: Cambridge University Press.
15. Ozlem Cetinoglu & Oflazer, K. (2006). ¨ Morphology-syntax interface for Turkish LFG. In ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pages 153–160.
16. Oflazer, K. & İlknur D. E. (2007). Exploring different representational units in English-to-Turkish statistical machine translation. In Proceedings of the Second Workshop on Statistical 65 Machine Translation, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.
17. Papineni, K., & Roukos, S., Ward, T., and Zhu, W. J. 2002. BLEU: A method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th

Annual Meeting on Association for Computational Linguistics, 311-318. Philadelphia, PA.
18. Sagay, Z. (1981). A computer translation from English to Turkish. Master's thesis, Middle East Technical University (METU), Department of Computer Engineering
19. Sajjadi, M. (2017) Evaluation of Machine Translation (Google Translate vs. Bing Translator) from English into Persian across Academic Fields/ Modern Journal of Language Teaching Methods (MJLTM) ISSN: 2251-6204 www.mjltm.org
20. Shankland, S. (2013). Google Translate Now Serves 200 Million People Daily. CNET. Retrieved from http://news.cnet.com/8301-1023_3-57585143-93/googletranslate-now-serves-200-million-people-daily/
21. Seljan, S., & Brkić, M., & Kučiš, V. (2011). Evaluation of Free Online Machine Translations for Croatian-English and EnglishCroatian Language Pairs. In Proceedings of the 3rd International Conference on the Future of Information Sciences: INFuture2011-Information Sciences and e-Society, 331-345. Zagreb, Croatia.
22. Tohmetov, T. A. et al. 2014. The Problems of Machine Translation. Молодежь и современные информационные технологии: сборник трудов XII Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых 2: 267- 268.
23. Turhan, C. K. (1997). An English to Turkish machine translation system using structural mapping. In Proceedings of the Fifth Conference on Applied Natural Language Processing, pages 320–323.
24. Vít Baisaa, & Vít Suchomela: (2015) 3rd International Conference on Computer Processing in Turkic Languages (TURKLANG 2015); a NLP Centre, Masaryk University, Brno, Czech Republic b Lexical Computing Ltd, Brighton, UK. Turkic language support in Sketch Engine.
25. Weaver, W. (1955). Translation (1949). Reproduced in W.N. Locke, A.D. Booth (eds.). Machine Translation of Languages, 15–23. MIT Press.
26. http://www.indiana.edu/~iaunrc/ceatiu/languages/kyrgyz-0
27. https://en.wikipedia.org/wiki/Machine_translation
28. https://tech.yandex.com/translate/doc/intro/concepts/how-works-machine-translation-docpage/
29. https://www.sketchengine.co.uk/kywac-kyrgyz-corpus/

COULD MACHINE TRANSLATION REPLACE TRANSLATORS

S .N. Bekniyazova, Samarkand state institute of foreign languages, Samarkand, Uzbekistan, sevara.bekniyazova@mail.ru

The article is devoted to the linguistic issues of oral and written machine translation, which is actual at the 21st century because of several factors as human mobility and process of globalization. It is known that machine translation comes across with many peculiarities of the translated language while decoding oral speech and transferring it into written form. The main issue discussed in this article is whether functional and software abilities of computers can replace educated translators in the nearest future.

Key words: *linguistic issue, machine translation, decode, software*

СМОЖЕТ ЛИ МАШИННЫЙ ПЕРЕВОД ЗАМЕНИТЬ ПЕРЕВОДЧИКОВ

С.Н. Бекниязова, Самаркандский государственный институт иностранных языков, Самарканд, Узбекистан, sevara.bekniyazova@mail.ru

Статья посвящена лингвистическим проблемам устного и письменного машинного перевода, который является актуальным в XXI веке из-за нескольких факторов, таких как человеческая мобильность и процесс глобализации. Как известно, машинный перевод сталкивается со многими особенностями при переводе на другой язык во время декодирования устной речи и трансформации ее в письменную речь. Основная цель данной статьи заключается в рассмотрении вопроса смогут ли функциональные и программные возможности компьютеров в ближайшее время заменить образованных переводчиков.

Ключевые слова: *лингвистическая проблема, машинный перевод, декодировать, программное обеспечение.*

MASHINA TARJIMA TARJIMONLARNING O‘RNINI EGALLAY OLADIMI?

S. N. Bekniyozova, Samarqand davlat chet tillar instituti, Samarqand, O‘zbekiston, sevara.bekniyazova@mail.ru

Maqola XXI asrda insoniy harakat va globallashuv jarayoni kabi bir nechta omillar ta’sirida shakllanayotgan og‘zaki va yozma mashina tarjimasining

lingvistik muammolariga bag‘ishlangan. Ma’lumki, kompyuter tilidagi tarjimalar boshqa tillarga tarjima qilinayotganda og‘zaki nutqni aniqlab, uni yozma nutqga aylantirganda ko‘plab qiyinchiliklarga duch keladi. Ushbu maqolaning asosiy maqsadi — bu yaqin kelajakda kompyuterlarning funktsional va dasturiy qobiliyatlari o‘qimishli tarjimonlarning o‘rnini bosa olish ehtimolini yoritishdan iborat.

Tayanch so‘zlar: *lingvistik muammolar, mashina tarjimasi, nutqni aniqlash, dasturiy qobiliyat.*

A modern world is characterized by globalization and integration of this process into the world economy. Most people are interested in establishing contacts for future business projects, in meeting and communicating with new interesting people, in general, for improving the quality of life. Getting what you want has becoming easier and much faster with the advent of computers and the Internet. All these have a tremendous impact on such field of study as translation, because it is becoming an integral part of communication and business with foreigners. As a result, at present, the services of translation companies, bureaus and agencies are increasingly in demand than they were before. On the other hand, not everyone has enough resources and time to afford to hire an interpreter. Demand creates supply, so many high-tech representatives also do not stand aside, making serious competition to translation firms. Companies such as Google, Microsoft and many others have expressed themselves in both written machine translation and oral, developing mobile applications that allow us to translate and reproduce sentences. Such a sharp jump happened due to technical innovations of the 21st century, which was initially influenced by the global coverage of the Internet, as well as a significant expansion of the capabilities of PC and new gadgets.

A new splash of interest in machine translation systems can be observed due to the development of the Internet network. This is because millions of people who speak in different languages have found themselves in a single information space. English dominates in the web, but there are users who do not know it, as, indeed, there are many Web pages written not in English. Additional functions of the browsers, which immediately translate selected fragments of the web page, viewed by the user, appeared to facilitate observing Internet pages in the unknown foreign languages. It is enough just to select a part of the text with a mouse and to transfer it to a special panel, or to press a pointer to a special menu button. Thus, the question arises: is it possible for a computer to replace an interpreter and how soon will this happen.

The answer is yes and no. With the help of machine translation, it is easy to translate texts, it can save a lot of time, if they are simple sentences with rather primitive lexicon. If the computer is used to translate literary texts, then it cannot convey the beauty of the author's written text, because the computer is not a

person, it is not able to recognize the feelings and meanings of the stated words. If we are talking about the translation of technical texts, choosing the right technical dictionary, we can have a good result, not requiring further editing. Meanwhile, need to edit computer translation appears because of the following reasons:

1. A computer can not think with images. Electronic translators adequately translate simple parts of speech, but make mistakes while translating cases, adjectives, speech turns, sentence constructions. The disadvantage of machine translations is inaccurate translation of words that have several meanings, in particular ambiguity of words.

2. A computer does not have the ability to operate with the realities of different cultures and epochs, as human brains can do.

3. A computer is not able to decode all capital letters and abbreviations. When a word begins with a capital letter, its translation will also begin with a capital letter. A word, consisting entirely of such letters, should be written with capital letters too. In English language literature, there are often externally effective abbreviations that can be read as one word. Such abbreviations will be translated with one word.

4. The computer can not always recognize spoken speech. The main component here is not only to catch words, but also to convey the right meaning, which can not be done with a set of specific grammatical rules and richness of computer vocabulary. For this reason, despite the fact that all words in the sentence will be familiar, machine translation can give a completely wrong translation. Therefore, it is difficult to imagine that it is possible to use machine translation during negotiations, where it is very important to translate, smoothing out sharp corners, because often the future of two or even more countries depends on such negotiations.

For the stated reasons, the use of modern machine translation systems is carried out within the framework of human-machine interaction. Not dwelling on the case, when a person uses various fields of knowledge (electronic dictionaries, encyclopedias, etc.) to translate texts "manually", let us consider the role of a human in the use of machine translation systems. In this case, human intervention into the translation process is possible at different stages of the translation process. The following strategies are distinguished:

1) automatic translation with pre-editing: includes transformation of the text before entering it into the translation system in the way that the edited text on the lexical-semantic and grammatical levels approximates to the translation language or the construction of the source language, the transformation rules of which are formalized and known to the used system;

2) automatic translation with post-editing includes work with a "rough" translation of the previously unedited text, made by the machine, to bring it into line with the norms of the target language;

3) automatic translation with inter-editing implies the interaction of the person and the machine in the process of the translation; the person in this case solves difficult cases "online" (for instance, translating lexically ambiguous units determines which unit should be used in each case). The editing process (on the each stage of automatic translation) is a similar process, but transformations are actually applied within the same language — the source language in the case of pre-editing or the target language in the case of inter and post-editing. At the same time, these processes differ a little from the realization of the linguistic personality of the translator, in particular on the verbal-semantic level. In the case of pre-editing, the translation is limited with the rules of the system — as a result, the output text does not show the result of the influence of different levels of the translator's language identity on the text, but rather the result of adapting the translator's language competence to the restrictions used in the translation system. On the verbal-semantic level, it is rigidly came out, because any translation system has a limited thesaurus and set of rules. The language model in systems takes into account only statistically most probable lexical units and language constructions, which are likely in the relationship of incomplete intersection with the linguistic competence of the translator. The process of inter-editing involves, as a rule, resolution of uncertainty in the case of polysemantic words, a choice of the most suitable grammatical constructions from possible ones. In this editing process, the verbal-semantic level of the translator’s linguistic personality is also largely limited with the thesaurus system. The translator only chooses the variant proposed by the system. At the same time, with the implementation of the inter-editing strategy, more flexibility and individual language features of the interpreter are realized. In order to optimize the process, some systems carry out parametrization of some transformations; machine translation before translating the text can adjust some settings (whether to translate it into English by using passive form, a sentence with a formal subject "it", an imperative and etc., if there is no subject in the sentence of the original language). It saves some efforts in translation, on the one hand and personalizes the translation of the text on the other. While post-editing, the linguistic personality of an interpreter fully manifests itself, as an interpreter has the right to edit the text by means of any ways, to change radically the structure and lexical composition of the final version of the translation. The only restriction here is the factor of optimality of the translation process in time and the effort spent on proofreading. Machine translation is used to save time and labor, and therefore, the realization of the linguistic identity of the translator at the verbal-semantic level is limited by an expediential factor. At the same time, with any machine translation strategy, the cognitive and motivational levels of the linguistic personality of the translator are

fully realized, because of the necessity of transformations at any stage of editing is the need to preserve the adequacy of the translation, its correspondence to the original. It happens, in turn due to the conceptual picture of the world of the translator. In general, modern machine translation systems should note two tendencies, the vectors of which are directed in 2 opposite sides: maximum automation of translation processes and personalization of synthesized translations. Each system finds some compromise between these polar principles. Optimization each of them is achieved with the use of systems with inter-editing, as well as the use of technology "translation memory." The latter allows preserving individual transformation solutions of a particular translator and shows an interest for researches of applied linguistics and linguopersonology.

The conclusion is if not to require too much from the MT, then perhaps the dreams of a qualitative machine translation will be soon fulfilled. Afterwards, we must remember that a good translation of the text is not only creative, but also quite labor-intensive work. Even a very good translation, as a rule, needs an editorial amendment. As for the creative part, in the nearest future computer-human competition will be probably won by the artificial intellect created by a man.

REFERENCES:

1. Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. Confidence Estimation for Machine Translation.- In Proceedings of COLING: Geneva, 2004.- P. 315–321
2. Byrne Jody Technical Translation: Usability Strategies for Translating Technical Documentation. -Dordrecht: Springer, 2006
3. Hutchins W. J. Current commercial machine translation systems and computer-based translation tools: system types and their uses // International J. of Translation, 2005. Vol. 17. № 1–2.- P. 5–38.
4. Turian J.P., Shen, L., and Melamed, I.D. Evaluation of Machine Translation and its Evaluation.- In Proceedings of MT Summit IX: New Orleans, USA, 2003.- P. 23-28
5. Ulitkin, I. Computer-assisted Translation Tools: A Brief Review. // Translation Journal, Vol. 15, No. 1, January 2011.
6. Марчук Ю. Н. Компьютерная лингвистика: Учеб. Пособие: М., 2007. — 317 с.

KOMPYUTER ANALIZI VA INGLIZ TILIDAGI GAPLARNI O‘ZBEK TILIGA TARJIMA QILISH ALGORITMI

S. Muhamedova, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti, Toshkent, prof.s.muhamedova@mail.ru

Mazkur maqolada kompyuter analizi va ingliz tilidagi gaplarni o‘zbek tiliga tarjima qilish algoritmlari haqida so‘z boradi. Unda ingliz tilidan o‘zbek tiliga kompyuterda tarjima qilish dasturi vositachi tilsiz amalga oshiriladi. Shuningdek, maqolada tarjima gaplarning sintaktik analiz qilish algoritmlari va formal modellari bazasi asosida ketma-ket amalga oshirilishi ko‘rsatib berilgan.

Tayanch so‘zlar: *kompyuter analizi, tarjima qilish algoritmi, vositachi til, sintaktik analiz, formal modellar bazasi.*

КОМПЬЮТЕРНЫЙ АНАЛИЗ И АЛГОРИТМ ПЕРЕВОДА ПРЕДЛОЖЕНИЙ С АНГЛИЙСКОГО НА УЗБЕКСКИЙ

С. Мухамедова,Ташкентский государственный университет узбекского языка и литературы им. Алишера Навои, Ташкент, Узбекистан, prof.s.muhamedova@mail.ru

В статье рассматривается алгоритм компьютерного анализа перевода предложений с английского на узбекский язык. В данной программе перевода предложений с английского на узбекский язык, перевод осуществляется без языка-посредника. Статья оптсывает базовые формальные модели и алгоритмы синтаксического анализа предложений.

Ключевые слова: *компьютерный анализ, алгоритм перевода, язык-посредник, синтаксический анализ, база формальных моделей.*

COMPUTATIONAL ANALYSIS AND ALGORITHM FOR TRANSLATING SENTENCES FROM ENGLISH INTO UZBEK

S. Muhamedova, Alisher Navo‘i Tashkent State University of the Uzbek Language and Literature, Tashkent, Uzbekistan, prof.s.muhamedova@mail.ru

This article contains algorithms for computer analysis of the translation of sentences from English into Uzbek. In the program of translating sentences from English into Uzbek, the translation is carried out without an intermediary language. The article also indicates on the basis of formal models a sequence of algorithms for the syntactic analysis of sentences

***Key words**: computer analysis, translation algorithm, intermediate language, syntactic analysis, base of formal models.*

Ingliz tilidagi matnlarni o‘zbek tiliga va aksincha, o‘zbek tilidagi matnlarni ingliz tiliga o‘girishning ommaviy kompyuter metodlarini qo‘llash, tillarni kompyuter yordamida o‘qitish, bilimlarni baholash, matnlarni tahrirlash eng dolzarb muammolar hisoblanadi.

Inglizcha-o‘zbekcha va o‘zbekcha-inglizcha kompyuter tarjimasi dasturlari ham juda katta ahamiyatga ega. Ma’lumki, ingliz va o‘zbek tillari leksik-grammatik xususiyatlariga ko‘ra bir-biridan tubdan farq qiladi. Shuning uchun ingliz tilidan o‘zbekchaga va o‘zbekchadan inglizchaga kompyuter tarjimasini yaratish o‘ziga xos qiyinchliklarni keltirib chiqaradi. Bugungi kunda rus tili vositasida ingliz tilidan o‘zbek tiliga avtomatik tarjima qiluvchi dasturlarning versiyalari e'lon qilingan.

Ammo biz taklif qilayotgan ingliz tilidan o‘zbek tiliga kompyuterda tarjima qilish dasturi vositachi tilsiz amalga oshiriladi. Ta’kidlash lozimki, mazkur ish gaplarning sintaktik analiz qilish algoritmlari va formal modellari bazasi asosida amalga oshiriladi.

INGLIZ TILIDAN O‘ZBEK TILIGA KOMPYUTERDA TARJIMA QILISH ALGORITMI

Algoritm quyidagi vazifalarni hal qilish uchun mo‘ljallangan:

I. Analiz-bunda ingliz tilidagi gap quyidagi soddalashtirilgan model ramkasida sintaktik tahlil qilinadi:

1. Ushbu model faqat sodda gaplarni qamrab oladi.
2.Gapning har bir bo‘lagi bitta so‘zdan iborat bo‘ladi.
3.Gaplarda aniqlovchilar bo‘lmaydi.
4.Gaplarning standart tiplari ko‘rib chiqiladi (darak gap (ega + kesim + to‘ldiruvchi + hol), so‘roq, inkor va so‘roq-inkor gaplar).
5. Fe'lning quyidagi tuslanishli shakllarni qamrab oluvchi kesimli gaplar ko‘rib chiqiladi:

a) shaxs (I, II, III shaxs);
b) son (birlik va ko‘plik);
d) zamon (Past, Present, Future);
e) harakat tipiga ko‘ra (Simple, Continuous, Perfect, Perfect-Continuous)
f) maylga ko‘ra ((Indicative mood)
g) nisbatga ko‘ra (Active i Passive voices).

II. Tarjimada gaplar ingliz tilidan o‘zbek tiliga o‘giriladi. Algoritm quyidagi etaplardan tashkil topadi:

1) gap kiritiladi;
2) gapning har bir so‘zi I massivining elementlariga qo‘shiladi;
3) a massivining elementlari yordamida lug‘at elementlari bilan taqqoslanadi, bu lug‘atda olmosh, ko‘makchilar, ko‘makchi va modal

fe'llar, artikllar va noto‘g‘ri fe'llar ro‘yxati mavjud bo‘ladi;
4) agar so‘zlar yordamchi lug‘atda topilmasa, unda taqqoslash maxsus lug‘at yordamida davom ettiriladi;
5) topilgan so‘zlar yordamchi lug‘atga beriladi, bu erda so‘zga ushbu so‘zni va uning tarjimasini saqlovchi kod beriladi;
6) bunday so‘z lug‘atlarda mavjud bo‘lmasa, so‘z shakl yasovchi affikslardan ajratib olinadi va 5-ish bajariladi;
7) agar so‘zlar yordamchi va maxsus lug‘atlardan topilmasa, ushbu so‘zning yo‘qligi haqida ma’lumot kiritiladi;
8) gap 2 guruhga bo‘linadi: kesimgacha bo‘lgan so‘zlar ega guruhiga kiradi;
9) kesimdan boshlanib gapning oxirgacha bo‘lgan so‘zlar kesim guruhi hisoblanadi (kesim guruhga: kesim, to‘ldiruvchi, hol);
10) kesim guruhidan kesim ajratib olinadi;
11) so‘ngra to‘ldiruvchi ajratiladi;
13) gapning qolgan qismi hol hisoblanadi;
14) gapning har bir bo‘lagi shakl yasovchi qo‘shimchalarsiz tarjima qilinadi;
tarjima qilingan gap bo‘laklaridan o‘zbek tilidagi gap tuziladi, u albatta ingliz tilidagi gap konstruktsiyasiga mutanosib bo‘ladi;
15) o‘zbek tilidagi so‘zlarga ingliz tilidagi so‘zlarga mutanosib ravishda affiks va qo‘shimchalar qo‘yib chiqiladi;
16) tarjima chiqarib beriladi («tarjima» rejimida);
17) analiz chiqarib beriladi («analiz» rejimida).

Dastur ishlashini misol yordamida ko‘rsatib beramiz:

We received a letter from school.

I. Morfologik tahlil.

1**) We —** kishilik olmoshi, ko‘plik birinchi shaxs, tarjimasi — **biz**;
2) **Received** – **receive+ ed**, fe'l, tarjimasi- **qabul qilmoq**;
3) **a**-noaniq artikl;
4) **letter**-birlikdagi ot, tarjimasi- **xat**;
5) **from**-ko‘makchi, tarjimasi- **(-dan);**
6) **school**-birlikdagi ot, tarjimasi-**maktab**.

II. Gap bo‘laklarini ajratish:

We Ega	Receiv ed kesim	a letter to‘ldiruvchi	from school hol

III. Gap tahlili:

1. Ijro mayli.
2. Aniq nisbat.
3. Simple (harakatlar).

4. O‘tgan zamon.
5. Darak gap.

VI. Tarjima.

Biz maktabdan xatni qabul qilgandik.

Taklif qilinayotgan dastur inglizcha–o‘zbekcha kompyuter lug‘atini yaratishning asosi (Computer Dased Dictonary) va undan effektiv va har tomonlanma foydalanish uchun kalit hisoblanadi. Avvalo shuni ta’kidlash zarurki, keng doiradagi mutaxassislar bilan bir qatorda tillarni o‘rganish va tarjimada har kuni muammolarga duch kelayotgan har qanday insonlarga mo‘ljallangandir. Mazkur lug‘at foydalanuvchiga bir necha marotaba vaqtni tejash imkoniyatini beradi.

Demak, dasturning ishlash algoritmi o‘zida quyidagi bosqichlarni qamrab oladi:

1.1. Boshlanish.
2.2. Rejimlarni tanlash.
3.3. So‘zni kiritish va uning kodini xotiradan qidirish.
4.4. Tanlangan rejimlarning maxsus dasturlari bilan topilgan kodni qayta ishlash va talab qilinayotgan ma’lumotlarni chiqarish.
5.5. Joriy rejimda ishni davom ettirish haqida so‘rash.
6.6. Ishni yakunlash haqida so‘rash.
7.7. Tamom.

Yaratilgan dastur versiyasi 10 000 ta umum iste'moldagi inglizcha so‘zlar bazasiga asoslanadi va Turbo Pascal 7.0 dasturlash tilida ishlab chiqiladi. U Windows, Norton Commander, Far larida va MS-DOS operatsiyasi sistemasida ekpluatatsiya qilinadi.

Kelajakda dasturning Delphi ga asoslangan versiyasini ishlab chiqish unga ovoz effektlarini qo‘shishni ishlab chiqish rejalashtirilgan.

ADABIYOTLAR:

1. http://elaticsearch.docwiki.ru
2. Po‘latov A., Muhamedova S. Kompyuter lingvistikasi.-Toshkent, 2009.
3. Po‘latov A. Kompyuter lingvistikasi.-Toshkent: Akademiknashr, 2011.

A NEW APPROACH TO AUTOMATED AZERBAIJANI-ENGLISH ATRANSLITERATION

S. Mammadzada, Institute of Information Technology of ANAS, Baku, Azerbaijan, sabina_ict_az@mail.ru

The paper highlights the importance and dominance of the English language in the globalizing world. English is explored as a lingua franca. Moreover, the necessity of a new approach to Azerbaijani-English transliteration is emphasized. In this regard, new conversion table for Azerbaijani-English language pair is set out. The transliteration principles and transformation rules for automated transliteration system are defined and the accuracy of this system is provided. The possible contribution of the proposed transliteration system for machine translation system of Azerbaijani-English language pair is shown.

Key words: *transliteration system; Romanization; conversion table; machine translation.*

НОВЫЙ ПОДХОД К АВТОМАТИЗИРОВАННОЙ АЗЕРБАЙДЖАНО-АНГЛИЙСКОЙ ТРАНСЛИТЕРАЦИИ

С. Маммадзадэ, ***Институт информационных технологий НАНА, Баку, Азербайджан,*** **sabina_ict_az@mail.ru**

В статье подчеркивается важность и доминирование английского языка в глобализирующемся мире. Английский изучается как лингва франка. Кроме того, подчеркивается необходимость нового подхода к азербайджано-английской транслитерации. В связи с этим намечена новая конверсионная таблица для азербайджано-английской языковой пары. Определены принципы транслитерации и правила трансформации для автоматизированной системы транслитерации и обеспечена точность этой системы. Показан возможный вклад предложенной системы транслитерации в систему машинного перевода азербайджано-английской языковой пары.

Ключевые слова: *система транслитерации; латинизация; таблица пересчета; машинный перевод.*

1. Introduction

Textual exchanges are performed through mail, instant messaging, real-time chat protocols, asynchronous discussion forums, Web pages, and etc. all of which constitute Computer-Mediated Conversation. In the global network, this ‘conversation’ is mainly realized in English. This fact has several reasons, one of

which is the emergence of the recent term ‘English as a lingua franca’ (ELF). This is a way of regarding the communication in English among the speakers, whose first language is non-English. According to Crystal’s estimations, approximately only one out of every four users of English in the world is a native speaker of the language (Crystal, 2003). It means that the greatest part of ELF communications occur among ‘non-native’ speakers of English. English, in most cases is ‘a ‘contact language’ between persons who share neither a common native tongue nor a common (national) culture, and for whom English is the chosen foreign language of communication’ (Firth, 1996).

Most linguists indicate that English is the most broadly used language for international and intercultural communication. The report presented by Ammon on the global usage of English illustrates the current prevailing position of English. He states that the dominance of English s confirmed by the statistics, (1) English is spoken by approximately 1,5 billion people worldwide, (2) English is adopted as an official language of 62 nations, (3) about 70-80 percent of academic publications in scientific communication is written in English, (4) in most international organizations, English is the de facto official and working language, and (5) English is the most taught foreign language throughout the world (Ammon, 1992).

At present, non-English-speaking users are eager to search information on the web, particularly, researchers are very enthusiastic to explore and publish articles to English-language journals. They try to access to a broader audiences and scientific academic community. Sunol and Saturno believe that the studies conducted in languages other than English are not accessible and cited as much as those written in English (Sunol, Saturno, 2008).

The notion of transliteration is explained as a process of replacing or supplementing the scripts or words of one language with the scripts or words of another. The key point here is the exact or at least the maximum equivalence of the phonology of the source language in the target language. Transliteration helps to establish the context in which cross-cultural translation is recognized. Nida and Taber state that untranslatable words usually appear when exact equivalence of the word does not exist in target language and comparative equivalence is needed (Nida, Taber, 1969). In this case, transliteration is used.

Obviously, most web resources are Romanized for being accessible for a wider audience. Numerous Romanization tables are adopted throughout the world in this regard. Although, several conversion tables for the Azerbaijani language are adopted by international and local organizations, the proper international standard for this language has not been accepted yet (Library and Archives Canada, Ottawa, 2006). The first of the three standard adopted by ISO9 is on the transliteration of Arabic characters (ISO 233: 1984), which were used in Azerbaijan at that period (1984), into Latin characters. The second standard was adopted in 1993 as the second edition of ISO 233 to simplify Arabic

transliteration; however these scripts had already been disabled in Azerbaijan prior to 1993. The third and the foremost transliteration standard ISO 9 was adopted in 1995 for the transliteration of Cyrillic characters into Latin characters (for Slavic and non-Slavic languages).

Since the problem of transliteration may have an effect of the research activity of many students and researchers, its study and development of methods becomes significant. Many researchers have proposed and developed several transliteration methods for separate languages so far. Though a few transliteration tables for Azerbaijani language are adopted by some organizations, their modification and standardization problems still remain. The modification of the transliteration tables comprises the exclusion of diacritics from the table and replacement of some letters, the phonology of which does not appropriate for the phonology of Azerbaijani language. This study attempts to fill the gaps existing in Azerbaijani-English transliteration.

2. Azerbaijani-English transliteration challenges and their solution

The alphabet of Azerbaijani language includes 32 Latin letters, 7 (ç, ğ, ı (lower case for «I»), İ (upper case for «i»), ö, ş, ü) of which are modified and 1 letter (ə) is adopted specifically for the Latin alphabet of Azerbaijani language. In fact, these letters become very challenging in the process of transliteration, since not only their graphical description, but also their pronunciations are unknown for English-speaking audience. Representation of these letters by the diacritical signs in adopted transliteration tables, such as ġ, ž, ǰ, ̌, ï, ̂, ù, š, č, ĉ, which are unfamiliar even for Azerbaijani reader, makes this problem even more complicated. The representation of the Azerbaijani letter «ə» is another problem. This letter is adopted in few languages, such as Abkhaz, Kazakh, Tatar, and Bashkir, although pronounced differently or not pronounced at all. Although, several attempts have been made for the unification of all Turkic alphabets during the past two decades in order to ease the writing system of all Turkic languages using Latin alphabet, adoption of common Turkic alphabets has failed (Bedii Duru Altuğ, 2014).

The problem of lack of letters also exists in the transliteration from English into Azerbaijani. For example, the letter «w» does not exist in Azerbaijani alphabet and has to be replaced by letter combination «ou» or «v».

One of the problems complicating transliteration issue of Azerbaijani-English language pair is the significant difference in the number of the scripts in both alphabets. Since the number of letters in the English alphabet is less than in the Latin alphabet used for Azerbaijani language, the use of letter combinations for the transliteration is required in order to accomplish a comprehensive transliteration (Weinberg, 1974).

Azerbaijani		English	Azerbaijani	English
A		A	Q	G
B		B	L	L
C		J	m	m
Ç		ch	n	n
D		D	o	o
E		e/a	ö	o
ə		E	p	p
F		F	r	r
G		G	s	s
Ğ		gh	ş	sh
H		H	t	t
X		kh	u	u
I		I	ü	u
I		I	v	v
J		Zh	y	y
K		K	z	z

The use of letter combination is another challenging issue. Thus, letter combinations adopted for the letter with the same sound vary depending on the language. For example, the Cyrillic letter «я» can be transliterated as «ia», «ja», and «ya» depending the conversion system. The point is that the characteristics of specific phonemes in Slavic languages are not written explicitly with only one Roman letter. Therefore, several equivalence for Azerbaijani letter «y» arises, which include «i», «j», and «y».

Other problem related to the transliteration is adoption of the conversion tables for the ancient alphabets, which are not used anymore. Adoption of these transliteration schedules is significant for the libraries, where numerous literary materials written in ancient alphabets are preserved. Although they have not been used for many decades, or even centuries, the phonological and graphic features of these letters are still taken into account in the design of the new conversion schedules. This fact complicates the modification, simplification, and finally, the improvement of available transliteration standards.

Transliteration problems also arise due to some political and social factors. For example, the territory of modern Azerbaijan has been invaded by different governments several times. Each invasion has brought new culture, traditions, languages and scripts with itself. With the conquest of Arabic Khalifat and the spread of Islam in Azerbaijan gave impetus to the rise of Arabic script here and the ancient Latin script used here was replaced by the Arabic one. Whereas in the period of Soviet Union, Arabic script was replaced by the Cyrillic. In addition, in different periods, the mentioned Cyrillic alphabet has been modified for several times to achieve the correspondence with the phonological features of Azerbaijani language. After gaining independence in 1991, Azerbaijani government adopted modified Latin script for Azerbaijani language (Law of the Republic of Azerbaijan on the renewal of the Azerbaijani alphabet with Latin graphic, 1991). The key point is that, in each period, numerous literary manuscripts, political documents and materials have been generated in each of these scripts, which increases the importance of transliteration problem for the next generation. Each of those transliterated materials becomes a valuable resource for the preservation of the national identity. However, due to the numerous changes made to the scripts, various transliteration versions have been generated and spread on the web. Solution of this problem requires the adoption of a unique transliteration standard for Azerbaijani language, including the adaptation of social and psychological and moral thinking of Azerbaijani people.

Table 1. Modified ISO Romanization table for Azerbaijani language used for the system

As mentioned above, the spread of various transliteration forms leads to confusion. This problem exists almost in all languages. For example, Azerbaijani last name «Quliyev», which has numerous transliterations as Kuliyev, Quliyev, Guliyev, Kuliev, Quliev, and Guliev. The reason for these variations is due to the long-term impact of Russian language of Azerbaijani language. Thus, the letter «Q», equivalence of which in accordance with the phonological features of this letter in Azerbaijani language is «G», was incorrectly transliterated as «K» into Russian language. Since the letter combination «ye» sounds same as the Russian letter «e», it was transliterated into «e», omitting the letter «y». Consequently, the identity of numerous personal names and place names is violated.

At present, transliteration issue is closely related to the multilingual nature of the global web, in which each of existing and endangered languages tries to be properly represented and not to be pressed by others, and to preserve the national identity. Many organizations dealing with the transliteration tables do not own transliteration systems, which can process different scripts. One of the reasons of this fact is insufficient funds, whereas another reason is related to the unwillingness or less concern to meet the multilingual needs of community (Simsova, 1988). Since most of organizations are restricted to the Roman characters and do not ensure transliteration tables for non-Roman scripts, users

cannot find or access the needed resources, materials and documents in their original scripts. Consequently, Romanization becomes urgent being provided with the automated cross-language transliteration system, which comprehensively represent the graphical and phonological features of certain language.

For comprehensive solution of transliteration problems, in this study, Romanization table for Azerbaijani language ISO is modified and provided in the Table 1. The modifications are made by the experts of certain fields of sciences in accordance with the grammatical, graphical and phonological features of Azerbaijani language. Based on this conversion table, an automated transliteration system for Azerbaijani language with other most commonly used languages by the Azerbaijanis, namely Russian, English, French, German and Farsi, is developed. Transliteration system is improved to achieve maximum transliteration accuracy based on feedbacks and proposals submitted to the system by the users.

Simple replacement principle is used for the system transliteration. The advantage of the system is its independence from any handmade corpus as the system is intended not only for the transliteration of proper names and terms but for all words and texts. For perfection of the automated transliteration the feedbacks and errors are regularly controlled, checked and corrected by the experts of the field.

CONCLUSION

The problem of transliteration of Azerbaijani language with other languages, particularly with English becomes topical, taking into account the number of Azerbaijanis living throughout the globe accounting for 30-35 million people (Sela, Avraham, 2002). Given the dominance and the status of international language of English, transliteration from Azerbaijani language into English and vice versa becomes an essential element of their communication with each other via the virtual environment.

This paper explored the growing dominance and status of English as an «international language» and «lingua franca». One of the key elements of communication on the web, that is transliteration, its importance and challenges were highlighted. The requirements for the users to transliterate necessary resources and information written in the modified Latin scripts were explained. Since the web resources include different transliterations of the same words, the factors affecting this situation were identified. This paper focused on and offered some strategies that could contribute to eliminate transliteration problems, preserve national identity and improve information retrieval.

REFERENCES:

1. Crystal, D. (2003). English as a Global Language (Second edition). Cambridge: Cambridge University Press.
2. Firth, A. (1996). The discursive accomplishment of normality. On «lingua franca» English and conversation analysis. Journal of Pragmatics 26: 237–59.
3. Ammon U.,, (1992), Gengo-to Sono Chii. Trans. by Y.Hieda and H.Yamashita Tokyo; Sangensha
4. Sunol, R., & Saturno, P. (2008). Challenge to overcome language barriers in scientific journals: Announcing a new initiative from the ISQua journal [Editorial]. International Journal for Quality in Health Care, 20(1), pp. 1–2.
5. Nida, E. A., & Taber, C. R. (1969). The theory and practice of translation. Leiden: EJ Brill
6. Library and Archives Canada, Ottawa, (2006) https://www.bac-lac.gc.ca/eng/services/cataloguing-metadata/Documents/040006-14-e.pdf
7. Bedii Duru Altuğ, (2014) The 1991 International Contemporary Turkic Alphabets Symposium and its Contributions to the Turkic Alphabet Reform, University of Washington, pp.8-11
8. B. Weinberg, (1974) "Transliteration in Documentation," Journal of Documentation 30: pp. 18-31.
9. Law of the Republic of Azerbaijan on the renewal of the Azerbaijani alphabet with Latin graphic (December, 25, 1991).
10. Sylva Simsova, (1988), "Coping with Foreign and Nonstandard Character in Libraries," in John Eyre, ed., Small Computers in Libraries, London: Meckler, pp.13-15
11. Sela, Avraham (2002). The Continuum Political Encyclopedia of the Middle East. Continuum. p.197. ISBN0-8264-1413-3.

TRANSLATION OF MULTIPLE SENSES IN UNRESTRICTED TEXTS

Y.Polat, S. Bacak, A. Zakirov, Ala-Too International University, Bishkek, Kirgizstan, yahya.polat@iaau.edu.kg

This paper addresses the problem of how to identify the primary and secondary senses in translating the various senses. To discriminate senses, a translator should consider the characteristic of words that a single lexical item may have several meanings other than that which most readily comes to mind. These meanings are often called secondary meanings, or secondary senses. Our discussion will include how the the meaning is suggested by the word when it is used alone, when the word is said in isolation. It is the meaning learned early in life and is likely to have reference to a physical situation. Here we will describe how the same word may have a different meaning when used in context with other words. We will also discuss ambiguity caused by senses in translation.

Key words: *Primary sense, secondary sense, multiple senses, ambiguity.*

ПЕРЕВОД МНОГОЗНАЧНОСТИ В НЕОГРАНИЧЕННЫХ ТЕКСТАХ

Я. Полат, С. Бачак, А. Закиров, Международный университет Ала-Тоо, Бишкек, Кыргызстан, yahya.polat@iaau.edu.kg

В данной статье подчеркивается важность перевода многозначности в неограниченных текстах. Как известно, перевод многозначности в неограниченных текстах сталкивается со многими особенностями при переводе на другой язык во время декодирования устной речи и трансформации ее в письменную речь. Основная цель данной статьи заключается в рассмотрении ряд проблем перевода многозначности в неограниченных текстах.

Ключевые слова: *первое значение, второе значение, многозначность, неоднозначность.*

INTRODUCTION

Primary sense is the core, basic, literal meaning of a lexeme. A primary sense is generally the first meaning that comes to mind for most people when a lexeme is uttered alone. Usually it refers to an actual physical thing, an action, or a characteristic of a referent [1]. The primary sense is the meaning suggested by the word when it is used alone. It is the first meaning or usage which a word will

suggest to most people when the word is said in isolation. It is the meaning learned early in life and is likely to have reference to a physical situation. But the same word may have a different meaning when used in context with other words.

A secondary sense is a meaning that is more abstract than a primary sense of a lexeme but still shares some of its semantic components. Because it has a different range of reference, its usage contexts and collocates are different from those of a primary sense. For example the word ‘okşamak’ in Turkish has a primary sense meaning 'to caress, to fondle.' As in the example «adam çocuğun başını okşadı», «The man caressed the child’s head». However, okşamak can also mean; ‘to resemble to someone’ as a secondary meaning. «Fatma teyzesine okşuyor», means «Fatma looks like her aunt».

The "unpacking" of the concepts or meaning components contained in a word all deal with the fact that the same meaning may occur as part of the meaning of various words [2,112]. In order to define the problem more clearly we can look at the ways of unpacking the word:

(a) By looking at Lexical items from the point of view of the meaning components of which a given word is composed.

(b) By contrasting one lexical item with another in a system.

(c) Pairs of words which have some meaning in common may be contrasted; whole semantic sets may be contrasted.

(d) Taxonomic studies, componential analyses, the study of antonyms and synonyms.

These ways given above are all about one sense of a given word, the primary meaning. However, most words have more than one sense.

For example the word run in isolation will mean something like move rapidly by moving the legs rapidly. But if the same word is used in the context of river as in the river runs, run has nothing to do with legs or rapidity, although the idea of motion is still there. Run in the context of river means to flow. Secondary senses are dependent on the context in which a word is used [2,112].

METHODOLGY

I collected the data and analyzed them descriptively by using both qualitative and quantitative methods. The analysis of this study took several steps. The data were numbered into a comparative chart, classified into a specific comparative chart to be the related data to the research subject, then the data were analyzed according to the theory used in this study.

SECONDARY MEANING OR SECONDARY SENSE

A speaker of Turkish (Turkey) will tell you that «yemek» means to eat. This is the primary meaning. But a speaker of Turkish will also use this same word in phrases as shown below [3, 32]. Examples:

Ayvayı yemek: (To eat quince) (Be screwed).

Başının etini yemek: (To eat smb’s skin of the head) (nag at smb)

Damga yemek: (Be branded, to be sealed). (to blacken smb's name)
Feleğin sillesini yemek: (To be hit by heavens). (Come down in the world)
Halt yemek: (Make a great blunder). (To do sth improper)
Hazırdan yemek: (Spend the money you saved for you do not work)
İçi içini yemek: (To be very anxious for sth bad will happen) (Eat one's heart out)
Kaymağını yemek: (To skim) (To get benefit from a good position)
Nane yemek: (Make a blunder) (Do sth stupid)
Papara yemek: (To be in the dog house) (To be told of, be reprimanded)
Başını yemek: (Cause the death of) (Get smb into trouble)
Bıçak yemek: (Be stabbed)
(bir iş birinin) vaktini almak (yemek): (Steal smn's time)
Birbirini yemek: (Go at it hammer and tongs) (To eat each other)
(birini) çiğ çiğ yemek: (eat smn alive) (be violently angry at)

Translating the primary sense of a lexical item is usually much easier than a secondary sense. This is because the receptor language will often have a lexical equivalent for the primary meaning which very nearly matches the meaning of the lexical item in the source language. However, the secondary senses of those same two words will probably not match [2,113]: A native speaker knows immediately by the other words which occur in the phrase or sentence which sense of the word is being signaled. Learners of a second language often have a great deal of trouble to use a word in its many secondary senses.

Turkish	**English**
Asker kaleye yürüdü:	Soldiers marched to the castle
Dedemiz Hakka yürüdü:	Our grandfather has passed away
Dallara su yürüdü:	Water moving up to the branches

Any word used in a non-primary sense will probably not be translated by the word in the receptor language which is equivalent to its primary sense, but by a different word. For example, the primary sense of key would be translated into Turkish with «anahtar.» But notice the following list which shows how they differ in translating secondary senses:

English to Turkish
key — **anahtar** (of a lock)
key — **şifre** (of a code)
key — **tuş** (of a typewriter)
Turkish to English
anahtar – **key**
anahtar – **switch**
anahtar — **clue, clef, cipher, cotter, cock, spanner, interrupter, wrench, toggle**

ANALYZING SENSES OF WORDS

The process for discovering the various senses of words is rather complicated but can be very crucial for making dictionaries, learning a second language, and may also be helpful to the translator when no dictionaries are available which give an adequate description of the senses of words in the language [4]. A translator who is truly bilingual in the source and receptor languages will usually recognize a non-primary sense. Nevertheless, there is always the possibility that a literal translation of a word may be used in a secondary sense. This literal translation sets up a strange collocation and wrong meaning [2,113].

Step 1. Collecting data.

One must first collect as many examples of the use of the word as possible. If a person knows the language he can simply think of all the possible combinations with other words. If not, he will need to find the word in as many texts as possible. A concordance done on the computer will greatly speed up the search, learning a language, or hoping to make a dictionary, will want to begin early in his research to collect data on each word of the language, building up more words and more examples of their co-occurrence with other words. The goal is to list as many collocate as possible. For our purposes, we shall now assume that we have found the following [4].

Cam kırdı	Broke the window	Fındık (ceviz) kırdı	mess around women
Ayağını kırdı	Broke his leg	Gurur kırdı	humiliated
Cesaretimi kırdı	Discouraged me	Onur kırdı	insulted
Kalbini kırdı	Broke heart	Hatırını kırdı	offended, worried
Fiyat kırdı	Made discount		
Tavlada pul kırdı	Hit a checker	Hevesini kırdı	dissuade,
Umudunu kırdı	Dashed my hopes	Aşkını, şevkini kırdı	dishearten
		İnadını kırdı	overcome his stubbornness (will)
Direksiyonu kırdı	Turn the wheel hard		
		Kabuğunu kırdı	broke the shell
Kemiklerini kırdı	Broke his bones	Kesek kırdı	harrow
Boynunu kırdı	Broke his neck	Kibrini kırdı	abase
Dersi kırdı	played truant	Kirişi kırdı	got away
Direncini kırdı	broke his resistance	Kod kırdı	broke the code
		Nefsini kırdı	mortify the flesh
Dümen kırdı	veered	Not kırdı	took points of a student
Rekor kırdı	broke the record		
		Pot kırdı	dropped a br

Step 2. Sort the collocates into generic classes.

Each grammatical form should be analyzed separately. In this example, we have used only intransitive verb forms. If the noun run occurred, this noun form would need to be separated and analyzed separately. One begins by making best guesses, refining the analysis as he goes.

(1) Human body: Leg, bone, neck
(2) Human senses: Courage, heart, hope, resistance, honor, pride
(3) Objects: Window,
(4) Run away: lesson (play truant), got away
(5) Change direction (car, ship): veer, wheel
(6) Decrease: note, price
(7) Having affair: mess with woman
(8) Mistake: drop a brick
(9) Game: Hit a checker
(10) Change sth: broke the shell
(11) Achievement: broke the record

Step 3. Regroup the contexts according to the collocates which belong to the same generic classes as follows

3.1. Human body

(1) Ayağını kırdı: Broke his leg
(2) Kemiklerini kırdı: Broke his bones
(3) Boynunu kırdı: Broke his neck

3.2. Human senses:

(1) Cesaretimi kırdı: Discouraged me
(2) Kalbini kırdı: Broke heart
(3) Umudunu kırdı: Dashed my hopes
(4) Direncini kırdı: Broke his resistance
(5) Gurur kırdı: Humiliated
(6) Onur kırdı: Insulted
(7) Hatırını kırdı: Offended, worried
(8) Hevesini kırdı: Dissuade,
(9) Aşkını, şevkini kırdı: Dishearten
(10) İnadını kırdı: Overcome his stubbornness (will)
(11) Kabuğunu kırdı: Broke the shell
(12) Kibrini kırdı: Abase
(13) Nefsini kırdı: Mortify the flesh

3.3. Objects

(1) Cam kırdı: Broke the window
(2) Kesek kırdı: Harrow

3.4. Running away

(1) Dersi kırdı: Played truant
(2) Kirişi kırdı: Got away

3.5. Change direction (car, ship)

(1) Dümen kırdı: Veered
(2) Direksiyonu kırdı: Turn the wheel hard

3.6. Earn or punish by decrease

(3) Fiyat kırdı: Made discount
(4) Not kırdı: Took points of a student

3.7. Having affair

(5) Fındık (ceviz) kırdı: Mess around women

3.8. Mistake: drop a brick

(6) Pot kırdı: Dropped a brick

3.9. Game:

(7) Tavlada pul kırdı: Hit a checker

3.10. Penetrate a secret

(8) Kod kırdı: Broke the code

3.11. Achievement

(9) Rekor kırdı: Broke the recor

Step 4. List and label the senses of the words.

Once the data is reorganized by the generic classes of the collocates, it is much easier to see the senses of the word. For animate beings with legs, the meaning seems to be to move oneself from one place to another rapidly; for liquids, simply to flow, for vines, the meaning is to grow, etc.

Sense 1: Changing the form of body in an unwanted way.
Sense 2: Changing the human senses
Sense 3: Changing the form of objects.
Sense 4: Running away from responsibility
Sense 5: Change direction
Sense 6: Changing value
Sense 7: Changing the value of heart
Sense 8: Mistake
Sense 9: Game
Sense 10: Penetrate a secret
Sense 11: Achievement

Translating the Various Senses

If the above analysis were of the receptor language word, that is, if one were translating into English, the analysis would point up the necessity of including, in the context of run, a collocate from the generic class mentioned in order to insure the correct meaning. When the meaning is signaled by the context in which the word occurs, it is very important that the context be built into the translation.

The word «uçmak» occurs in the following contexts, each signaling a different sense of the Turkish word. It is possible to restate the meaning in Turkish.

Kuş uçtu: A bird flew
Uçak uçtu: Plane took off
Gaz, buhar uçtu: Gas, steam evaporated
Rengi, benzi uçtu: He grew pale
Çatı uçtu: The roof structure was uplifted by the hurricane
Toprak, evin üstüne uçtu: Soil eroded over the house
Patlamadan dolayı bina havaya uçtu: Building exploded
Saçları havada uçuyor: Her hair was flown in air

Araba çok hızlı gidiyor, uçuyor: The car was very fast
Yarın istanbul'a uçuyorum: I am flying to İstanbul tomorrow
Yok oldu sanki havaya uçtu: Lost, disappeared suddenly
Sevinçten havalara uçtu: He was very very happy
Uyuşturucu almış uçuyor: He tripped out
Bizim kitaplar uçmuş: Our books were stolen
Cennete uçtu: Flew to haven, died

The idea of "flying in a presentable form" is common to all the senses. The common thread of meaning shows that we are dealing with a single word rather than with two or more separate words [4, 97], but each sense will result in a different form for the translation.

A secondary sense will almost always need to be translated by a different word than the word which denotes the primary sense. In English there are many synonyms of the word powerful.» They include strong, muscular, muscly, sturdy, strapping, robust, brawny, burly, heavily built, athletic, manly, well built, solid; and others. All belong to a common semantic set and can be contrasted and components of meaning analyzed as presented in the previous chapter. The nuclear component of each would be POWER. «Tiger» has the contrastive component «animal, wild»; «King» has the contrastive component of being a human, authority, etc. That is, each of these contrasts with the others in the semantic set. But in addition, each of these words has a primary sense and a number of secondary senses. Some of them are being used in a secondary sense when they are included as part of the semantic set, POWERFUL BEING. For example, "king» has the primary meaning of head of a country. However, it also has a secondary meaning of «the most important chess piece.» A word may be a member of various semantic sets. In some, it will be used in its primary sense and in others in one of its secondary senses. This, of course, adds to the complications of translation.

In the display which diagram the senses of the word «powerful», notice that ten senses have been identified. (The kind of analysis which leads to this type of charting is described in [5, 99-113].

POWERFUL

Animate								Inanimate
God	Man				Animal			Nature
Abso-lute	Natural	Wealth	group	Authority	Wild		domes-tic	Water, rain flood
One	children	man	army	police	land	Water	ram	wind hurricane
1. Omni potent	man	woman	media	King	Tiger	Crocodile	rooster	earthquake

2. Sturdy	3. Wealthy	4. Influential	5 Authoritative	6. Strong	7. powerful	8. Fast	9. Devastating

In the analysis of the English word «powerful,» the senses are numbered at the bottom with the primary sense «strong» as number six. In the discussion of secondary senses above, we showed how the sense is signaled by the collocates that go with the word. However, it may not always be a specific word that signals the meaning but the presence of some signal of the components of meaning within the word when used in that sense. For example, to signal the sense of «woman,» rather than «shark» for «powerful» something in the context must signal «human» rather than «wild water mammal» since «wild» (powerful) is not the primary meaning of powerful.

«Powerful» has at least ten senses. But the meaning will be signaled only if the translation into English has built into the context the semantic components that will trigger the meaning. If not, the wrong meaning may result even when the right word is used. For example, if we use the collocate «woman» for the context, the meaning would still be ambiguous. It could refer to her body power, or a wealth of her. If an English said, "they have a powerful media," it would immediately be understood that their media has a lot of influence on people of that country. If someone said, "There is a powerful leader in our country," it would immediately signal a man, since it must be animate. The collocate powerful leader signaled this. The choice of meaning is signaled by including in the context some other lexical item which will activate the semantic components indicated at the nodes of the chart. The king had a powerful time on his throne is understood to be a rule over people because the throne indicates this sense. John saw a powerful bite in the body would mean tracks because of the collocate body.

The two main rules about secondary senses are

1) the secondary senses of the source language can probably not be translated literally but will need to be understood in order to find a good equivalent, and

2) the secondary sense of words in the receptor language will only mean what they are intended to mean if the context includes collocates which will signal the sense desired.

Ambiguity Caused by Senses not Clearly Signaled

Something is ambiguous when it can be understood in two or more possible senses or ways. If the ambiguity is in a single word it is called lexical ambiguity. In a sentence or clause, structural ambiguity [6].

It is important to know the meaning components of the primary sense. For example, in the Chuj language of Guatemala, the word say turned out to be a problem for the translator. The word say was used in the sentence, «The people said, «This man is God.»« In the story where this was used, the man was not

God. The people said it, but it was not true. However, what the translator did not know was that the word say in its primary sense includes the component of the truth. The word say in Chuj means to say the truth; that is, the unmarked meaning. In order to indicate that what they said was not true, say must be marked. So it had to be translated «The people said falsely, «He is a God,» to avoid wrong meaning. [4, 115]

It should also be noted that lack of context will lead to ambiguity in many cases example, the sentence «study like your brother, do not be lazy!» is ambiguous. It could mean that his brother is hard working or lazy. The ambiguity comes because of the two senses, and lack of context to make it unambiguous. It would be possible to simply say «study like your brother, do not be lazy» or «study, like your brother do not be lazy!» Here, change of comma, changes the meaning of the sentence. No equivalent lexical items will have the same senses from language to language. Even primary meanings that look the same at first may have additional components that can distort the meaning if used without care. One of the most important things in translation is to be sure that the context is sufficient to mark the meaning desired. Ambiguities often arise when the translator knows only one or two senses of a word and does not know the context needed to signal the correct meaning. Notice the three Turkish sentences below:

1. Kol yedi (he ate meat)
2. Kol gerdi (he protected)
3. Kol bastı (soldiers swooped, busted, raided)

The first means that he ate the meat of the front arm, the second that he protected someone and the third that police or soldiers raided a house or some place. All of them use the word kol which has the primary sense of arm. This is the unmarked meaning which all native speakers would give as the meaning of kol.

When Zahir Faryabi a Persian poet was praising the king Kılıcharslan, he said «when contemplating to kiss his feet, they had put the seven heaven under his feet like a stool. But Kılıchaslan's one foot was shorter. Zahir's enemies told him that «the poet meant that you were lame, limping. So they made him butcher the poet [7, 14].

Conclusion

Above we have talked about the problem of how to identify the primary and secondary senses in translating the various senses. We have considered the characteristic of words that a single lexical item may have several meanings other than that which most readily comes to mind. We have seen the examples of secondary meanings, or secondary senses.

We have discussed how the the meaning is suggested by the word when it is used alone, when the word is said in isolation. How meaning learned early in life and is likely to have reference to a physical situation. We have described how

the same word may have a different meaning when used in context with other words.We have discussed ambiguity caused by senses in translation.

REFERENCES:

1. http://www.glossary.sil.org/term/secondary-sense
2. Larson, M, L. «Meaning Based Translation» University Press of America. New York, 2012. — 589 p.
3. Barnwell K. «An analysis of strategies used in translating the short story.» High Wycombe: Summer Institute of Linguistics. 1980.
4. Beekman, J, and Callow J. «Translating the Word of God.» Zondervan Pub. House, Jun 1, 1974 — Religion — 399 p.
5. Eugene A. N. «Toward a Science of Translating» Leiden, E, J Brill Netherlands, (1964) 331 p.
6. Quiroga-Clare, C. «Language Ambiguity: A Curse and a Blessing» Translation Journal, Literary Translations. (2003) http://translationjournal.net/journal/23ambiguity.htm
7. Shirazy S. «Bostan ve Gulistan» İstanbul, Cagaloglu, 1984. 473 p.

HOMONOMY IN MACHINE TRANSLATION

U. Akhmadova, D. Isrofilov, M. Amirkulov, TSUULL
Tashkent, Uzbekistan, komp_ling@mail.ru

The article is about homonomy in machine translation, some researches in the sphere and applied proposals for special aim. Furthermore, it represents implementation of resolving problems of homonymy for mother tongue and given brief suggestion how to apply for machine translation in perspective.

***Key words**: lexical homonomy, machine translation, problems, proposal.*

ОМОНИМИЯ В МАШИННОМ ПЕРЕВОДЕ

У. Ахмадова, Д. Исрофилов, М. Амиркулов, Ташкентский государственный университет узбекского языка и литературы им. Алишера Навои, Ташкент, Узбекистан, komp_ling@mail.ru

Статья посвящена вопросам омонимии в машинном переводе, отдельным исследованиям в этой области и прикладным программам. Также в статье представлена реализация решения проблем омонимии для родного языка и дается краткое описание, как использовать в машинном переводе.

***Ключевые слова**: лексический омоним, машинный перевод, проблемы, рекомендация, предложение.*

MASHINA TARJIMA TIZIMIDA OMONIMIYA

Axmadova O‘., Isrofilov D., Amirkulov M.,
Alisher Navoiy nomidagi TDO‘AU, Toshkent, O‘zbekiston, komp_ling@mail.ru

Ushbu maqolada mashina tarjima tizimida omonimiya, u bilan bog‘liq tadqiqotlar, omonimiya muammosining yechimlari uchun amaliy tavsiyalar va takliflar hamda ularni ona tilimizga tadbiq etish haqida qisqacha to‘xtalib o‘tiladi.

***Tayanch so‘zlar:** leksik omonimiya, mashina tarjimasi, muammolar, tavsiyalar, takliflar.*

As a result of long historical and evolutionary processes, the need for human to utilize science and technology has grown. At the same time, the volume of human receiving or transmitting information resources has risen considerably. In some point it is expected to impact on the natural language. The use of a new

name for each conception has led to a sharp increase in the number of lexical units in the language. As a result some lexical items are used to express the meaning of new or existed concepts. Thus almost every language has appeared a phenomenon of homonomy

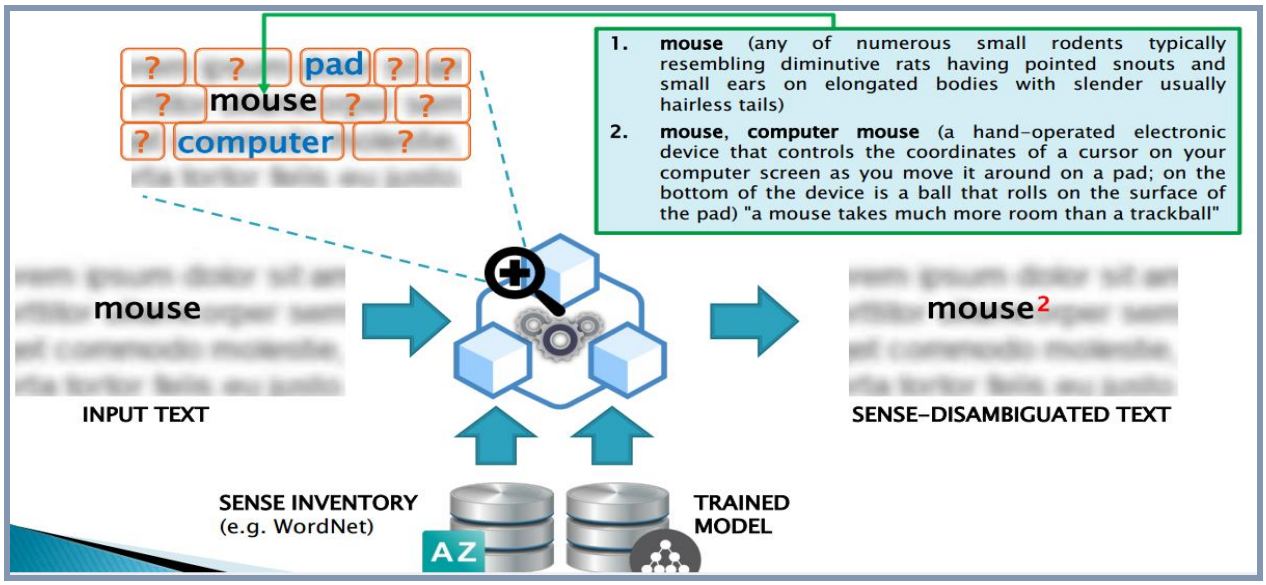
It is time to talk about the machine translation has been one of the latest achievements of human intelligence. As well as it becomes one of the fastest growing spheres in the recent decades. The key to this development is the vital demand to translate economic, social, military, and political texts in fast and quality manner. However, it is quite challenging process to obtain brillant translation product. There are some problems to find particular solution. Apparently, one of them is homonomy. How to determine appropriate meaning.

The term of homonomy is derived from Greek (homos 'similar' and onoma 'name'). According to LONGMAN dictionary it is defined as a word that is spelled the same and sounds the same as another, but is different in meaning or origin. For example, the noun 'bear' and the verb 'bear' are homonyms. Indeed, homonomy can be one or more part of speech at a time. For example: uzbek words **qirq** expresses one more meaning. If it is considered noun it means the name of funeral ceremony in uzbek culture, but it is verb, it means **to cut off smth/ to trim/ to shorten** as well it also means the number **forty** (40). To extend should we draw attention to phraseological homonomy in uzbek language to asses homonomy phrase. For instance: in uzbek if «U xonani **boshiga ko‘tardi»** phraseological homonomy is translated from uzbek into english by GOOGLE translation system, output will be «He raised **the room upside down».** But this phraseological item also contains following senses. 1. **value, estimate-** e’zozlamoq, qadrlamoq; 2. **make a noise, make a bustle, noise loudly —** shovqin qilmoq, to‘polon qilmoq, janjallashib baqir-chaqir qilmoq; 3. As a set expression – **raise to head** boshiga ko‘tarmoq (boshiga joylashtirmoq, boshining tepa qismiga qo‘ymoq. Obviously, in this sample phraseological item is used second meaning in source language that is make a noise, make a bustle, noise loudly. But GOOGLE translation system analyses it as a set expression.

The people can consciously use appropriate meaning of the homonomies in utterance by understooding common contextual meaning. However, the main point in this article is that the correct translation of the homonomy items in the machine translation. Because giving another one meaning in the translation process may have a negative impact on the meaning of the whole context and may change the content of the text.

Many scientistc and researchers in machine translation sphere put forward to some ideas and opinion. One of them is Robert Krovetz. He offered this idea to tackle the problem: « One response to this problem is to use phrases to reduce homonomy and ambiguity. It is not always possible, however, to provide phrases in which the word occurs only with the desired sense.» [19-1] Another one researcher Jamie Faulkner mentioned the following idea in his article named

"The Dangers of Polysemes and Homonyms,": "How can such threats threaten? Good question. This danger is reflected in the translation of the language. If you translate your information to anyone, you probably will not experience this problem, but if you use the machine translation, errors will be noticeable. When conversion occurs, car translators allow different errors, because the language variants are not the same. It is difficult to grasp the meaning of a verb (homonym) or a noun for a machine". And Prof. Jesús Vilares from Coruna university puts forward to **Lesk algorithm.** According to this method the translation system choose the sense whose dictionary gloss or definition shares the most words with the target word's neighbourhood. He explanes his idea a sample of word «mouse». Below we attach the process of extraction homonomy from context.



Below we attach some examples of homonomy in English to explain.

There is only a little beer left. = Bu yerda ozgina piva qolgan. I was only to pleased to leave that place. = Bu yerni tark etishdan yagona mamnun odam menman.

If only you had been, you could given me a chance to run business. = Qaniydi, sen bo'lganingda edi, menga biznesimni boshlash uchun imkon bergan bo'larding.

In these examples, the word only has three different meanings. If we translate the second sentence with the meaning of the first sentence, it is translated as: "I'm a little satisfied with leaving the place," and the meaning of this totally changes Not only does this translate into mechanical translation, but it can also lead to serious mistake in human translation. In our next example, is one of english words, board. It can express these meaning.

1. Long standart wood used for floor and other building materials.
2. Internal part of air vessels, boats and undergrounds

3. food ordered when you stay in place

4. management.

We use these meanings in the context;

1. Loose boards creaked as I walked on them – Ustida yurganimda bo‘shab qolgan pol g‘ichirladi.

2. They would not be able to board without a ticket – Ular chiptasiz bortga chiqa olishmaydi.

3. I have to pay for room and board – Men xona va ovqat uchun to‘lashimga to‘g‘ri keladi.

4. He sits on the board of directors – U direktorlar kengashida o‘tiradi.

In this circumstance classifying homonomy in the database of machine translation system may be solution. If we approach to the issue in terms of uzbek philology, we should classify according to part of speech, use of style, and range of sphere.

In our opinion, the following recommendations may also be partly a solution to the homonomy phenomenon of the machine translation.

1. Create and add to the database the set-combinations that can be combined with each other according to morphological features of the words and part of speech.
2. Create homonomy formalized forms of speech with the level of syntactic expression of words in the lexical units.
3. Huge corpora or the ready collections of homonomy must be inputted in database.
4. Use of specific electronic dictionary in the translation process, taking into account the general context of the context and coverage area.
5. Provide all meanings of homonomy words in a translation simultaneously. (Post editor selects right meaning depending on context)

If we approach to solve the homonomy problem exactly, it will be better to form some bilingual or multilingual dictionaries. They may be english-uzbek, uzbek-english lexical homonomy dictionary, english-uzbek or uzbek-english phrasal homonomy dictionary or english-uzbek or uzbek-english morphological homonomy dictionary and etc.

1. Taking into account that it is possible to translate audio texts in modern machine translation systems, the correct pronunciation of the Uzbek-English and English-Uzbek dictionary is essential one to input the database..

2. In both languages, a collection of different style texts which include of also homonomy collected in the database.

Some machine translation systems have accomplished a number of successful results and translate more satisfactory translations. However, this issue has not been fully solved in some of the English translation programs. It is important to note that in order to develop concrete and effective solutions for the

translation of the machine and to use in practice them into our native language is one of the most important tasks in uzbek computational linguistics.

REFERENCES:

1. citeseerx.ist.psu.edu/viewdoc/ Robert Krovetz «Homonomy and polysemy in information retrievel»
2. www.aclweb.org/anthology/C86 Karole Fabricz «Particle Homonomy and machine translation» maqolasi. JATE Universitet of Szeged.
3. kb.psu.ac.th/psukb/bitstream/2010 Jitsuda Laongphol. «Developing ability to translate homonomy and homographs via training in part of speech identification and dictionary use.»
4. https://www.grin.com/document/40316 Sasha Andreyeva. «Lexical Functions and Homonymy» Moscow State University
5. https://en.wikipedia.org/wiki/ List_of_true_homonyms.
6. examples.yourdictionary.com › ... › Examples of Homonyms
7. «Word sense disambiguation.» ppt. Prof. Jesús Vilares from Coruna university.

СЕКЦИЯ 3.
КОРПУСНАЯ ЛИНГВИСТИКА

ГРАММАТИЧЕСКАЯ РАЗМЕТКА КРЫМСКОТАТАРСКОГО ЭЛЕКТРОННОГО КОРПУСА (СУЩЕСТВИТЕЛЬНОЕ, ГЛАГОЛ): СРАВНЕНИЕ С РАЗМЕТКОЙ ЭЛЕКТРОННОГО КОРПУСА ТУРЕЦКОГО ЯЗЫКА

Л. Кубединова, Крымский федеральный университет им. В. И. Вернадского, Симферополь, Россия, kubedinova@gmail.com

Е. Адалы, Стамбульский технический университет, Стамбул, Турция, adali@itu.edu.tr

Морфологическая разметка является основным типом разметки в текстах, так как она рассматривается как основа для дальнейших этапов анализа – синтаксического и семантического. Создание морфологического анализатора является одной из первоочередных задач для крымскотатарского электронного корпуса. Так как исследования в области корпусной лингвистики для крымскотатарского языка являются довольно новыми по сравнению с такими тюркскими языками как татарский, турецкий, якутский, казахский, башкирский и другие, то считается приемлемым изначально обратить внимание на унификацию систем грамматической разметки в корпусах тюркских языков. В статье описываются общие и отличительные черты турецкого и крымскотатарского языков на морфологическом уровне. Предметом исследования послужили категории единственного и множественного числа, категория принадлежности, а также грамматические категории глагола в обоих исследуемых языках.

Ключевые слова: *крымскотатарский язык, турецкий язык, морфологическая разметка, тэг, существительное, глагол.*

GRAMMATICAL TAGGING OF THE CRIMEAN TATAR ELECTRONIC CORPUS (NOUN, VERB): THE COMPARISON WITH THE MORPHOLOGICAL TAGGING OF THE ELECTRONIC CORPUS OF THE TURKISH LANGUAGE

L.Kubedinova, V.I. Vernadsky Crimean Federal University
Simferopol, Russia, kubedinova@gmail.com

E. Adali, Technical University of Istanbul,
Istanbul, Turkey, adali@itu.edu.tr

Morphological tagging is the main type of markup in texts, as it is considered as the basis for the further stages of the analysis – syntactic and semantic. Creating a morphological analyzer is one of the priorities for the Crimean Tatar electronic corpus. Since the researches in the field of corpus linguistics for the Crimean Tatar language is quite new compared to such Turkic languages as Tatar, Turkish, Yakut, Kazakh, Bashkir and others, it is considered acceptable to initially pay attention to the unification of grammar annotation systems in the Turkic languages. The article describes the common and distinctive features of the Turkish and Crimean Tatar languages at the morphological level. The subject of the study was the singular and plural categories, the category of belonging, and the grammatical categories of the verb in both studied languages.

Key words: *the Crimean Tatar language, the Turkish language, morphological annotation, tag, noun, verb.*

Развитие тюркских языков и их взаимодействие является одной из самых важных задач тюркского мира. Проблемы разработки унифицированной морфологической разметки текстов на тюркских языках для использования в корпусах и других системах автоматической обработки текста обсуждались неоднократно на семинарах UniTurk в рамках ежегодной Международной конференции по компьютерной обработки тюркских языков Turklang, которая имеет довольно широкую географию проведения конференции и ее участников. Подобная унифицированная система разметки могла бы также служить в качестве универсального средства глоссирования текстовых примеров (например, в международных публикациях).

В данной статье мы рассмотрим особенности грамматической разметки существительного и глагола для электронного корпуса крымскотатарского языка в сравнении с имеющейся разметкой для корпуса турецкого языка.

В исторических условиях формирования тюркского населения Крымского полуострова крымскотатарский язык сложился как язык

типологически неоднородный. Разделяется на степной, средний и южный диалекты, относящиеся соответственно к кыпчакско-ногайскому, кыпчакско-половецкому и огузскому типам тюркских языков. Неоднородность выявляется на фонетическом, морфологическом и лексическом уровнях [1].

Турецкий язык относится к огузской подгруппе, в результате чего эти языки объединяет много общего, хотя мы можем наблюдать ряд различий.

Как известно, турецкий язык использует алфавит в 29 букв, основанный на латинской графике. Что касается крымскотатарского языка, то до 1928 г. как и большинство тюркских языков, он пользовался арабским письмом, с 1928 г. — латинизированным, а с 1938 г. — русской графикой (35 букв), которая официально используется и по сей день. В настоящее время существует и его аналог, основанный на латинской графике (31 буква).

Существительное

Категория числа в крымскотатарском и турецком языках выражается аффиксом **-лар/-лер** и означает множественность и совокупность: **эв-лер / evler** 'дома', **дуйгъулар / duygular** 'чувства'. Именные основы без **–лар / -lar** обозначают отдельный предмет и весь его класс, а также парные органы: **айван / ayvan** 'животное', 'животные вообще', **козь / göz** 'глаз', 'глаза'. При названиях неисчисляемых предметов и явлений **-лар** означает их обилие: **къар-лар / karlar** 'снега', **ягъмур-лар / yağmurlar** 'дожди'. При собственных именах лиц **-лар** означает их окружение: **Асан-лар / Asanlar** 'Асан и его окружение' (семья или друзья), иногда со вставкой **-а-**: **Илъяс-а-лар / İlyasalar** 'Ильяс и его окружение'.

Категория множественного числа в обоих языках образуется одинаковым набором аффиксов.

Таблица 1. Категория множественного числа

	КРЫМСКОТАТАРСКИЙ ЯЗЫК			**ТУРЕЦКИЙ ЯЗЫК**	
Pl	-лАр -lAr[3]		**Pl**	-lAr	
	-лар -lar	алмалар (almalar) – яблоки		-lar	- kitaplar
	-лер -ler	эриклер (erikler) – сливы		-ler	- evler

[3] Здесь и далее аффиксы в крымскотатарском языке даны на кириллице и латинице, для более наглядного сравнения с турецким языком.

Категория принадлежности также практически полностью совпадает. Отличием является то, что в первом и во втором лице множественного лица в крымскотатарском языке неизменно присоединяются аффиксы –**мыз /–миз, -нъыз / -нъиз** в соответствии с правилами сингармонизма, тогда как в турецком языке после **у/u** и **ю/ü** данные аффиксы меняются на **–muz /-müz**, **-nuz / -nüz**.

Таблица 2.

КАТЕГОРИЯ ПРИНАДЛЕЖНОСТИ

TAGS	КРЫМСКОТАТАРСКИЙ ЯЗЫК		TAGS	ТУРЕЦКИЙ ЯЗЫК	
P1sg	-[Ы]м -Im -ым/ -им/ -ум/ -юм/ -м -ım/-im -um/-üm -m	торуным (torunım) – мой внук денъизим (denizim) – мое море ёлум (yolum) – мой путь юкюм (yüküm) – мой груз/багаж анам (anam) – моя мама	**P1sg**	-Hm -m -ım -im -um -üm	-babam - ağacım - kalemim -yolum -yüküm
P2sg	-[Ы]нъ — Iñ -ынъ/ -инъ/ -унъ/ -юнъ -нъ -ıñ/ -iñ -uñ/ -üñ -ñ	торунынъ (torunıñ) – твой внук денъизинъ (deniziñ) – твое море ёлунъ (yoluñ) – твой путь юкюнъ (yüküñ) – твой груз/багаж ананъ (anañ) – твоя мама	**P2sg**	-Hn -n -ın -in -un -ün	-baban - ağacın - kalemin - yolun - yükün
P3.sg	-[с]Ы -(s)I -ы/ -и/ -сы/ -си -у/ -ю	торуны (torunı) – его(ее) внук денъизи (deñizi) — его(ее) море анасы (anası) – его(ее) мама эмджеси (emcesi) – его(ее) дядя ёлу (yolu) – его(ее) путь юкю (yükü) – его(ее) груз/багаж	**P3.sg**	-(s)H -ı - i -sı -si -u -ü	- ağacı - denizi - babası - gazetesi - yolu - yükü

P1.pl	-[Ы]мЫз -(I)mIz -мыз/ -миз/ -ымыз/ -имиз -умыз/ -юмиз -mız/-miz -ımız/-imiz -umız/-yümiz	анамыз (anamız) – наша мама эмджемиз (emcemiz) – наш дядя торунымыз (torunımız) – наш внук денъизимиз (deñizimiz) – наше море ёлумыз (yolumız) – наш путь юкюмиз (yükümiz) – наш груз/багаж	**P1.pl**	- (H)mHz -mız -miz -ımız -imiz -umuz -ümüz	- babamız - annemiz - kitabımız - gazetimiz - yolumuz - yükümüz
P2.pl	-[Ы]нъЫз -[I] ñIz -нъыз/ -нъиз -ынъыз/ -инъиз -унъыз/ -юнъиз -ñız/ — ñiz - ıñız / -iñiz -uñız/ -yüñiz	торунынъыз (torunıñız) – ваш внук денъизинъиз (deniziñiz) – ваше море ёлунъыз (yoluñız) – ваш путь юкюнъиз (yüküñiz) – ваш груз/багаж ананъыз (anañız) – ваша мама эмдженъиз (emceñiz) – ваш дядя	**P2.pl**	- (H)nHz -nız -niz - ınız -iniz - unuz - ünüz	- babanız - anneniz - kitabınız - gazetiniz - yolunuz - yükünüz
P3.pl	-[лар]Ы -(lar)I -(лар)ы/ (лер)и	торунлары (torunları) – их внуки юклери – их грузы/багаж	**P3.pl**	-lArH -ları -leri	kitapları gazeteleri

Обратившись к падежам, мы представим их группами в зависимости от степени различий. Именительный, исходный и местно-временной падежи не имеют отличий.

Таблица 3.

ИМЕНИТЕЛЬНЫЙ, ИСХОДНЫЙ И МЕСТНО-ВРЕМЕННОЙ ПАДЕЖИ

Tags	**КРЫМСКОТАТАРСКИЙ ЯЗЫК**		**Tags**	**ТУРЕЦКИЙ ЯЗЫК**	
Nom	-	терек (terek), долап (dolap)	**Nom**	-	sepet, dolap

Abl	- ДАн -DAn -дан/ -ден/ -тан/ -тен/ -dan/-den -tan/-ten	сувдан (suvdan) – от (из) воды эвден (evden) – от (из) дома долаптан (dolaptan) – от (из) шкафа теректен (terekten) – от (из) дерева	**Abl**	- DAn - dan /den - tan/ten	- sudan - evden - dolaptan - sepetten
Loc	-ДА -DA -да/ -де/ -та/ -те -da/-de -ta/-te	сувда (suvda) – в (на) воде долапта (dolapta) – в шкафу эвде (evde) – в доме теректе (terekte) – в (на) дереве	**Loc**	-DA - da /de - ta/te	- suvda - evde - dolapta - sepette

Родительный и винительный падежи в крымскотатарском языке характеризуются фиксированным, менее широким набором аффиксов – по 2 аффикса (Gen. – **-нынъ/-нинъ**, Acc. – **-ны/-ни**), тогда как в турецком языке первая **n** в аффиксах **–nHn** и **–nH** выпадает при окончании слова на согласную, а также гласные **u** и **ü** влияют на появление новых форм данных аффиксов **–nun**, **-nün**, **-nu**, **-nü**. Следовательно в турецком языке мы наблюдаем в четыре раза больше вариантов аффиксов **–nHn** и **–nH**, чем в крымскотатарском (Gen. **-(n)in**, **-(n)ün**, **-(n)ın**, **-(n)un**; Acc. **-(y/n)i, -(y/n)ü, -(y/n)ı, -(y/n)u)**.

Таблица 4.

РОДИТЕЛЬНЫЙ И ВИНИТЕЛЬНЫЙ ПАДЕЖИ

	КРЫМСКОТАТАРСКИЙ ЯЗЫК			**ТУРЕЦКИЙ ЯЗЫК**	
Gen	-нЫнъ -nI -нынъ/ -нинъ	къуюнынъ — колодца эвнинъ — дома	**Gen**	-nHn -(n) in -(n) ün -(n) ın -(n) un	- evin - gülün -masanın -kuyunun
Acc	-н[Ы] -ны/ -ни/	къуюны — колодца эвни — дома	**Acc**	-(n)I -(y/n) i -(y/n) ü -(y/n) ı -(y/n) u	- evi - gülü -masanı -kuyunu

Направительный падеж в обоих языках имеет наибольшее количество различий, как в морфемах, так и послелогах. В данном падеже в крымскотатарском и турецком языках морфемы отличаются друг от друга (КТЯ **–ГЪА**, ТЯ **–(y)A**) и набор аффиксов в крымскотатарском языке шире, что необычно, так как до этого мы наблюдали обратное. Также следует

отметить большое количество послелогов в турецком языке – 8, тогда как в крымскотатарском языке их всего 4 (хотя все они имеют соответствия в турецком). При разметке направительный падеж в крымскотатарском языке, как и в татарском языке, делиться на DirDat и Dir_Lim, что мы не наблюдаем в турецком языке. Причиной этому в крымскотатарском языке является присоединяемая к основе слова морфема –**ГЪАдже**, которая свидетельствует о временном или пространственном ограничении направительного падежа. Ограничителями времени и пространства в турецком языке выступают только послелоги (всего их 5), два из которых имеют соответствия в крымскотатарском языке. Следует отметить, что турецкими исследователями в области корпусной лингвистики данные послелоги не рассматриваются как показатели направительного падежа. В данной таблице они представлены для сравнения с крымскотатарским языком.

Таблица 5.

НАПРАВИТЕЛЬНЫЙ ПАДЕЖ

	КРЫМСКОТАТАРСКИЙ ЯЗЫК			**ТУРЕЦКИЙ ЯЗЫК**	
Dir_Dat	-ГА -гъа -къа -ге -ке - А догъру - ДАн тышары	сувгъа – (к) воде, в воду долапкъа – (к) шкафу, в шкаф эвге – (к) дому, в дом терекке – (к) дереву, в дерево - эвге догъру - мектептен тышары	**Dat**	-(y)A -(y)a -(y)e -a doğru -DAn dışarı -A içeri	-duvara, -kediye -eve doğru - okuldan dışarı - okuldan içeri
Dir_Lim	-ГЪАдже - гъадже - гедже - къадже - кедже - А къадар -ДАн берли	акъшамгъадже – до вечера о ергедже – до того места терекке къадар – до дерева - сабадан берли		A kadar -DAn beri -A dek -A değin -DAn öte	- duvara kadar - 50’lerden beri - ölünceye dek - gülünceye değin - İzmir’den öteye

Инструментальный падеж в крымскотатарском и турецком языках совпадает только по единственному послелогу в обоих языках – **иле**. Данный послелог употребляется реже в турецком языке, так как предпочтение отдаётся аффиксу **–(y)lA**. В крымскотатарском языке он также редко употребляется в силу предпочтения ему аффикса **–нен**.

Таблица 6.

ИНСТРУМЕНТАЛЬНЫЙ ПАДЕЖ

	Крымскотатарский язык			Турецкий язык	
Ins	-нЕн иле	къолнен – рукой таякънен – палкой/ с палкой таякъ иле – палкой/ с палкой	**Ins**	- (y) lA ile	- trenle, - kediyle köpek, - duvar ile, - kedi ile köpek

Аффиксы именных аттрибутивов практически полностью совпадают в крымскотатарском и турецком языках.

Таблица 7.

ИМЕННЫЕ АТТРИБУТИВЫ

Тэги	Крымскотатарский язык		Тэги	Турецкий язык	
ATTR_ABES	-лЫ -лы -ли -лу -лю	- китаплы - буберли - сувлу - сютлю	*Minor-POS* **With**	-lH -lı -li -lu -lü	- kitaplı - biberli - sulu - sütlü
ATTR_MUN	-сЫз -сыз -сиз -суз -сюз	- китапсыз - буберсиз - сютсюз	*Minor-POS* **Without**	-sHz -sız -siz -suz -süz	- kitapsız - bibersiz - susuz - sütsüz
ATTR_LOC	-ДА+кИ -даки/ -таки/ -деки/ -теки/	китаптаки	*Minor-POS* **Rel**	-DAkH	- kitaptaki - evdeki
ATTR_GEN	-[ны] ЫНЪ+кЫ -нынъки -нинъки	китабынъки		- [n]HnkH	- kitabınki - masanınki

ГЛАГОЛ

Настоящее время

Времена глагола в крымскотатарском языке имеют ряд отличий от таковых в турецком языке. В турецком языке существует Geniş zaman

(«широкое время») – время, которое обозначает действие или состояние, которое началось в прошлом, происходит или является действительным в настоящем и будет происходить в будущем – ***yazar, okur, gelir***. В крымскотатарском языке отсутствует специальное время для обозначения подобного действия. Простое настоящее время в крымскотатарском языке может иметь как значение повторяющегося действия в настоящем, так и настоящего длительного действия в зависимости от контекста – ***яза, окъуй, келе***.

Настоящее длительное действие в турецком языке образуется с помощью аффикса **-(H)yor** ***(geliyor, yazıyor)***, тогда как в крымскотатарском языке посредством аффикса **–мАктА** (***кельмекте/ kelmekte***). Данная форма имеет множество ограничений в употреблении. Она в основном функционирует в положительном аспекте и фактически лишена форм других аспектов: отрицательного, возможности, невозможности. В обоих языках данная форма настоящего времени не является распространенной и в крымскотатарском языке ей предпочтут форму ***келе (kele)***. Некоторые турецкие грамматики пишут, что настоящее длительное время также можно выразить с помощью суффикса **–mAktA** ***(gelmekte)***, но специалисты утверждают, что в турецком языке посредством данного суффикса можно образовать только деепричастие несовершенного вида.

Таблица 8.

Tags	Crımean Tatar	Examples	Tags	Turkish	Examples
PRES	-E / -E -a (-a) -e (-e) -й (-y)	яза (yaza) юре (yure) окъуй (oquy)		- Ar - ar - er - ir - ır - ur	yazar keser gelir kırır okur
PRES_CONT	-мАкъта / -mAkta -макъта (-makta) - мекте (-mekte)	язмакъта (yazmakta) юрмекте (yurmekte)	Pres	-(H)yor -ıyor -iyor -uyor -üyor	- yazıyor - geliyor - uyuyor - gülüyor

Будущее время

Будущее постоянное время в крымскотатарском и турецком языках образуется с помощью разных морфем: КТЯ – **(Ы)Р** ***отурыр***, ТЯ **–AcAk** ***oturacak***. Но будущее категорическое действие в обоих языках имеет одинаковую форму образования: **АджАкъ / AcAk** ***пишиджек – pişicek***.

Следует отметить, что в турецком языке будущее категорическое время также может быть образовано с помощью аффикса настоящего длительного действия **-(H)yor *(geliyor)***. Можно провести параллель между турецким и английским языками, где глагол в длительном настоящем времени (the Present Continuous) в зависимости от контекста может также иметь значение будущего действия (the Future Indefinite). Например: I am going to school now (the Present Continuous). I am going to the doctor tomorrow.

Следует отметить, что будущее постоянное время (Future Indefinite) в крымскотатарском языке и простое настоящее время (the Present Simple) в турецком языке образуется с помощью одной и той же морфемы с её алломорфами – **(Ы)Р** – **(Н)R**. Например: (ТЯ) Ben her gün yazarım. (Я пишу каждый день) (КТЯ) Мен эр кунь язарым / Men er kün yazarım. (Я буду писать каждый день).

В отрицательной форме будущего времени аффикс **–(Ы)Р** в крымскотатарском языке во 2-м и 3-м лице единственного и множественного числа меняется на аффикс **–З**: ***язарсынъ – язмазсынъ***, а в 1-м лице единственного и множественного числа вообще выпадает: ***язарым – язмам*** (тоже самое происходит в турецком языке в простом настоящем времени – ***yazarım – yazmam***).

Таблица 9.

Tags	Crımean Tatar	Examples	Tags	Turkish	Examples
FutDef	-джАкъ / -cAq -джакъ (-caq) -джек (-cek)	язаджакъ (yazacaq) юреджек (yurecek) пишиджек (pişicek)	**Fut**	- (H) yor - AcAk	yazacak gelecek gidecek
FutIndf	-[Ы]Р / (I)R -р (-r) -ар (-ar) -ер (-er) -ыр (-ır) -ир (-ir)	язар (yazar) юрер (yurer) пишер (pişer) къырыр (kırır)		- AcAk -acak -ecek	yapacak gelecek
FutIndfNeg	мА (mA) (-З) (Z) (во 2-м и 3-м лице ед. и мн. числа) -ма (-ma) -ме (-me)	язмаз (yazmaz) юрмез (yurmez)		-mA-Z -maz -mez	yapmaz gelmez

Прошедшее время

Прошедшее категорическое время в крымскотатарском и турецком языках совпадает по морфемам, но не по количеству алломорфов, и описывает действие, которое уже совершилось к моменту речи и говорящий является очевидцем этого. Оно образуется путем прибавления аффикса **Д(Ы) – D(H)**: (КТЯ) ***язды (yazdı)*** – (ТЯ) ***yazdı***.

Прошедшее неочевидное время в обоих языках образуется с помощью разных аффиксов: в крымскотатарском языке – **-ГЪАн**, в турецком – **-mHş**. Оно выражает действие, которое совершилось, однако говорящий знает об этом только по словам других или видел уже результат действия. Понятие неочевидности в дальнейшем будет указываться в тэгах как **Indf**, соответствующий в английском языке indefinite (неопределенный).

Таблица 10.

Tags	Crımean Tatar	Examples	Tags	Turkish	Examples
PstDef	-ДЫ / -DI -ды (-dı) -ди (-di) -ты (-tı) -ти (-ti)	- язды (yazdı) - юрди (yurdi) - сатты (sattı) - пишти (pişti)	**Past**	-DH -dı -di -tı -ti -du -dü -tu -tü	- yazdı - geldi - attı - kesti - vurdu - yüzdü - tuttu - öptü
PstIndf	-гъАн / -ĞAn -гъан (-ğan) -ген (-gen) -къан (-qan) -кен (- ken)	- язгъан (yazğan) -юрген (yurgen) -саткъан (satkan) -пишкен (pişken)	**Narr**	-mHş - mış - miş - muş - müş	- kırmış - gelmiş - koşmuş - öpmüş

Аналитические формы прошедшего времени с глаголом *эди*.

Следующие пять прошедших времен в крымскотатарском языке, в отличие от турецкого языка, образованы аналитическим способом, который представляет собой соединение двух глагольных форм: основа простого времени + прошедшее время от глагола **–э (эди)**. В турецком же языке эти глагольные формы объединены и данные времена образуются синтетическим способом.

Время определенный имперфект (К. Мусаев) [2;с.215-216] или «прошедшее время данного момента» образуется путем присоединения к основе глагола аффикса настоящего времени данного момента **–а/ -е/ -и** + **эди** + аффиксы лица и числа (***яза эдим*** (*в тот момент* брала), ***яза эдинъ*** и

т.д.) В турецком языке ему соответствует настоящее историческое время, образованное с помощью аффиксов **–yor** + **-du/-dü** (***yazıyordu, yazıyordum***). Можно провести параллель с английским языком, где данное время будет соответствовать времени the Past Continuous.

Неопределенный имперфект, обозначающий действие, которое в прошлом совершалось продолжительно, неопределенно длительно или постоянно, к моменту речи завершенное или незавершенное, в обоих языках практически совпадает (КТЯ ***язар эди (yazar edi)*** – ТЯ ***yazardı***). Тоже самое можно сказать и о форме отрицательного аспекта, в которой форма аффикса времени изменяется с **-Р** на **-З** (КТЯ ***язмаз эди (yazmaz edi)*** – ТЯ ***yazmazdı***). В обоих языках мы наблюдаем данные изменения в форме отрицания во всех временах, где фигурирует аффикс **–ар (-ar)**. В английском языке данное время соответствует конструкции **used to**, которое также обозначает повторяющееся или длительное действие в прошлом.

Одним из основных отличий давнопрошедшего времени в крымскотатарском и турецком языках являются суффиксы **–гъАн** и **–mHş**, к которым прибавляется в первом случае служебное слово **эди**, а во втором случае суффикс **DH (-tı/-ti)** (КТЯ – ***язгъан эди***, ТЯ – ***yazmıştı***). Данное время обозначает действие, которое по мнению говорящего, совершилось задолго до момента речи или прежде другого прошедшего действия. Таким образом, употребление этой формы предполагает наличие другого действия или события, которое совершается после обозначенного формой действия, что является весьма схожим со временем the Past Perfect в английском языке.

Прошедшее длительное время в крымскотатарском и турецком языках имеет одинаковый аффикс **мАктА (макта/-мекте) – mAktA (-makta/-mekte)** и отличается способом образования с использованием в первом случае вспомогательный глагол **эди (*язмакта эди, пиширмекте эди*)**, а во втором случае – аффикса **DH (-(y)dı/(y)di) (*yazmaktaydı, pişirmekteydi*)**. Следует отметить, что и прошедшее длительное время и определенный имперфект обозначают длительное, незаконченное действие в прошлом, но и в крымскотатарском и в турецком языках прошедшее длительное время употребляется крайне редко, отдавая предпочтение определенному имперфекту.

Будущее-прошедшее время в крымскотатарском языке образуется путем присоединения к глагольной основе аффикса будущего категорического времени **-АджАкъ (-аджакъ/-еджек)** и глагола **эди (*язаджакъ эди, пиширеджек эди*)**. В турецком же языке – по модели глагол + **-AcAk (-acak/ecek)** + **DH (-tı/-ti) (*yazacaktı, pişirecekti*)**. В обоих языках данное время обозначает действие, которое говорящий намеревался совершить в ближайшем будущем, но не смог осуществить его. В английском языке данному времени соответствует время Future-in-the-Past,

в одном из его значений, вследствие чего это отражено в тэге для обозначения будущее-прошедшего времени в крымскотатарском языке.

Таблица 11.

Tags	Crımean Tatar	Examples	Tags	Turkish	Examples
PstCont1	- A (A) + эди (edi) -a (a) -e (e) -и (i)	яза эди (yaza edi) пишире эди (pişire edi)	**ProgPast**	-(H)yor + -DH	yazıyordu pişiriyordü
UsedTo	- Ap (Ar) + эди (edi) -p (r) -ep (er) -ap (ar)	язар эдим (yazar edim) пиширер эдим (pişirer edim)	**AorPast**	-(A)r, -(H)r + -DH -dı -di	yazardı pişirerdi
UsedToNeg	-мА/-мЕ (mA/mE) + -з(z) + эди (edi)	язмаз эдим (yazmaz edim) пиширмез эдим (pişirmez edim)		-mA/-mH + -z + -DH	yazmazdı pişirmezdi
PastPast	-гъАн (ĞAn) + эди (edi) -ğan -gen	язгъан эди (yazğan edi) пиширген эди (pişirgen edi)	**Pqp**	-mHş + -DH -mış -miş+ -tı- ti	yazmıştı pişirmişti
PstCont2	-мАкъта + эди (-mAkta+edi) -макъта (-makta) - мекте (-mekte)	язмакъта эди (yazmakta edi) пиширмекте эди (pişirmakte edi)		-mAktA + -(y)DH -makta -mekte -dı -di	yazmaktaydı pişirmekteydi
FutPast	-АджАкъ (AcAk) + эди (edi) -аджакъ (acak) -еджек (ecek)	язаджакъ эди (yazacak edi) пиширеджек эди (pişirecek edi)	**FutPast**	-AcAk + -DH -acak -ecek	yazacaktı pişiricekti

Аналитические формы прошедшего времени с глаголом *экен*.

Следующие четыре формы прошедшего неочевидного времени в крымскотатарском языке образованы аналитическим способом с помощью формы служебного глагола **э – экен**. В результате этого данные времена

крымскотатарского языка в своей форме имеют наибольшие отличия от турецкого языка.

Прошедшее-неочевидное время данного момента в крымскотатарском языке образуется путем присоединения к основе глагола аффикса настоящего времени **–а, -е, -й** и глагола **экен** (***яза экен, яза экенсинъ*** и т.д.).

В турецком языке оно образуется путем прибавления к основе аффиксов **-(ı)yor** + **-muş** (***yaz-ı-yor-muş, yaz-ı-yor-muş-sun*** и т.д.), в третьем лице множественного числа **-(ı)yor + -lar + -mış (yaz-ı-yor-lar-mış)**.

Прошедшее неочевидное многократно-длительное время создается путем присоединения к основе глагола аффикса **–р, -ар, -ер** + служебный глагол **экен** (***язар экен, язар экенсинъ*** и т.д.).

Данное время выражает многократное, длительное действие, имевшее место в прошлом, но говорящий знает и судит о нем со слов другого лица, других лиц (2; с. 221). В турецком языке данное время образуется с помощью аффиксов **–r, -er/-ar** + **-mış** (***yazarmış, yazarmışsınız*** и т.д.). Отрицательная форма в обоих языках образуется нестандартным способом, так как после отрицательного аффикса **–ма/-ме** аффикс **–р, -ар, -ер** заменяется на **– з** (***язмаз экен, кельмез экен***). Такая же ситуация и в турецком языке – ***yazmazmış, pişirmezmış***.

Давнопрошедшее неочевидное время, обозначающее действие, которое действительно имело место в прошлом, однако говорящий не был его очевидцем, образуется в крымскотатарском языке при помощи аффикса **–гъан/ -ген** и служебного слова **экен** (***язгъан экен, пишырген экен***).

В турецком языке данное время образуется при помощи двойного аффикса **–mış (-mış/-miş + -mış)** (***yazmışmış, pişirmişmış***), что свидетельствует об абсолютной разнице в форме в сравнении с крымскотатарским языком.

Таблица 12

Tags	Crımean Tatar	Examples	Tags	Turkish	Examples
PstContIndf	- А (А) + экен (eken) -а (а) -е (е) -и (i)	яза экен (yaza eken) пишире экен (pişire eken)	**ProgPqp**	-(H)yor + -mHş -muş -müş	yazıyormuş üzülüyormüş
UsedToIndf	- Ар(Ar) + экен (eken) -р (r) -ер (er) -ар (ar)	язар экен (yazar eken) пиширер экен (pişirer eken)	**PresPqp**	-(A)r, -(H)r + -mHş -mış -miş	yazarmış pişirermiş

UsedToIndf	-мА/-мЕ (ma/me) + -з (z) + экен (eken)	язмаз экен (yazmaz eken) пиширмез экен (pişirmez eken)		-mA/-mH + -z + -mHş -mış -miş	yazmazmış pişirmezmiş
PstPstIndf	-гъАн (ğAn) + экен (eken) -гъан (ğan) -ген (gen)	язгъан экен (yazğan eken) пишырген экен (pişirgen eken)	**NarrPqp**	-mHş + -mHş -mış -miş	yazmışmış pişirmişmış
FutPastIndf	-АджАкъ (AcAk) + экен (eken) -аджакъ (acak) -еджек (ecek)	язаджакъ экен (yazacak eken) пишыреджек экен (pişirecek eken)	**FutPqp**	-AcAk + -mDş -acak -mış -ecek -miş	yazacakmış pişiricekmiş

Заключение

В данной работе представлен подробный сравнительный анализ крымскотатарского и турецкого языков в плане грамматической разметки электронных корпусов исследуемых языков. Данное исследование затрагивает только такие части речи как существительное и глагол, что свидетельствует о необходимости изучения и других частей речи в обоих языках.

ЛИТЕРАТУРА:

1. Лингвистический энциклопедический словарь http: // tapemark.narod.ru/les/246d.html
2. Меметов А., Мусаев К. Крымтатарский язык. Ч. I. Общие сведения о языке. Ч. II Морфология. Учебное пособие. – Симферополь: Крымское учебно-педагогическое государственное издательство, 2003. – 288 с. – На русском языке. ISBN 966-8025-35-0
3. Mustafa Erkan The comparative essay of Crimean Tatar Turkish and Turkey Turkish. The Journal of International Social Research. – April 2015. Volume 8. Issue 37.
4. Kemal Oflazer, Bilge Sayö Dilek Zeynep Hakkani-Tür Building a Turkish Treebank. http://www.andrew.cmu.edu/user/ko/downloads/Papers/TurkishTreebank-Chapter.pdf
5. Lenara Kubedinova, Ayrat Gatiatullin Morphological tagging of Crimean Tatar electronic corpus / Proceedings of the International Conference «Turkic Languages Processing: Turklang-2015». – Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. – 488 с. ISBN 978-5-9690-0262-3

FORMATION OF THE SYNTHETIC CORPORA FOR KAZAKH ON THE BASE OF ENDINGS COMPLETE SYSTEM

A. Karibayeva, B.Abduali, D. Amirova, Al-Farabi Kazakh National University,
Institute of Information and Computational Technologies, Almaty, Kazakhstan,
a.s.karibayeva@gmail.com , balzhanabdualy@gmail.com, amirovatdina@gmail.com

The problem of absence of parallel corpora are actual for a large number of language pairs and can severely detriment the quality of neural machine translation systems. The lack of parallel corpora for Kazakh is actual in machine translation system. The creation and collection of corpus limits the creation of a neural machine translation with a good quality of translation. Since, the mentioned kind of machine translation needs large data for system training. For this reason was created synthetic corpora to extend the number of sentences to training Kazakh neural machine translation system(NMT). As, synthetic corpora is mentioned the sentences that automatically created from program that generated construction by part of speech and their description of changing by person, case and in number. The method is a language-dependent to enable machine translation between a low-resource language and a high-resource language, e.g. English and Russian. Kazakh language has 8 types of changing by person, 2 types of changing in number, 7 types of connection dependency and 6 cases. 'Мен [Men] (I)', 'Сен [sen] (you)', 'Сіз [Siz] (you)', 'Ол [ol] (he)', 'Біз [biz] (we)', 'Сендер [sender] (you)', 'Сіздер [sizder] (you)', 'Олар [olar] (they)' — the types of person. The singular and plural is type of number. The connection dependency based on the possessive form of nouns. Belonging in the Kazakh language is expressed with the help of the endings of belonging — 'тәуелдік жалғау'. Such a construction "noun" + "end of belonging" is also called the possessive form of nouns. The word in the Kazakh language is based on adding an ending to the stem. Taking into account the number of types of change, a complete system of word endings was created, which consists of 3550 possible combinations of endings structures. Synthetic corpora is created from the longest construction of the offer to the shortest. The novelty of approach in generation synthetic corpora by using sentence structure pattern and complete set of endings. In this paper will be shown the creating process of synthetic corpora of Kazakh language by sentence construction. The results are shown in number of created sentences for Kazakh-English, Kazakh-Russian language pairs.

Key words: *synthetic corpora; parallel corpora; neural machine translation system; set of endings, Kazakh language.*

ФОРМИРОВАНИЕ СИНТЕТИЧЕСКОЙ КОРПОРА ДЛЯ КАЗАХА НА ОСНОВЕ ЗАВЕРШЕНИЯ СИСТЕМЫ

А. Карибаева, Б. Абдуали, Д. Амирова,
Казахский национальный университет им. Аль-Фараби, факультет информационных систем, Пр-т Аль-Фараби, 71, 050040, Алматы, Казахстан;
Институт Информационных и Вычислительных Технологий, ул. Пушкина, 125, 050010, Алматы, Казахстан, a.s.karibayeva@gmail.com , balzhanabdualy@gmail.com, amirovatdina@gmail.com

Проблема отсутствия параллельных корпусов актуальна для большого числа языковых пар и может серьезно ухудшить качество систем нейронного машинного перевода. Отсутствие параллельных корпусов для казахского языка актуально в системе машинного перевода. Создание и сбор корпусов ограничивают создание нейронного машинного перевода с хорошим качеством перевода. Поскольку упомянутый вид машинного перевода требует больших данных для обучения системы. По этой причине были созданы синтетические корпуса для расширения количества предложений для обучения казахской системе нейронного машинного перевода (НМП). В качестве синтетических корпусов упоминаются предложения, которые автоматически создаются из программы, которая генерировала конструкцию по части речи, и их описанию, изменяющееся по лицу, падежу и количеству.

Способ зависит от языка, чтобы обеспечить возможность машинного перевода между языком с низким уровнем ресурсов и языком с высоким ресурсом, например английский и русский. Казахский язык имеет 8 типов изменения по лицу, 2 типа изменения по количеству, 7 типов притяжания и 6 падежей. «Мен (Я)», «Сен (Ты)», «Сіз (вы)», «Ол (он, она)», «Біз (мы)», «Сендер (вы)», «Сіздер (вы)», «Олар (они)» — типы по лицу. Единственное и множественное число является типом числа. Зависимость соединения основана на притяжательной форме существительных. Принадлежность в казахском языке выражается с помощью окончаний принадлежности — ‘тәуелдік жалғау’. Такая конструкция «существительное» + «конец принадлежности» также называется притяжательной формой существительных. Слово в казахском языке образуется при добавлении окончания к основанию. С учетом количества типов изменений была создана полная система окончаний слов, состоящая из 3550 возможных комбинаций структур

окончаний. Синтетические корпуса создаются от самой длинной конструкции предложения до самой короткой. Новизна подхода в генерации синтетических корпусов с использованием шаблона структуры предложений и полного набора окончаний. В данной статье будет показан процесс создания синтетических корпусов казахского языка по конструкции предложения. Результаты показаны в количестве созданных предложений для казахско-английских, казахско-русских языковых пар. Этот метод зависит от языка, обеспечивающий машинный перевод между языком низкого ресурса и языком высокого ресурса, например, английский и русский.

***Ключевые слова**: синтетические корпуса; параллельные корпуса; система нейронного машинного перевода; система окончаний, казахский язык.*

Introduction

The creating the neural machine translation system required a big number of data. For low-resources languages, like a Kazakh needs to qualitative parallel corpora.

Kazakh language is agglutinative language with rich morphology with various combinations of suffixes. This language doesn’t have enough resources like linguistic resources and parallel corpora. So, Kazakh Language is low-resource language. Lack of data is the main problem for creating a neural machine translation with high quality for the Kazakh language. For having good translation with NMT it should to train significant number of data.

For that reason we present method to create synthetic corpora for Kazakh language on the base of complete set of endings[1]. Each part of speech has its own characteristics and its kinds of endings, which it can have. There are about 3550 combinations of endings. Based on this complete set of endings, tables were created for all parts of the speech of the Kazakh language.This paper is structured as follows. Related works are described in section 2. In section 3 we present method of generating synthetic corpora. Results are discussed in section 4. Finally, conclusion is given in section 5.

Related works

The most of work were considered with researchers. The absences of parallel data were inspired researchers to creating and investigation the low-resources languages. The Kazakh language related to low resources languages too. Under synthetic corpora the most considers the automatically-generated corpora, translated texts from different translation systems, and etc. Anna Currey and et al. used monolingual data with mixing main corpora in target language, namely to Romanian and Turkish languages to train the NMT system for low-resource. This method improved the BLEU to 1.2 for latter languages[2].

One of the methods of generating synthetic parallel corpora is using back-translation. That means NMT system is trained in the reverse translation direction

(target-to-source), and is then used to translate target-side monolingual data back into the source language (in the backward direction, hence the name back translation)[3]. The received sentences can be added to the existing training data and increase a volume of synthetic parallel corpus. In [3] authors for training NMT systems use iterative back-translation for generating synthetic parallel corpora from monolingual data. They used method to both high (German-English) and low (English-French, English-Farsi) resourced scenarios.

In [4] presented dual learning method on English-French language pairs.They develop a dual-learning mechanism, which can enable an NMT system to automatically learn from unlabeled data through a dual-learning game[4].

Generation process of synthetic corpora for Kazakh language

The process of generation depend on language direction. The proposed method of synthetic generation based on part of speech and Kazakh endings. As all Turkic languages Kazakh is agglutinative language. The word forms constituted from adding suffixes to the base of word.

The complete set of endings used for create synthetic corpora. Based on complete set of Kazakh endings was created morphological language. It consists about 3550 various combinations of endings.

The structure of sentences changed by person, case and etc. For example one of the part of speech presented in table 1.

Table 1.

The tense of Kazakh language, structural forms, examples and their communication with English language.

The tense of Kazakh language and transliteration	Grammar structure for Kazakh	Example for Kazakh and transliteration	English names of the tense	Grammar structure for English	English translation
Нақосы шақ (Nakosy shak)	V+A(PresSm)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істеп жүрмін (Men istep zhurmin)	Present Simple	V	I work

Нақ осы шақтың күрделі түрі (Nakosy shaktyng kurdeli turi)	V+A(PresComp(PresSm))+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істеп жатырмын (Men istep zhatyrmyn)	Present Continuous	to be + V + ing	I am working
Ауыспалы осы шақ (Auyspaly osy shak)	V+A(PresNow)+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істедім (Men istedim)	Present Perfect	to be + V + ed	I have worked
Жедел өткен шақ (Zhedel otken shak)	V+A(PastOper)+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істедім (Men istedim)	Past Simple	V + ed	I worked
Бұрынғы өткен шақ (burynggy otken shak)	V+A(PastOld)+(Sg,Pl)+(P1,P2,P2в,P3)	Мен істегенмін (Men istegenmin)	Past Continuous	to be + V + ing	I was working

Ауыс палы өткен шақ (Auys paly otken shak)	V+A(PastMay)+(Sg,Pl)+(P 1,P2,P2в,P3)	Мен істеп отырд ым (Men istep otyrdy m)	Past Perfe ct	to be + V + ed	I had wor ked
Болжа лды келер шақ (Bolzh aldy keler shak)	V+A(FutCast)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істей мін (Men isteimi n)	Futur e Simp le	to be + V	I shall wor k
Мақса тты келер шақ (Maks atty keler Shak)	V+A(FutObj)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істеп отырм ын (Men istep otyrm yn)	Futur e Cont inuo us	to be + be + V + ing	I shall be wor king
Ауыс палы келер шақ (Auys paly keler shak)	V+A(FutSub)+(Sg,Pl)+(P1, P2,P2в,P3)	Мен істей мін (Men isteimi n)	Futur e Perfe ct	to be + hav e + V+ ed	I shall have wor ked

Similarly, we fill out the table of all parts of speech and get many options for ending. Then with helping this complete set of endings created files, and prepared sentence structure. Each part of speech are in different files and connect to the software part. For example one structure of sentences «Сіз мектепке бүгін ерте келдіңіз», for this sentence created 6 files:

- pronoun (мен, сен, сіз, ол, біз, сендер, сіздер, олар);
- nouns (мектепке, жұмысқа, сабаққа, бақшаға, үйге, паркке);
- adverb1 (бүгін, кеше, арғыкүні, таңертең);
- adverb2 (ерте, кеш, асығып, жүгіріп, баяу);

- verb (кел, келме);
- endings (дім, дің, діңіз, ді, дік, діңдер, діңіздер, ді).

And similarly, create exactly the same files in English, but the sixth files must be empty, because in English does not has a suffixes. There is first and fifth files does not has many types of variants, but other nouns, adverbs can be filled more, it is help to create lots of options of sentences.

Then through the automatic generated was get following structure of sentences with changing context of words in the following table 2:

Table 2.

Automatic generated sentences.

Sentences in Kazakh	Sentences in English
Сіз мектепке бүгін ерте келдіңіз	You come to school early today
Сіз мектепке бүгін ерте келмедіңіз	You did not come to school early today
Сіз мектепке бүгін келдіңіз.	You come to school today
Сіз мектепке бүгін келмедіңіз.	You did not come to school today
Сіз мектепке ерте келдіңіз.	You come to school early
Сіз мектепке ерте келмедіңіз.	You did not come to school early
Сіз бүгін ерте келдіңіз.	You come early today
Сіз бүгін ерте келмедіңіз.	You did not come early today
Сіз мектепке келдіңіз.	You come to school
Сіз мектепке келмедіңіз.	You did not come to school
Сіз ерте келдіңіз	You come early
Сіз ерте келмедіңіз.	You did not come early
Сіз бүгін келдіңіз.	You come today
Сіз бүгін келмедіңіз.	You did not come today
Сіз келдіңіз.	You come
Сіз келмедіңіз.	You did not come

It was for one case, and exactly the same for other structural proposals, the files are created and by using automatic generation, was created the parallel synthetic corpora.

Results

The automatic generation of the sentence helps to increase the volume of corpora, and in the future, use a variety of options and structured proposal to get more sentences. Thus, the volume of the corpora increases.

The results of translation are shown in the next table 3 below.

Table 3.

Automatic generated synthetic corpora.

Corpora	Number of generated sentences
Kazakh-English	600K
Kazakh-Russian	700K

Conclusion

In this paper was considered the generation synthetic corpora of Kazakh language by using complete set endings and logical structure of sentences. The reason of using it is understand with absences of resources. By using this method we will try to extend the number of parallel corpora.

Acknowledgements

This research performed and financed by the grant Project IRN AP05132950 "Development of an information-analytical search system of data in the Kazakh language", awarded to The Republican State Enterprise (RGP) on the right of economic management (PVC) «Institute of Information and Computational Technologies».

REFERENCES:

1. Tukeyev, U., Automaton models of the morphology analysis and the completeness of the endings of the kazakh language. Proceedings of the international conference «Turkic languages processing» TURKLANG-2015 September 17–19, Kazan, Tatarstan, Russia, 2015. 91-100 pp.

2. Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In Proceedings of the Second Conference on Machine Translation, pp. 148–156, Copenhagen, Denmark, September 2017. Association for Computational Linguistics..

3. Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, Trevor Cohn.Iterative Back-Translation for Neural Machine Translation. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 18–24. Melbourne, Australia, July 20, 2018. Association for Computational Linguistics.

4. Di He, Yingce Xia, Tao Qin, LiweiWang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In D. D. Lee, M.

Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 820–828..

КЛАССИФИКАЦИЯ СЕМАНТИЧЕСКИХ РОЛЕЙ В СЕМАНТИЧЕСКОЙ РАЗМЕТКЕ ЭЛЕКТРОННОГО КОРПУСА ТЕКСТОВ ТУВИНСКОГО ЯЗЫКА[4]

А.Б. Хертек, В.С. Ондар, Тувинский государственный университет,
г. Кызыл, Россия, khertek.ab@yandex.ru

В статья содержит описание классификации семантических ролей для семантической разметки Электронного корпуса текстов тувинского языка. Для составления инвентаря семантических ролей были использованы понятия классов семантических функций: адъекты, актанты и сирконстанты. Классы актантов и сирконстантов представлены подклассами, каждый из которых включает по несколько конкретных семантических ролей.

Ключевые слова: *семантические роли, актанты, сирконстанты, адъекты, семантическая разметка.*

THE CLASSIFICATION OF SEMANTIC ROLES IN THE SEMANTIC MARKUP OF THE ELECTRONIC TEXT CORPUS OF TUVAN LANGUAGE

A.B. Khertek, V.S. Ondar, Tuvan State University,
Kyzyl, Russia, khertek.ab@yandex.ru

In this article provides a description of the classification of semantic roles to semantic markup of the Electronic text corpus of Tuvan language. To compile the inventory of semantic roles has been used the concept of classes of semantic features: adjecti, actants and circostanti. Classes of actants and circonstances represented by subclasses, each of which includes several specific semantic roles.

Key words: *semantic role, actant, circonstance, objects, semantic markup.*

[4] Работа выполнена в рамках Госзадания Министерства науки и высшего образования РФ № 34.3876.2017/4.6

Данная статья посвящена описанию классификации семантических ролей для семантической разметки Электронного корпуса текстов тувинского языка (ЭКТТЯ).

В современной лингвистике принято связывать семантические функции падежей с семантической и синтаксической ролью имени в составе определенной ситуации. Известно, что между семантическими и синтаксическими ролями имени есть зависимость, но нет прямого соответствия.

Понятие семантической роли в современную лингвистику было введено Ч. Филлмором [Филлмор 1981], использовавшим первоначально термин «глубинный падеж». Семантическая роль имени при данном предикате является частью семантики этого предиката и отражает общие свойства участников определенных групп ситуаций.

В теории структурного синтаксиса Л. Теньером была введена оппозиция в сфере участников ситуации: выделяются два класса семантических ролей – *актантные* и *сирконстантные* [Теньер 1988]. Предложенное разграничение актантов и сирконстантов было именно семантическим и не содержало указания на синтаксическую обязательность или необязательность этих участников ситуации.

Актантами он называл предметных участников события, чаще всего выражающихся именами существительными, а *сирконстантами* – непредметных участников события, или различные обстоятельства, выражающиеся как именами, так и наречиями.

«Грамматика зависимостей» Л. Теньера представила предложение как структуру, имеющую иерархию связей и отношений, которые представляются в виде узлов. Предикат, расположенный в предикативном узле, в соответствии со своими валентностными способностями распределяет места, или позиции, актантам и сирконстантам. Валентность того или иного глагола определяется количеством актантов, которыми он способен управлять [Теньер 1988: 250].

Поэтому при классификации предложений, в том числе и в тюркских языках, исследователи за «точку отсчета» брали глагол и его модель управления именными компонентами [Предикативное склонение 1984: 77; Черемисина, Озонова 2005: 15, 55].

Именно в связи с типом управления предиката представление о сущности оппозиции актантов и сирконстантов изменилось. Актанты стали связываться с обязательным управлением «сильными» валентностями глагола, и на первый план в разграничении актантов и сирконстантов вышла вовлеченность в ситуацию. Актанты стали трактоваться как обязательные участники ситуации, ее неотъемлемые составляющие, которые своим существованием создают ее, естественно, вместе с

предикатом. А сирконстанты лишь дополняют, «украшают» уже созданную ситуацию, являясь факультативными.

В связи с пониманием актантов как обязательных предметных участников ситуации возникла проблема толкования в этой системе локализаторов, которые, с одной стороны, не являются предметными участниками ситуации, а с другой стороны обязательны для ее реализации при глаголах местонахождения и движения.

В Новосибирской синтаксической школе для таких участников ситуации используется термин *актант-локализатор* (см., например, [Черемисина, Скрибник 1996: 50; Черемисина, Озонова 2005: 21]).

Ж. Лазар предложил трехчленную классификацию, которая строится на типе управления, которая нарушает бинарность противопоставления актантов и сирконстантов, [Тестелец 2001: 188–189, со ссылкой на Lazard 1998]. В этой классификации вводится понятие *адъекта*, такой синтаксической единицы, которая обязательна, но может выражаться при одном и том же предикате более одного раза, причем форма ее вариативна. Этот термин полностью соответствует понятию актанта-локализатора.

Предложенная Ж.Лазаром классификация является основой для предлагаемой нами классификации семантических ролей, где адъекты, актанты и сирконстанты представляют собой не конкретных участников конкретных ситуаций, а обобщенные, типизированные классы участников ситуации, задаваемой предикатом также определенного семантического класса.

Названия классов и подклассов участников в основном заимствованы из работ М. В. Всеволодовой [2000: 141–148], В. А. Плунгяна [2000: 165–166], Н. А. Лысковой [2003: 105−121], М. И. Черемисиной, А. А. Озоновой [2005: 150−152]. Некоторые роли, не получившие до настоящего времени специальных терминов, передаются описательно, они будут отмечены звездочкой (*).

Класс *актантов*, наиболее обширный, делится на подклассы с меньшим уровнем обобщения:

- *протагонист* – основной участник ситуации, «организующий» ее (***ректор*** *указал его фамилию в указе*);
- *пациенс* – второй по значимости после протагониста участник ситуации, подвергающийся воздействию протагониста (*старик срубил* ***дерево***);
- *адресат* – третий участник, которому протагонист направляет материальные объекты или информацию, желая, чтобы он их получил (*покажите* ***гостю*** *его комнату*; *бабушка рассказала* ***внучке*** *сказку*);
- *источник** – участник, у которого протагонист получает материальные объекты или информацию (*взять деньги* ***в банке***; *спросить* ***у друга***);

- *инструмент* – участник (обычно неодушевленный), которого протагонист использует для осуществления своей деятельности (*разбил окно* ***камнем***; *слова подчеркни* ***по линейке***);
- *ситуант* – участник ситуации, осложняющий основную пропозицию или конкретизирующий отношения части и целого (***по просьбе*** *отца*, *выпить* ***чаю***).

Каждый подкласс актантов представляет набор конкретных семантических ролей. Семантические роли реализуются разными падежными формами, поэтому в последнем столбце таблиц даются примеры с разными падежными формами, которые выражают определенную роль. Интерпретация ролей зависит от семантики предиката.

Таблица № 1.

СЕМАНТИЧЕСКИЕ РОЛИ АКТАНТОВ

Подкласс	Семантические роли	Примеры
Протагонист	Агенс	***Авам*** *ыры ырлап турду* ‘ Моя мама спела песню’; ***Оолчук*** *ойнап олур* ‘мальчик играет’; *Башкы* ***уругларга*** *шүлүк доктааттырган* ‘Учительница велела ученикам выучить стихотворение’. ***Эжимден*** *меңээ байыр чедирткен болду.* ‘Мой друг, оказывается, мне привет послал.’
	Экспериенцер	***Уруглар*** *ыттан корга берди* ‘Девочки испугались собаки’; ***Диис*** *шылыраан дааштыдыңнаалап чыткан* ‘Кот прислушивался к шуршанию’; ***Ачазы*** *аарый берген* ‘Его отец заболел’; *Ол чугаа* ***силерге*** *чиктиг кылдыр дыңналган боор* ‘Вероятно, та речь послышалась вам странной’.
	Каузатор	***Уруг*** *авазынга тон дааратканы* ‘Девочка попросила мать сшить ей пальто’.
Пациенс	Перцептив	*Сугда* ***балыкты*** *көрүп калдывыс* ‘Мы увидели рыбу в воде’; *Эрес Долаанадан карак салбайн олурган* ‘Эрес не сводил глаз с Долааны’; *Кырган-ачам ыракта* ***дагларже*** *көрүп олурган* ‘Мой дедушка смотрел на горы вдалеке’.
	Объект социального контакта*	***эжинге*** *ужурашкан* ‘встретил друга’; ***кырганнарга*** *дузалашканнар* ‘помогли старикам’.
	Объект отрицательного	***Менден*** *дезип турар апарган* ‘Он стал меня

	социального контакта*	избегать;’ *Миша* ***эжинден*** *яблогун харамнанган* ‘Миша пожадничал для друга яблоко’.
	Объект интеллектуального избегания*	*Бодунуң* ***өчүүнден*** *ойталап эгелээн* ‘он начал отказываться от своих показаний’.
	Объект адаптации*	*Ыт чаа* ***ээлеринге*** *өөрени берген* ‘Собака привыкла к новым хозяевам’.
Адресат	Реципиент	*Миша конфеталарны* ***эштеринге*** *улеп берген* ‘Миша раздал друзьям конфеты’; ***кижиже*** *даңза сунар* ‘подносят человеку трубку’.
	Адресат	*Башкы* ***өөреникчилерге*** *чаа тема тайылбырлаан* ‘Учительница объяснила новую темы ученикам’; *Саида кежээ* ***авазынче*** *долгаар болган* ‘Саида вечером должна позвонить своей маме’.
	Бенефактив	*Чазак* ***хойжуларга*** *чаа кыштаглар тудуп берген* ‘Правительство чабанам построило новые зимовья.’
Источник	Источник получения предмета*	*Бо бичии аптараны хем* ***кыдыындан*** *тып алдым* ‘Я нашел этот маленький сундук на берегу реки’; ***Акымдан*** *телефонну ап алдым* ‘Я взял у своего брата телефон.’
	Источник получения информации*	***Эжиңден*** *айтырып ал!* ‘Спроси у своего друга!’; *Медээни* ***солуннардан*** *номчаан* ‘вычитал новость из газет’.
	Источник получения навыков*	*Танцылаарынга Оля* ***Катядан*** *өөренип алган* ‘Танцевать научилась у Оли’
	Источник негативного социального контакта*	***Менче*** *халдап, шурап чүзүл?* ‘почему на меня нападает?’; *Болат-оол тенек* ***оолдардан*** *бисти камгалап турду* ‘Болат-оол защищал нас от озорных мальчиков.
	источник негативного физического воздействия*	*Хөй чигирзиг* ***чемден*** *шеглээр болза эки* ‘Лучше будет воздержаться от сладкого’; *Сигенни* ***чаашкындан*** *чаглактап каан* ‘Укрыли сено от дождя’.
Инструмент	Собственно инструмент*	*Грядкаларны* ***лейкадан*** *суггарып кал.* ‘Полей грядки из лейки.’ *адыгны* ***боо-биле*** *адар* ‘стрелять в медведя из ружья’
	Случайный инструмент *(нежелательного*	*Салаазын* ***бижекке*** *кезип алган* ‘он поранил палец ножом’

	*действия)**	
	Мобилитив	*Оглунуң **машиназынга** аай-дедир халдып чоруп турган* ‘Катался на машине своего сына’
	Ингредиентное средство*	*Дыка хөй **эътти** чипкен* ‘съел очень много мяса’ *Оглуң хөй **конфет-чигирге** идээлээн эвеспе оң?* ‘Может быть, твой сын переел сладостей?’
	Фабрикатив	*Шиви **дазылындан** сыын чазап каан* ‘Выстрогал оленя из корня ели’.
Ситуант	Посредник	*Акымга акшаны ачам **чолаачыдан** чорудупкан* ‘Мой отец отправил деньги брату через шофера’.
	Эталон сравнения	*Ажылдап турар черим **негелдеге** дүүшпес* ‘Мое рабочее место не соответствует требованию’; *Өске **оолдардан** бис канчап дорайтаар бис* ‘Чем мы хуже других мальчиков’; *Дилгиден арай улуг* ‘чуть больше лисы’
	Комплетив	*Шыырак өөреникчилерни **класстан** шилээш…* ‘Выбрав лучших учеников из класса…’; *Бопуй-оол суурнуң **оолдарындан** кежээзи-биле онзаланып аңгыланып турган* ‘Бопуй-оол выделялся среди парней села трудолюбием.’
	Деструктив	*Оолдар будуктарны **ыяштардан** сыйып алган.* ‘Мальчики оторвали ветки от деревьев.’
	Композитив	*Шериг кезээ хая-даш эвес, дириг **кижилерден** тургустунган* ‘Военная часть состоит не из камней, а из живых людей.’
	Партитив	*Ытка **хлебтен** кезип бер.* ‘Отрежь собаке хлеба.’
	Часть тела*	*Ыт оолдуң **будундан** ызырыпкан.* ‘Собака укусила мальчика за ногу.’; *Буянның **иштинче** теп, **арнынче** шанчып тургаш…* ‘Пиная в живот, ударяя по лицу Буяна…’; *Авазы Гуляны **хавааанче** чыттап каан* ‘Мама Гули поцеловала ее в лоб’.
	Часть временного целого *(удаляемый объект)**	*Миша чудуктарны **чугундан** адырып каапкан.*‘Миша оскоблил бревна от смолы.’
	Дистрибутив	*Башкы өөреникчилерге **номнарны** үштеп берген.* ‘Учитель ученикам дал по три книги.’

Класс адъектов представлен следующим набором ролей: локатив (место), директив-старт (исходная точка), директив-финиш (конечная точка), направление, траектория. Семантическая роль места в качестве адъекта реализуется в предложениях статической локализации при предикатах бытия-местонахождения и ненаправленного движения.

Остальные семантические роли являются обязательными при глаголах движения и перемещения.

Таблица № 2.

СЕМАНТИЧЕСКИЕ РОЛИ АДЪЕКТОВ

Семантические роли	**Примеры**
Локатив (место)	*Мен* ***хоорайда*** *чурттап тур мен* ‘Я живу в городе’; ***Бажыңга*** *мени манап олур* ‘Сиди и жди меня дома’.
Директив-старт	*Бедик* ***черден*** *аңдарылган* ‘Упал с высокого места’; *Катер* ***эриктен*** *улам-на ырап бар чораан* ‘Катер еще больше отдалялся от берега’; *Куш оолдары* ***уязындан*** *ужуп үнүпкеннер* ‘Птенцы улетели из своего гнезда’.
Директив-финиш	*Машина* ***чанынче*** *халып чеде бер* ‘Подбеги к машине’; *Эрткеш,* ***сандайга*** *оожум олуруп алган* ‘Она прошла и тихо села на стул’.
Направление	*Башкы бөлүк уруглар-биле* ***эжикче*** *углапкан* ‘Учитель с группой детей направилась к двери’; *Ол* ***аргаже*** *чааскаан агаарлаан* ‘Он один направился в лес прогуляться’.
Траектория	*Мен ол орук-****биле*** *чедип келдим* ‘Я пришел той дорогой.’; *Арганың ишти-****биле*** *чоруптаалыңар.* ‘Давайте пойдем лесом.’; ***Соңгадан*** *өдүп кирип кел* ‘Пролезай через окно!’

Если актант является непосредственным участником ситуации, то *сирконстант* несет дополнительную обстоятельственную характеристику предиката. Его функциональное назначение состоит в том, чтобы дополнять, уточнять глагольное действие. Среди сирконстантов различают пространственные, временны́е и каузальные [Всеволодова 2000: 148−152].

К пространственным сирконстантам может относиться только *локатив* (см. выше), который включается как в класс адъектов, так и в класс сирконстантов, в зависимости от класса управляющего предиката.

Таблица № 3.

СЕМАНТИЧЕСКИЕ РОЛИ СИРКОНСТАНТОВ

Подкласс	Семантические роли	Примеры
Пространственные	Локатив	***Дагларда*** *хар чаапты* ‘В горах выпал снег’
Временные	Темпоратив	*Кежээ сес* ***шакта*** *хурал болур* ‘Вечером в восемь часов будет собрание’
	Время, занятое временем	*Ийи* ***хонук ишти*** *ажылдадым* ‘я работал два дня’; *Бо ажыл* ***беш хонукка*** *четчир* ‘Этой работы хватит на пять дней’
	Точка отсчета во времени*	*Караңгы* ***эртенден*** *караңгы дүнге чедир ажылдаар* ‘Работать с раннего утра до глубокой ночи’
	Срок выполнения действия	***Күске чедир*** *тудугну доозар херек* ‘Надо закончить стройку к осени’
Каузальные	Стимул *(эмоции, отношения, поведения)*	*Ынчалза-даа ол* ***Кимге*** *бүзүрээн* ‘Несмотря на это она поверила Киму’; *Ак-Төш бир-ле* ***чүвеге*** *хомудаан.* ‘Ак-Тош на что-то обиделся.’; *Ээзи* ***ыдынче*** *хорадап алгырган* ‘Хозяин в гневе накричал на свою собаку’; *Бөрү* ***оттан*** *коргар* ‘Волк боится огня’; *Ол бодунуң* ***күжүнге****-даа,* ***туразынга****-даа менээргенмейн чораан* ‘Он не зазнавался своей силой, своей волей’.
	Причина	*Кижилер* ***аарыгдан*** *өлүп турган.* ‘Люди умирали от болезни.’
	Ситуатив	*Алешаның арны* ***соокка*** *хорлай берген.* ‘Лицо Алеши на морозе потрескалось.’
	Мотив	*Буян караңгыда кижини* ***үнүнден*** *танып каан* ‘По голосу человека в темноте Буян узнал

Таким образом, для составления инвентаря семантических ролей для семантической разметки Электронного корпуса текстов тувинского языка были использованы понятия классов семантических функций: адъекты, актанты и сирконстанты. Классы актантов и сирконстантов представлены подклассами, каждый из которых включает по несколько конкретных семантических ролей. Данный инвентарь семантических ролей построен в виде иерархии с классами и подклассами для создания инструмента поиска и кластеризации. Список ролей можно свести к основным базовым ролям или расширить до нескольких десятков ролей в зависимости от валентности предиката.

ЛИТЕРАТУРА:

1. Всеволодова М.В. Теория функционально-коммуникативного синтаксиса. М., 2000. С. 141–148.
2. Лыскова Н.А. Семантика падежа в обско-угорских языках. СПб: РГПУ, 2003. С. 105−121.
3. Плунгян В. А. Общая морфология. М, 2000. С. 165–166.
4. Теньер Л. Основы структурного синтаксиса. М., 1988. С. 250.
5. Тестелец Я.Г. Введение в общий синтаксис. М., 2001. С. 188–189.
6. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. М., 1981. С. 168−370.
7. Черемисина М.И., Озонова А.А. Синтаксис тюркских языков Южной Сибири. Простое предложение. Новосибирск, 2005. С. 150−152.
8. Черемисина М. И., Скрибник Е.К. О системе моделей элементарных простых предложений в языках Сибири // Гуманитарные науки в Сибири. Новосибирск, 1996. Вып. 3. С. 50.
9. Lazard G. Actancy. Berlin, NewYork: Mouton de Gruyter. 1998.

ИССЛЕДОВАНИЕ И СОЗДАНИЕ РАЗМЕЧЕННОГО КОРПУСА ТЕКСТОВ ДЛЯ КАЗАХСКОГО ЯЗЫКА

Нурхан[1] А.К., Рахимова[2] Д.Р., [1] Казахский национальный университет имени Ал-Фараби, Алматы, Казахстан, [2] Институт информационных и вычислительных технологий, Алматы, Казахстан,
diana.rakhimova@kaznu.kz

Растет согласие на то, что значительный, быстрый прогресс может быть достигнут, а также в текстовой форме и на устном языке, изучив автоматическое извлечение информации о языке из очень больших корпусов. Такие корпорации начинают служить важными инструментами исследования в области обработки естественного языка, признания речи и интегрированных систем разговорного языка и теоретической лингвистики. Аннотированный корпус обещает быть ценным для предприятий, столь разнообразным, как автоматическое составление статистических моделей для грамматики письменного и разговорного языка, разработка очевидных формальных теорий разных грамматик письма и речи, исследование просодических явлений в речах, а также оценки и сопоставление соответствия синтаксического анализа моделей. В этой статье рассматривается создание одного такого аннотированного случая. В этой статье рассмотрены проблемы и решения по созданию аннотированного текстового корпуса для казахского языка. Правильно собранный веб-корпус потенциально имеет набор приложений. Парадигма «сеть как корпус» с ее логическим продолжением «сеть как набор для обучения» породила широкий спектр возможностей для разработчиков в области обработки естественного языка и вычислительных лингвистов, которые, тем не менее, часто должны собирать все необходимые массивы Интернет-текстов самостоятельно. Статья дает ответы на такие вопросы, как создание корпуса из космоса во всемирной паутине, сравнение существующих инструментов и объяснение выбора одного из них. Кроме того, выбор стандарта аннотации универсальных зависимостей, выбор инструмента аннотации, который является парсером udpipe. Объяснение аннотирующего корпуса с использованием стандарта UD по сравнению с аннотацией собственным тегом. По крайней мере, для каких дальнейших исследований этот корпус будет использоваться. В этой статье описывается создание QWaC (Qazaq Web как Corpus) — большого здания, состоящего из нескольких миллионов слов, взятых из Интернета на казахском языке. По сравнению со многими мировыми языками на казахском языке нет общедоступного объекта поиска. Несмотря на то, что проблема утратила свою актуальность для мировых языков, таких

как английский, немецкий и т. д., для казахского языка это актуально и по сей день из-за закрытого доступа многих корпусов. Создание полезного учебного корпуса требует постоянного пополнения коллекции текстов. По этой причине во время исследования было принято решение о написании программы, которая автоматически загружает сразу несколько статей с установленного сайта, а также мгновенно обрабатывать и маркировать их. Существует множество инструментов для реализации этой задачи. Вы можете использовать готовое программное обеспечение, которое очень много в интернете, путем поиска казахского языка, вы также можете писать сценарии в Python. В данной работе представлены некоторые сценарии и использование подходящих библиотек.

***Ключевые слова**: машинное обучение; текстовые корпуса; обработка естественного языка; аннотированный корпус; POS-теги; UDPipe;*

RESEARCH AND CREATION OF A MARKED TEXT FOR THE KAZAKH LANGUAGE

Nurkhan[1] A.K., Rakhimova[2] D., [1]Kazakh national university named after Al-Faroby, Almati, Kazakhstan[2]; Institute of information and Institute of Information and Computational Technologies, Almaty, Republic of Kazakhstan diana.rakhimova@kaznu.kz

There is a growing consent that considerable, fast progresses can be made and in text form and in a spoken language, having investigated by automatic extraction of information on language from very big corpora. Such corpora begin to serve as important tools of a research for investigators in natural language processing, recognition of the speech and the integrated systems of a spoken language and in theoretical linguistics. The annotated corpus promises to be valuable to the enterprises, so various as automatic drawing up statistical models for grammar of a written and spoken language, development of obvious formal theories of different grammars of writing and the speech, investigation of prosodic of the phenomena in speeches, and estimates and comparisons of compliance of parsing of models. In this article we consider creation of one such annotated case. In this article considered problems and solutions of creating an annotated text corpus for Kazakh language. Correctly assembled web corpus potentially has a set of applications. The paradigm "a web as the corpus" with her logical continuation "a web as the training set" has generated a wide range of possibilities for developers in the field of natural language processing and computational linguists who, nevertheless, often should collect all necessary arrays of Internet texts independently. Article provides answers to such questions as creation of corpus from space the world wide web, comparison of the existing tools and explanation of the choice of one of them. Also, choice of the Universal

dependencies' annotation standard, choice of the annotating tool that is udpipe parser. Explanation of annotating corpus using UD standard in comparison with annotating by own tagset. At least, for what further researches this corpus will be used. This article describes the creation of QWaC (Qazaq Web as Corpus), a large building consisting of several million words, taken from the Internet in the Kazakh language. Compared with many world languages, the Kazakh language does not have a public, accessible search facility. Even though the problem has outlived its relevance for world languages such as, English, German, etc., for the Kazakh language it is relevant to this day, due to the closed access of many buildings. Creation of the useful training case requires constant replenishment of a collection of texts. For this reason, during the research the decision on writing of the program which automatically will load several articles from the set website at once and also to instantly process and mark them has been made. There are many tools for implementing this task. You can use ready-made software, which is very much in the network, by searching for the Kazakh language, you can also write scripts in Python. This article justifies the choice of writing a script and the use of suitable libraries.

Key words: *Machine learning; text corpora; NLP; annotated corpus; POS-tagging; UDPipe;*

Веб как корпус

В последнее десятилетие проекты корпусов, объем которых превышает миллиард слов, быстро развиваются. Для казахского языка существуют такие проекты, как, Корпус Казахского Языка, АККЯ (Алматинский корпус казахского языка), Национальный Корпус Казахского Языка, Kazcorpus. Несмотря на то, что объем этих корпусов является их сильной стороной, составляющие материалы корпуса недоступны для самостоятельной разработки или независимого анализа, даже если они обеспечивают доступ к поиску по нему. Таким образом, задача корпуса для многих языков стран Средней Азии, в том числе и для казахского языка образует собой «черный ящик». Способ использования только открытых данных может стать решением данной задачи. В этом случае каждый пользователь отвечает за интерпретацию результатов поиска, но может выполнять любой анализ на них. Единственный момент в том, что эти данные должны быть открытыми и достаточными для представления определенной части языка. Основываясь на проблемах репрезентативности в веб-корпусах и больших лингвистических данных, мы можем переформулировать его на основе статистической идеи «язык — это большой набор редких событий» и достаточный объем данных по каждому неязыковому событию. Подводя итоги, мы можем выделить 3 современных подхода к набору больших веб-корпусов:

- классический — обход каждого ресурса с использованием сканеров, которые обращаются к поисковым машинам и просматривают страницы; и после того, как материал очищается от спама и дедуплицируется. Такой подход не позволяет сэкономить максимум метатекстовой информации, так как все шаблоны удалены, но это позволяет собирать много за короткое время (пример — Common Crawl).
- монтированный — все материалы из перечисленных тысяч URL-адресов сканируются. Иногда используется подходящая функция, которая декодирует, в то время как URL-адрес подходит или нет, в то время как искатель обращается к поисковым системам, как в первом подходе (например, Aranea corpora).
- дифференциальный — выполняется сканирование небольшого количества больших ресурсов, но они загружаются как можно полнее, полностью, если это возможно. Эта загрузка позволяет нам заявить, что лингвистическая вариация ресурса полностью покрыта.

Для использования корпуса текстов как инструмента для обработки языка и решать на его основе инженерные задачи разной сложности, корпус, во-первых, должен иметь открытый исходный код. Это позволит использовать материал не только для своих специфичных задач, но и позволит другим пользователям модифицировать и дополнять его. К примеру, пользователи смогут по своему усмотрению разработать неплохую векторную модель по данным. Еще одним преимуществом является то, что корпус можно пополнять из открытых источников, которых много в пространстве интернета, что поможет облегчить такую объемную задачу, как сбор тела корпуса. Вторым не менее важным фактором является объем данных. Чем больше данных, тем выше качество анализа. Также, это позволит собрать больше информации о словоформах, редких словах, омонимах и неологизмах. В-третьих, должны быть минимизированы ошибки в корпусе, что поможет качественной разметке корпуса и сохранять баланс исследования. Стоит отметить, что немаловажно отделять метаданными отдельные ресурсы, их лингвистические особенности, а также принять к сведению возрастные и социальные категории автора текста.

1.1. Алгоритм

Для реализации данной задачи можно использовать два известных подхода:

- Использование специальных программных обеспечений
- Написание скрипта, обрабатывающего тексты в веб-пространстве

В первом случае, существует 111 программных обеспечений, из которых 56% является платными (такие как Sketch Engine), 67% требуют значительной доработки в чужом программном коде, так как у каждого языка свои особенности, и более 60% предназначены для решения

определенной задачи обработки языка, например, только POS-тэгирование, только морфологическая разметка, только парсинг и т.д. Стоит заметить, что в Sketch Engine уже есть корпус казахского языка, который идет в составе с остальными языками тюркского семейства (Turkic corpora from the web) состоящий из 139 миллионов слов и обновленный в последний раз в январе 2012 года. К тому же, в Sketch Engine есть инструменты для выявления и анализа совпадений, синонимов и антонимов, примеры использования в контексте, ключевые слова или термины, там можно генерировать списки частотных слов казахского однословного или многословного выражения различных типов.

К сожалению, платформа Sketch Engine является платной, поэтому выбор пал на более бюджетный вариант. Было решено использовать программирование на языке Python для обработки естественного языка. Пакет BeautifulSoup был выбран для быстрого скрипирования страницы.

Алгоритм получения тестовых данных приведен на рисунке 1.

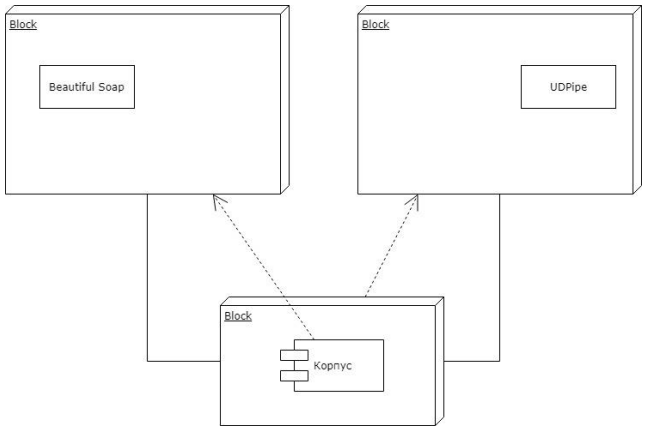
Рис 1.

АЛГОРИТМ GRABBING-А ТЕКСТА С WEB

Корпус хранится в текстовом формате, закодированном в UTF-8, со всеми соответствующими мета-информационными тегами, дублируемыми как база данных sqlite. Для каждого текста структура отступов и абзацев сохраняется как в источнике. Все тексты из каждого источника были отдельно дедуплицированы по URL-адресу, а также отфильтрованы для символов, отличных от UTF, html-тегов, неразрывных пробелов и т.д. Каждый текст может быть найден как в обычном тексте, так и с морфологической и синтаксической аннотацией, отмеченной парсером Udpipe как показано на рисунке 2.

Рис 2.

СОСТАВЛЯЮЩИЕ ИНСТРУМЕНТЫ КОРПУСА



UDPipe — обучающий аннотационный конвейер, маркировка, выведение и разбор файлов формата CoNLL-U. UDPipe является языковым агностиком и может обучать только аннотированные данные в формате CoNLL-U. (Тем не менее, для подготовки функции SpaceAfter должен быть, по крайней мере, некоторым простым текстом, также морфологическим анализатором и lemmatizer можно улучшить, если получить морфологический словарь.) Обученные модели доступны почти для всех деревьев дерева, доступных как UDPipe UD, двоичный файл как библиотека для C ++, Python, Perl, Java, C # и как веб-сервис. В настоящее время пакет позволяет подгонять модель текстовой аннотации, используя функцию udpipe_train . Входными данными будет вектор символов файлов, которые находятся в формате CONLL-U.

file_conllu <- system.file(package = "udpipe", "dummydata", "traindata.conllu") file_conllulibrary (udpipe) m <- udpipe_train(file = "Kazmodel.udpipe", files_conllu_training = file_conllu, annotation_tokenizer = list(dimension = 16, epochs = 1, batch_size = 100, dropout = 0.7), annotation_tagger = list(iterations = 1, models = 1, provide_xpostag = 1, provide_lemma = 0, provide_feats = 0), annotation_parser = "none")

Training tokenizer with the following options: tokenize_url =1, allow_spaces =0, dimension =16 epochs =1, batch_size =100, learning_rate =0.0050, dropout =0.7000, early_stopping =0 Epoch 1, logprob: -2.1721e+005, training acc: 84.20% Tagger model 1 columns: lemma use =1/provide=0, xpostag use =1/provide=1, feats use =1/provide=0 Creating morphological dictionary for tagger model 1. Tagger model 1 dictionary options: max_form_analyses =0, custom dictionary_file =none Tagger model 1 guesser options: suffix_rules =8, prefixes_max =0, prefix_min_count =10, enrich_dictionary =6 Tagger model 1 options: iterations =1, early_stopping =0, templates =tagger Training tagger model 1. Iteration 1: done, accuracy 44.44% m$file_model [1] "Kazmodel.udpipe"

The model is now trained and saved in file toymodel.udpipe in the current working directory ## Now we can use the model to annotate some text mymodel <- udpipe_load_model("Kazmodel.udpipe") x <- udpipe_annotate(object = mymodel, x = "Dit is een tokenizer met POS tagging, zonder lemmatisation noch laat deze dependency parsing toe.", parser = "none") str(as.data.frame(x))

Пример аннотированного текста на UDpipe:

sent_id = akorda-random.tagged.txt:54:936

text = Түстен кейін Президент Павлодар ауданындағы емхананы аралады.

1	Түстен	түс	NOUN	n	Case=Abl	7	obl		
2	кейін	кейін	ADP	post	_	1	case	_	_
3	Президент	президент	NOUN	n	Case=Nom	7	nsubj	_	_
4	Павлодар	Павлодар	PROPN	np	Case=Nom	5	nmod:poss		
5	ауданындағы	аудан	NOUN	n	Case=Loc|Number[psor]=Plur,Sing|Person[psor]=3	6	amod	_	_
6	емхананы	емхана	NOUN	n	Case=Acc	7	obj	_	_
7	аралады	арала	VERB	v	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root	_	SpaceAfter=No
8	.	.	PUNCT	sent	_	7	punct	_	_

К настоящему времени наш корпус содержит данные из сайтов, 13 источников и составляет 1,2 млн. слов или 1,3 млн. токенов, все документируют современный казахский язык. Все источники условно разделены на 5 сегментов: новости, газеты/журналы/радио, поэзия, субтитры, социальные сети, как показано в таблице 1 для распределения данных между сегментами.

Таблица 1.

СЕГМЕНТЫ КОРПУСА

Сегменты	Количество токенов (млн)	Процент (%)
Новости	0.43	33.07%
Медиа	0.62	46.69%
Поэзия	0.24	18.17%
Субтитры	0.012	0.92%
Социальные сети	0.015	1.15%
Итого	1.3	100.00%

Заключение

Мы надеемся, что наша работа будет полезна для обработки казахского языка и поможет разработать новые инструменты и проекты. В ближайшем будущем основными задачами являются создание сообщества пользователей и получение отзывов, отчетов об ошибках, предложений для новых сегментов и т.д. Новая цель — увеличить объем нашего корпуса до 10 миллионов слов за счет других ресурсов и предоставить нашим пользователям больше наборов данных для простой подготовки моделей, тестирования, организации треков и т.д.

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:

1. Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S., (2013) Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In Web as Corpus Workshop (WAC-8).
2. Yih W., Goodman J., Carvalho V. R. Finding advertising keywords on web pages //Proceedings of the 15th international conference on World Wide Web. – ACM, 2006. – С. 213-222.
3. Straka M., Hajic J., Straková J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing //LREC. – 2016.
4. Straka M., Straková J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe //Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. – 2017. – С. 88-99.
5. Kilgarriff A., Grefenstette G. Web as corpus //Proceedings of Corpus Linguistics 2001. – Corpus Linguistics. Readings in a Widening Discipline, 2001. – С. 342-344.
6. Liu V., Curran J. R. Web text corpus for natural language processing //11th Conference of the European Chapter of the Association for Computational Linguistics. – 2006.
7. Guevara E. NoWaC: a large web-based corpus for Norwegian //Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop. – Association for Computational Linguistics, 2010. – С. 1-7.
8. Srdanovic I. et al. A web corpus and word sketches for Japanese //Information and Media Technologies. – 2008. – Т. 3. – №. 3. – С. 529-551.
9. Kilgarriff A., Grefenstette G. Introduction to the special issue on the web as corpus //Computational linguistics. – 2003. – Т. 29. – №. 3. – С. 333-347.
10. Shavrina T., HSE N. R. U. Differential Approach to Web-Corpus Construction. – 2018.

УДК 81’32, 811.512.145

РАЗРЕШЕНИЕ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ В КОРПУСЕ ТАТАРСКОГО ЯЗЫКА НА ОСНОВЕ СТАТИСТИКО-ВЕРОЯТНОСТНОЙ МОДЕЛИ PUREPOS И НЕЙРОСЕТЕВОЙ МОДЕЛИ LSTM

Р. Р. Гатауллин, Р. А. Гильмуллин, Б. Э. Хакимов, Академия наук Республики Татарстан, Казанский федеральный университет, Казань, Россия,
ramil.gata@gmail.com, rinatgilmullin@gmail.com, khakeem@yandex.ru

В работе представлены результаты исследований, посвященных разрешению морфологической многозначности в национальном корпусе татарского языка «Туган тел». В качестве моделей для разрешения многозначности выбраны нейросетевая модель на основе LSTM и статистико-вероятностная модель на основе скрытых Марковских моделей Purepos. В качестве данных для обучения использовался размеченный общественно-политический подкорпус национального корпуса татарского языка «Туган тел» со снятой морфологической многозначностью объемом 2.4M лексических единиц. Эксперименты показали, что модели PurePos и LSTM языкнезависима и дает достаточно высокие результаты и для татарского языка, которые сопоставимы с результатами для других агглютинативных языков, таких как венгерский и турецкий языки.

Ключевые слова: *морфологическая многозначность, татарский язык, Татарский национальный корпус, нейросетевая модель, LSTM, PurePos*

MORPHOLOGICAL DISAMBIGUATION IN THE NATIONAL CORPUS OF TATAR LANGUAGE USING PUREPOS AND LSTM MODELS

R. R. Gataullin, R. A. Gilmullin, B. E. Khakimov, Academy of Sciences of the Republic of Tatarstan, Kazan Federal University, Kazan, Russia,
ramil.gata@gmail.com, rinatgilmullin@gmail.com, khakeem@yandex.ru

This paper presents the results of experiments on morphological disambiguation in the national corpus of the Tatar language «Tugan tel». The experiments were conducted using two technologies. One is PurePos, an open-source HMM-based automatic morphological annotation tool. Other is LSTM based neural network model. As training data tagged socio-political sub-corpus of the National corpus of the Tatar language «Tugan tel» with a volume of 2,4

million lexical units was used. Experiments have shown that the PurePos and LSTM models are language-independent and can be applied to Tatar language too. The results for the Tatar language are on comparable level with the results for other agglutinative languages, such as Hungarian and Turkish.

Key words: *morphological disambiguation, Tatar language, Tatar National Corpus, corpus data, morphological tagging, PurePos, LSTM.*

1. ВВЕДЕНИЕ

Разрешение многозначности является одной из основных задач автоматической обработки естественного языка. Результаты разрешения могут использоваться для повышения точности и улучшения качества применяемых методов в таких задачах как классификация и кластеризации текстов, машинный перевод, информационный поиск.

Сложность и особенности разрешения многозначности для каждого конкретного языка проявляются по-разному. Например, для английского языка с бедной морфологией и жестким порядком слов в предложении разрешение морфологической многозначности, как правило, сводится к задаче POS-теггинга и решается применением достаточно простых методов. Для русского языка морфологическая многозначность не столь характерна, как для английского и татарского, но, тем не менее, присуща. Дополнительную сложность добавляет свободный порядок слов в русском языке. В татарском языке, как и в других агглютинативных языках тюркской группы морфемы являются важнейшими значащими языковыми единицами, которые несут как семантическую, так и синтаксическую информацию. Имея теоретически неограниченное количество присоединяемых к основе морфем, морфологическая многозначность приобретает разнообразные формы, что значительно усложняет задачу разрешения.

К настоящему времени сформирована основная парадигма методов снятия многозначности, которая включает методы, основанные на правилах; методы машинного обучения, использующие вероятностные модели; гибридные методы *(Гатауллин Р.Р., 2016; 4.Gataullin R., 2017)*. Создание электронного корпуса татарского языка «Туган тел» *(http://tugantel.tatar/)* и общественно-политического подкорпуса со снятой вручную морфологической многозначностью данных дали возможность исследования данной задачи с применением статистических методов на основе технологий машинного обучения *(Хакимов Б.Э. и др., 2014; Гатауллин Р.Р. и др., 2016; Гильмуллин Р.А. и др., 2017)*.

Анализ открытых программных кодов, разработанных для этой задачи в последние несколько лет, показал, что одними из эффективных являются инструментарий PurePos 2.0 *(Orosz, G. and Novák, 2013)*, реализующая гибридную модель на основе скрытых Марковских моделей, а также

нейросетевая модель на основе рекуррентных нейросетей с долгой краткосрочной памятью LSTM *(Qinlan Shen and other, 2016)*. Скрытая Марковская модель − модель процесса, в которой процесс считается Марковским, причем неизвестно, в каком состоянии находится система (состояния скрыты), но каждое состояние может с некоторой вероятностью произвести событие, которое можно наблюдать. Другими словами, изучается Марковский процесс с неизвестными параметрами, и задачей является распознавание неизвестных параметров на основе наблюдаемых. Инструмент PurePos разрабатывался для языков со сложной морфологией, в том числе для агглютинативных языков, а также для языков с малой ресурсной базой guages). Основные эксперименты проводились для венгерского яхыка (*Orosz, G., Novák, 2012*; *Orosz, G. and Novák, A., 2013*). Результаты по распознаванию POS-тегов слов показали точность 97%. Авторами было выдвинуто предположение, что модель способна работать и для других языков, в том числе для языков тюркской группы.

Эксперименты с нейросетевой моделью на основе рекуррентной нейронной сети с долгой краткосрочной памятью (англ., Long short-term memory, LSTM) описываются в работе *(Qinlan Shen and other, 2016)*. Для обучения модели требуются размеченные тексты со снятой многозначностью. Идея метода сводится к тому, что каждому разбору многозначного слова и окружающему его контексту сопоставляются вектора. В первом случае вектор строится на основе его леммы и морфологических признаков, во-втором на основе поверхностных форм окружающих слов; дополнительно вектор можно расширить и за счет морфологических признаков. При этом контекст не ограничивается несколькими словами непосредственной близости слов и может достигать размеров всего предложения. После этого на основе полученной пары векторов, строится распределение условных вероятностей, из которых выбирается наиболее вероятный разбор в качестве правильного.

Авторы также отмечают, что модель языканезависима. В качестве демонстрации этого утверждения модель применили для разрешения морфологической многозначности в турецком, русском и арабском языках. Результаты LSTM практически близки к результатам других методов разрешения, а в некоторых случаях и превосходят, например, для турецкого языка, как показано в таблице 1, авторы добились результата в 96,41% точности разрешения.

Таблица 1.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ПРИМЕНЕНИЯ НЕЙРОСЕТЕВОЙ АРХИТЕКТУРЫ LSTM ДЛЯ РАЗРЕШЕНИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ

Язык	Турецкий		Русский		Арабский	
	% от много-значных слов	% от всех токенов	% от много-значных слов	% от всех токенов	% от много-значных слов	% от всех токенов
Без контекста (baseline)	88.65	95.45	64.97	88.58	72.22	78.06
Локальный контекст	89.18	95.67	71.56	90.72	80.10	84.29
Все предложение (поверхностная форма)	91.03	96.41	69.49	90.05	86.45	88.95
Left-to-Right	90.50	96.19	68.55	89.75	89.30	91.27
CRF	90.24	96.09	72.78	91.13	-	-

В работе заслуживает внимание и анализ размера используемого контекста – авторы сравнивали разные размеры и типы контекста, и экспериментально выявили наиболее подходящий под каждый язык тип. Оказалось, что для турецкого языка достаточным является построение векторов на основе поверхностных форм слов без явного определения морфологических признаков, но используя все слова в предложении. Тогда как для русского важным моментов является согласованность в роде, числе и падеже, что в свою очередь требует наличия не только поверхностной формы слова, но и морфологических признаков слов в контексте. При этом добиться лучших результатов (точность разрешения 91.13%) помогает оптимизация с помощью метода условных случайных полей (англ., Conditional Random, CRF). Похожая ситуация и с арабским языком, когда поверхностных форм слов недостаточно для полного разрешения многозначности. Это можно объяснить тем, что в арабском доля многозначности больше, чем турецком. Если, например, в турецком языке в среднем на одно слова приходится 2,81 вариантов разбора, в русском языке – 5,81, то в арабском языке – 11,31. Поэтому для правильного обучения модели требуется размеченный контекст с полностью снятой омонимией.

2. ПОДГОТОВКА ДАННЫХ

На начальном этапе работы из базы текстов национального корпуса татарского языка «Туган тел» *(Хакимов и др., 2014)* были получены статистические данные о частотности словоформ с альтернативными разборами, приведенные в таблице 2. Для морфологической разметки

корпуса используется морфологический модуль, реализованный на основе инструментария HFST *(Gilmullin R. and Gataullin R., 2017)*.

Таблица 2.

РАСПРЕДЕЛЕНИЕ ВАРИАНТОВ МОРФОЛОГИЧЕСКОГО РАЗБОРА

Варианты разборов	Количество	Доля в корпусе
Всего словоформ с альтернативными разборами	5.650.820	25,75%
2 разбора	4.282.108	19,51%
3 разбора	1.045.392	4,76%
4 разбора	296.547	1,35%
5 и более разборов	26.773	0,12%
Всего в выборке	21.940.452	100%

Общий объем корпуса на этом этапе составлял 21.940.452 словоупотреблений, доля словоупотреблений с альтернативными разборами составила 25,75%.

При этом максимальная длина представленной в корпусе словоформы состоит из основы и двенадцати грамматических аффиксов.

Для проведения экспериментов с обучением моделей необходимо создание корпуса со снятой морфологической многозначностью. В качестве данных для обучения использовался морфологически размеченный со снятой вручную морфологической многозначностью общественно-политический подкорпус национального корпуса татарского языка «Туган тел». Статистика подкорпуса приведена в таблице 3.

Таблица 3.

СТАТИСТИКА ОБУЧАЮЩЕЙ И ТЕСТОВОЙ ВЫБОРКИ ПО ОБЩЕСТВЕННО-ПОЛИТИЧЕСКОМУ КОРПУСУ

	Обучающая выборка	Тестовая выборка
Количество контекстов (предложений)	54.580	944
Количество токенов (включая пунктуацию)	600.480	11.655
Количество многозначных разборов	125.480 (21%)	2.527 (21%)
Количество уникальных	29.953	2.788

словоформ		
Количество уникальных лемм	7.117	1.226
Количество уникальных морфологических форм	1.898	346

Ручное снятие морфологической многозначности общественно-политического корпуса выполнялось лингвистами и экспертами с помощью Веб-инструментарий для снятия морфологической многозначности в текстовом корпусе татарского языка *(Гатауллин, 2014)*. По результатам проведенной работы было подготовлено 56.524 предложения со снятой морфологической многозначностью.

3. ЭКСПЕРИМЕНТ

Как видно из таблицы 3, размеченная выборка данных была разделена на выборку для обучения и тестовую выборку. Модели на основе PurePos и LSTM обучались только на обучающей выборке, тестовая выборка использовалась только для тестирования. Каждая модель обучалась на одной и той же обучающей выборке и проходила валидацию на одной и той же тестовой выборке. В таблице 4 приведены оценка точности по нескольким показателям: распознавания леммы, аффиксальной цепочки и разрешения многозначности.

Таблица 4.

ПОКАЗАТЕЛИ ТОЧНОСТИ РАСПОЗНАВАНИЯ ЛЕММ/ И АФФИКСАЛЬНОЙ ЦЕПОЧКИ

Показатели	PurePos 2.0	LSTM NN
Точность распознавания леммы (без морфологической разметки)	11163 / 11655 = 95.77%	не поддерживается
Точность распознавания аффиксальной цепочки (без морфологической разметки)	10889 / 11655 = 93.42%	не поддерживается
Точность распознавания леммы (с морфологической разметкой)	11399 / 11655 = 97.80%	11299 / 11655 = 96.94%
Точность распознавания аффиксальной цепочки (с морфологической разметкой)	11243 / 11655 = 96.46%	11127 / 11655 = 95.46%

Таблица 5.

КОЛИЧЕСТВО ВАРИАНТОВ МОРФОЛОГИЧЕСКОГО РАЗБОРА И ТОЧНОСТЬ РАЗРЕШЕНИЯ МОДЕЛЕЙ

Количество вариантов	PurePos 2.0	LSTM NN
n=2	1613 / 1826 = 88.33%	1545 / 1826 = 84.61 %
n=3	337 / 424 = 79.48%	268 / 424 = 63.21 %

n=4	125 / 192 = 65.10%	141 / 192 = 73.44 %
n=5	7 / 9 = 77.78 %	7 / 9 = 77.78 %
n=6	50 / 72 = 67.57 %	37 / 72 = 51.39 %
n=7	0 / 2 = 0.00 %	0 / 2 = 0.00 %
n=8	0 / 1 = 0.00 %	0 / 1 = 0.00 %
Общее	2132/ 2527 = 84.36%	1999 / 2527 = 79.10%

ЗАКЛЮЧЕНИЕ

В данной работе представлены результаты работ по разрешению морфологической многозначности татарского языка с использованием инструментария PurePos 2.0 и нейросетевой модели на основе LSTM. Учитывая ограниченный набор корпусных данных для обучения, результаты экспериментов показали достаточно хороший уровень точности для разрешения морфологической многозначности 84,36% и 79.10% соответственно. Разрешение с помощью инструментария PurePos 2.0 показало более высокую точность разрешения практически для всех вариантов разбора, кроме n=4. Также PurePos может быть использован для распознавания нераспознанных морфологическим анализатором словоформ с тегом NR (Not Recognized) с точностью 79% для разметки леммы.

По мнению авторов, более низкие показатели точности нейросетевой модели прежде всего связаны с объемом обучаемых данных. Поскольку системы с нейросетями недостаточно эффективны при обучении на ограниченном наборе данных. Полученные результаты будут использоваться в процессе ручного снятия морфологической многозначности, что безусловно позволит повысить производительность труда при создании «золотого» подкорпуса с разрешенной многозначностью.

Мы планируем повышать качество и объем обучаемый данных, на основе которых будут продолжены исследования по повышению точности разрешения морфологической многозначности в корпусе татарского языка «Туган тел», в том числе за счет комбинирования различных методов.

ЛИТЕРАТУРА:

1. Гатауллин, Р.Р. Аналитический обзор методов разрешения морфологической многозначности. / Р.Р. Гатауллин // Российский научный электронный журнал (Электронные библиотеки). Том 19, № 2 (2016). – С. 98-114

2. Gataullin R., Khakimov B., Suleymanov D., Gilmullin R. (2017) Context-Based Rules for Grammatical Disambiguation in the Tatar Language. // N.T. Nguen et al. (Eds). ICCCI 2017, Part II, LNAI 10449, pp. 529-537, 2017

3. Хакимов, Б.Э. Разрешение грамматической многозначности в корпусе

татарского языка / Б.Э.Хакимов, Р.А.Гильмуллин, Р.Р.Гатауллин // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. - 2014. - Т. 156, кн. 5. -С. 236-244
4. Гатауллин, Р.Р. Контекстные правила для разрешения морфологической многозначности в корпусе татарского языка / Р.Р. Гатауллин, Р.А. Гильмуллин // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2016 OpenSemantic Technologies for Intelligent Systems МАТЕРИАЛЫ V МЕЖДУНАРОДНОЙ НАУЧНО-ТЕХНИЧЕСКОЙ КОНФЕРЕНЦИИ (Минск, 18-20 февраля 2016 года), - Минск. : БГУИР, 2016. — С. 389-392
5. Гильмуллин, Р.А. Разрешение морфологической многозначности текстов на татарском языке на основе инструментария PurePos. / Р.А. Гильмуллин, Р.Р. Гатауллин // V МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ «TURKLANG 2017». – Труды конференции. В 2-х томах. Т – Казань: Издательство Академии наук Республики Татарстан, – С. 30-37
6. Orosz, G. and Novák, A. 2013. PurePos 2.0: a hybrid tool for morphological disambiguation. Proceedings of Recent Advances in Natural Language Processing, pages 539–545, Hissar, Bulgaria, 7-13 September 2013. Online version: http://aclweb.org/anthology//R/R13/R13-1071.pdf
7. Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, Chris Dyer. 2016. The Role of Context in Neural Morphological Disambiguation. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 181–191, Osaka, Japan, December 11-17, 2016. http://aclweb.org/anthology/C16-1018
8. Гатауллин, Р.Р. Веб-инструментарий для снятия морфологической многозначности в текстовом корпусе татарского языка / Р.Р. Гатауллин // Сохранение и развитие родных языков в условиях многонационального государства: проблемы и перспективы: материалы V Международной научно-практической конференции (Казань, 19-22 ноября 2014 г.). – Казань: Отечество, — С. 71-73
9. Gilmullin R., Gataullin R. (2017) Morphological Analysis System of the Tatar Language // N.T. Nguen et al. (Eds). ICCCI 2017, Part II, LNAI 10449, pp. 519-528, 2017

НАЦИОНАЛЬНЫЙ МНОГОЯЗЫЧНЫЙ КОРПУС ИМЕНИ АБУСУПЬЯНА АКАЕВА: ВОПРОС РЕПРЕЗЕНТАТИВНОСТИ ВЫБОРКИ

Д.А.Темирова, Московский государственный университет имени М.В. Ломоносова, Москва, Россия, dzhannett@bk.ru

В статье дается описание предварительного этапа создания НМК имени Абусупьяна Акаева, который включает отбор материала и его технологическую обработку. В настоящий момент происходит разработка кумыкской части корпуса, пополняющегося лирическими и прозаическими произведениями кумыкских авторов. В будущем развитие получат и другие подкорпуса, в которые, помимо произведений известных авторов, также войдут творческие работы школьников и текстовые сообщения пользователей социальных сетей/блоггеров.

Ключевые слова: *национальный корпус, региональные языки, корпусный подход, кумыкский язык, история, религия.*

NATIONAL MULTILINGUAL CORPUS NAMED AFTER ABUSUPYAN AKAYEV: THE PROBLEM OF SELECTION'S REPRESENTATIVENESS

D.A.Temirova, Lomonosov Moscow State University, Moscow, Russia, dzhannett@bk.ru

The article describes the preliminary stage of the NMC's creation (National Multilingual Corpus named after Abusupyan Akaev), which includes the selection of the texts and their technical processing. At present the Kumyk part of the Corpus is developed, replenishing with Kumyk authors' lyric and prosaic works. Other subcorpora also will be developed, which besides the well-known authors' works, will include the students' creative works and different users'/bloggers' text-messages.

Keyword: *national corpus, regional languages, corpus approach, Kumyk language, history, religion.*

В настоящее время в свете развития информационных технологий, все большее внимание завоевывает раздел лингвистики, «…занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий» [Захаров, Богданова, 2011: 7] и получивший название **корпусная лингвистика.**

На сегодняшний день уже разработано достаточно большое количество национальных корпусов, как в зарубежной, так и в отечественной лингвистике, среди которых наиболее известны НКРЯ, Открытый корпус

русского языка, Британский Национальный Корпус (BNC), Корпус современного американского английского (COCA), Турецкий Национальный Корпус (TNC) и другие.

Однако такой подход в большей степени ориентирован на языки мирового уровня, в то время как региональные языки остаются без должного внимания, обучения и изучения.

В связи с такой ситуацией целесообразным кажется создание корпуса НМК имени Абусупьяна Акаева, объединившего 4 языка тюркской группы: кумыкский, карачаево-балкарский, крымскотатарский , ногайский [3], родство которых получило отражение в стихотворении одного из современных кумыкских поэтов, Магомедмурада Гаджиева (кум. Мугьаммадмурад Гъажилени), «Бирлик»:

Дёрт баш, дёрт къол, дёрт юрек,
Бирлик уьчюн дуньягъа яратылгъан,
Бирлик булан татывлукъда да яшап,
Оьзге тюрлю миллетлеге айтылгъан!
Дёрт гёз, дёрт сан, дёртде жан,
Бирлик булан эл намусларын кютеген,
Оьзге халкълар айры кюйде турса да
Шулай гючлю татывлукъ юрютеген!
Къумукъ, Ногъай, Алан, Къырымлы
Унутмагъыз бирде буланы!
Булар бары бири-бирисин тутгъан
Тутагъандай Ана баланы!
Мени аявлу агъам ва къызашым!
Къаст эт, бирлик булан турмагъа!
Дёртде башгъа гелеген балагьланы
Дёрт къол булан башын тутуп бурмагъа!

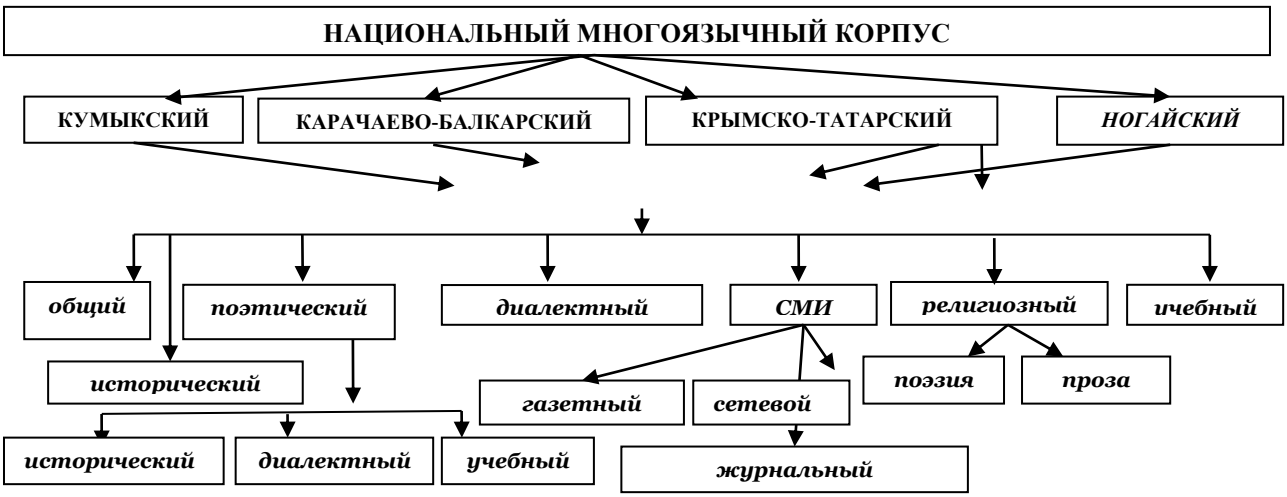
Концептуальная модель НМК имени А.Акаева представляет собой дифференцированную систему подкорпусов, которая показана на макете.

Таким образом, помимо подкорпусов, НМК имени А.Акаева предполагает наличие их модифицированных форм, включающих в себя коллекции текстов в соответствии с указанными тематическими группами.

Одной из первостепенных задач при создании и развитии корпуса является вопрос репрезентативности, т.е. достаточно большой его объем для возможности проведения дальнейших исследований, достоверности полученных данных, а также для получения цельной языковой картины мира. Вследствие этого важно отметить слова Э. Финегана, который утверждает, что «корпус – репрезентативное собрание текстов, обычно в машиночитаемом формате и включающее информацию о ситуации, в

которой текст был произведен, такую как информация о говорящем, авторе, адресате или аудитории».

Ввиду такого понимания проблемы в целом, в октябре 2017 года был начат сбор текстов в кумыкскую часть корпуса имени А. Акаева.



Он включает произведения 7 авторов и содержит их металингвистическую разметку. Так, в **поэтический подкорпус** вошли произведения Ирчи Казака (кум. Йырчы Къазакъ), Анвара Аджиева (кум. Гь. Анвар), Салавата Салаватова (кум. Салават Бойнакълы), Магомедмурада Гаджиева (кум. Мугьаммадмурад Гъажилени), которые объединены концептом «Родина».

Религиозный подкорпус представлен учениями Абусупьяна Акаева («Пайхамарны ёлу булан», «Ислам динни кюрчюлери: Намаз, Ораза, Гьаж Байрамлар»), чьи труды «известны широкому кругу читателей благодаря титаническому труду Г.М.-Р Оразаева, который собрал его работы на кумыкском языке и издал с комментариями в транслитерации на кириллице» [4], а также религиозной лирикой Магомедмурада Гаджиева. В дальнейшем планируется продолжение сбора текстового материала, а также проведение творческих работ среди учащихся – школьников для пополнения учебных подкорпусов.

ЛИТЕРАТУРА:

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. Иркутск: ИГЛУ, 2011. 161с.
2. Finegan E. LANGUAGE: its structure and use. – N.Y.: Harcourt Brace College Publishers, 2004
3. http://ojs.ifmo.ru/index.php/IMS/article/view/524
4. http://www.kumyki.ru/pages/people/AbusufyanAkaev.php

УДК81'322.2

ЛИНГВИСТИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЙ ЯЗЫКА САХА

А.Н. Ноговицына, Северо-Восточный федеральный университет имени М.К.Аммосова, Якутск, Россия, erkin2007@mail.ru

В данной статье рассматривается проблема лингвистического аннотирования причастий языка саха. Для аннотирования причастий автором рассмотрены разные подходы глоссирования, при этом соблюдается основное условие аннотирования – совместимость предлагаемых разметок с тэгами размеченных корпусов других тюркских языков.

***Ключевые слова**: корпусная лингвистика, якутский язык, причастия, морфологическая разметка текстов.*

LINGUISTIC ANNOTATION OF THE VERBS VOICE FORMS IN THE SAKHA LANGUAGE

A.N. Nogovitsyna, M. K. Ammosov North-Eastern Federal University, Yakutsk, Russia, erkin2007@mail.ru

The article discusses linguistic annotation of the Sakha participles. To annotate the participles, the author has considered various approaches for glossing, while complying with the main condition for annotation is the compatibility of the proposed markup with tags of the marked corps of other Turkic languages.

***Key words:** corpus linguistics, the Yakut language, participles, linguistic annotation.*

Создание электронных корпусов миноритарных тюркских языков является актуальной задачей современной тюркологии. В ближайшей перспективе в сравнительно-сопоставительных исследованиях тюркских языков будет применен метод автоматического лингвистического анализа, что требует унификации систем грамматической разметки в корпусах тюркских языков. Нами при аннотировании словоизменительных морфологических показателей языка саха выделены 4 типа подхода к морфологической разметке языка саха:

1) использование тэгов Лейпцигской системы глоссирования.

2) заимствование тэгов, употребленных при описании структуры татарской словоформы. Одним из главных условий при аннотировании

языка саха является совместимость предлагаемых разметок с тэгами размеченных корпусов других тюркских языков.

3) использование разметки морфологического анализатора, разработанной для хакасского языка А.В. Дыбо.

4) использование собственной разметки. Проживание носителей якутского языка на географической периферии тюркоязычного мира способствовало развитию грамматических особенностей языка саха, например, появление третичных причастий или поздних отпричастных образований на почве самого якутского языка.

«По составу живые причастия якутского языка делятся на первичные, по происхождению на более древние и имеющие соответствия во многих тюркских языках (к ним относятся –быт/-батах, -ар/-бат, -ыах/-мыах); вторичные, по происхождению более поздние, на состоящие из двух компонентов, образованные во времена сибиро-монгольского единства тюркских языков (к ним относятся –ааччы, -ааччыта суох, — а илик); третичные или поздние отпричастные образования, которые появились на почве самого якутского языка (к ним относятся формы на –ыахтаах, -ымыахтаах, -ыа суохтаах, -ардаах, -баттаах, -быттаах, -батахтаах)» [6, с. 98].

Таблица 1.

КЛАССИФИКАЦИЯ ПРИЧАСТИЙ ЯКУТСКОГО ЯЗЫКА

Виды причас-тий	**Первичные причастия**		**Вторичные причастия**		**Третичные причастия**	
	Положительная форма	Отрицательная форма	Положительная форма	Отрицательная форма	Положительная форма	Отрицательная форма
Причастие прошедшего времени	-быт	-батах			-быттаах	-батахтаах
Причастие настоящего времени	-ар/-ыыр	-бат	-арга/-ыырга	-бакка	-ардаах	-баттаах
Причастие будущего времени	-ыах	-мыах	-ыахха	-мыахха	-ыахтаах	-ымыахтаах -ыа суохтаах
Древнее причастие	-тах	-батах	-тахха	-батахха		
Хабитуальное причастие			-ааччы	-ааччы[та] суох		
Кункативное причастие				-а илик -ыы илик		

ПЕРВИЧНЫЕ ПРИЧАСТИЯ

Причастие на –ар/-бат.

Причастие на *–ар/-ыыр* образует две формы времени – настояще-будущее и совместно с недостаточным глаголом э- прошедшее незаконченное время в системе индикатива. Отрицательной формой причастия на –ар/-ыыр является причастие на –бат с двенадцатью вариантами алломорфов.

«Причастие на –ар в якутском языкознании принято называть причастием настоящего времени (Харитонов Л.Н., 1947. С. 224; Коркина Е.И., 1970. С. 34; ГСЯЛЯ, 1982. С. 228) с оттенками будущего, длительно-настоящего, постоянно происходящего времен» [6, с. 264].

В практике аннотирования причастия настоящего времени в других тюркских языках приняты следующие условные символы:

PrtAuct – турецкий язык;

PCP_PR – татарский язык.

Для разрабатываемой модели аннотирования грамматических (морфологических) категорий языка саха нами для причастия настоящего времени предлагается использование разметки PCP_PR (present participle), она соответствует причастию настоящего времени, аннотированной как PCP_PR в татарском корпусе (форма на –У+чЫ).

«Форма причастия настоящего времени на *–учы* и *–а торган*. Различия между формами относятся к функциональной стороне причастий: форма *–учы* может быть определением лишь субъекта действия, выраженного данной формой: *укучы бала* «читающий ребенок»; форма на *–а торган* может определять субъект, прямой и косвенный объекты, место, время действия, выраженного данным причастием: *укый торган бала* «читающий ребенок», *укый торган китап* – «читаемая книга», *укый торган мәктәп* – «школа, в которой учатся», *укый торган чак* – «время, когда учатся» [5, с. 82].

Таблица 2

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ НАСТОЯЩЕГО ВРЕМЕНИ

Сокращения Tags	**Расшифровка сокращений Full term**	**Название категории Name of the category**	**Алломорфы Allomorphs**	**Морфемы Morphemes**
PCP_PR	present participle	причастие настоящего времени	-ар/ -эр/ -ор/ өр/ -ыыр/-иир/-уур/ — үүр	-Ар -ЫЫр
PCP_PR_NEG	present participle negative	причастие настоящего времени отрицательная	-бат/-бэт/-бот/-бөт -пат/-пэт/-пот/-пөт	-БАт

		форма	-мат/-мэт/-мот/-мөт	

Причастие на –быт/-батах

Причастие на –быт лежит в основе нескольких форм прошедшего времени в системе индикатива:

1. Преждепрошедшее повествовательное время, например, барбытым – ‘(я) ушел’, барбыккыт – ‘(вы) ушли’;

2. Прошедшее результативное время первое, например, барбыппын – ‘(я) оказывается, ушел’, барбатахпыт – ‘(мы) оказывается, не ушли’;

3. Давнопрошедшее время (аналитическая форма), например, барбыт этим – ‘(я) ушел тогда’, барбатах этигит – ‘(вы) тогда еще не ушли’;

4. Прошедшее эпизодическое время (аналитическая форма), например, барбытым баар – ‘(я) был как-то раз’, барбыттара суох – ‘(они) не были ни разу’.

В практике аннотирования причастия настоящего времени в других тюркских языках приняты следующие условные символы:

PrtHab – турецкий язык;

PCP_PS – татарский язык.

Для разрабатываемой модели аннотирования грамматических (морфологических) категорий языка саха нами для причастия прошедшего времени предлагается использование разметки PCP_PS (past participle), она соответствует причастию прошедшего времени, аннотированной как PCP_PS в татарском корпусе (форма на — ГАн).

«Форма на *–ган* выражает процессуальный признак предмета, соотнесенный с планом прошлого и настоящего времени: *сөйгән кызы* «его любимая девушка» и «девушка, которую он любил», *укыган китап* «книга, которую мы читаем (читали)» [5, с. 80].

Таблица 3

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ ПРОШЕДШЕГО ВРЕМЕНИ

Сокращения Tags	**Расшифровка сокращений Full term**	**Название категории Name of the category**	**Алломорфы Allomorphs**	**Морфемы Morphemes**
PCP_PS	past participle	причастие прошедшего времени	- быт/-бит/-бүт/-бут -пыт/-пит/-пүт/-пут -мыт/-мит/-мүт/-мут	-БЫт
PCP_PS_NEG	past participle negative	причастие прошедшего времени	-батах/-бэтэх/ -ботох/-бөтөх -патах/-пэтэх/	-БАтАх

		отрицательная форма	-потох/-пөтөх -матах/-мэтэх/ -мотох/-мөтөх	

Причастие на –ыах/-мыах.

«Причастие на *–ыах* по типу спряжения относится к первичным причастиям и обозначает будущее время» [6, с. 129].

При аннотировании причастия будущего времени в корпусе татарского языка исследователями применена помета PCP_FUT с дополнительной разметкой DEF для категорического будущего и INDF для неопределенного будущего.

Таблица 4

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ БУДУЩЕГО ВРЕМЕНИ В КОРПУСЕ ТАТАРСКОГО ЯЗЫКА

Сокращения Tags	**Расшифровка сокращений Full term**	**Название категории Name of the category**	**Алломорфы Allomorphs**	**Морфемы Morphemes**
PCP_FUT	Future participle	причастие будущего времени	-ыах/-иэх/-уох/-үөх	-ЫАх
PCP_FUT_NEG	Future participle negative	причастие будущего времени отрицательная форма	-мыах/-миэх/-муох/-мүөх	-мЫАх

Кроме того, в практике аннотирования других тюркских языков (турецкий, хакасский языки) не принято выделять причастия по признаку темпоральности. Приняты следующие обозначения:

Таблица 5

ПРАКТИКА АННОТИРОВАНИЯ ПРИЧАСТИЙ В ТУРЕЦКОМ, ХАКАССКОМ ЯЗЫКАХ

Обозначения	**Турецкий язык**	**Обозначения**	**Хакассий язык**
PrtAuct (gUcI)	агентивное причастие		
PrtHab (gAn)	хабитуальное причастие	PrtHab	хабитуальное причастие
PrtAct ((X)glI)	активное	PrtAct	активное причастие

	причастие		
PrtProsp (sXk)	проспективное причастие	PrtProsp	проспективное причастие
PrtProj (gU)	проективное причастие	PrtProj	проективное причастие
PrtNecess (gU.lXk)	причастие необходимого действия	PrtNecess	причастие необходимого следствия
PrtImpf ((X)gmA)	имперфективное причастие	PrtImpf	имперфективное причастие

Причастие на –тах

Образуется путем присоединения к глагольной основе морфемы на *–TAx* и его алломорфов. «Л.Н.Харитонов и Е.И. Убрятова справедливо указывают на то, что форма на *–max* в якутском языке существенно отличается от других причастных форм, что она в значительной степени стала собственно-глагольной формой. Действительно, причастие на *–max* легло в основу двух наклонений в якутском языке: предположительного и второго условного» [6, с. 236].

Поэтому нами предлагается использование пометы PCP_COND (participle conditional).

Таблица 6

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ НА –ТАХ

Сокращения Tags	**Расшифровка сокращений Full term**	**Название категории Name of the category**	**Алломорфы Allomorphs**	**Морфемы Morphemes**
PCP_COND	participle conditional	причастие условное	-тах/-тэх/-тох/-төх -дах/-дэх/-дох/-дөх -лах/-лэх/-лох/-лөх -нах/-нэх/-нох/-нөх	-TAx

ВТОРИЧНЫЕ ПРИЧАСТИЯ

Причастие на –ааччы/-ааччы суох.

«Причастие на *–ааччы* в якутском языке выступает как: 1) как имя действующего лица (деятеля) *суруйааччы* ‘обычно пишущий, писатель’ 2) как финитная форма глагола со значением обычно совершаемого действия *суруйааччы* ‘обычно писал’ 3) редко в качестве определения как причастие *суруйааччы киһи* ‘пишущий человек’» [6, с. 118].

Форма на *–ааччы* нами была рассмотрена для наклонения обычно совершаемого действия. Для аннотирования этого наклонения нами была применена разметка HAB (Habitualis). Похожее обозначение принято А.В. Дыбо для хабитуального причастия хакасского языка – PrtHab. При этом необходимо отметить, что хотя аффикс –ааччы в якутском языке и хакасское причастие на –ааччы/еечі являются параллелями, в якутском языке аффикс –ааччы «полностью сохранило свое глагольное действие и легло в основу наклонения обычности» [2, с. 224], в хакасском языке «рассматривается как форма, утратившая глагольные признаки полностью перешедшая в разряд имен прилагательных и существительных, обозначающих действие в качестве постоянного признака предмета, а также наименование деятеля [2, с. 224].

При аннотировании аффикса причастия регулярно совершаемого действия в татарском корпусе использована помета USIT (Usitative). Под узитативом в современной аспектологии понимается действие «иметь обыкновение [постоянно]» [3, с.298]. В.Дресслер среди континуативных нюансов включал «узитатив – при помощи которого выражается повторяющееся событие, являющееся одновременно способностью, склонностью или привычкой участника» [8, с. 17] С этой точки зрения, нам кажется, что помета USIT наиболее подходит для причастия на –ааччы, которое выступает и «глаголом со значением обычно совершаемого действия», и «как определение». Здесь и далее мы сталкиваемся с таким понятием как омонимичность аффиксов.

Разработчики татарского корпуса А.М. Галиева, Б.Э. Хакимов, А.Р. Гатиатуллин в статье «Метаязык описания структуры татарской словоформы для корпусной грамматической аннотации» при рассмотрении омонимии аффисов отмечают, что «возможны две стратегии обозначения этой формы (*омонимии, прим. автора*) в корпусной аннотации: объединение всех случаев под одной пометой либо присвоение отдельных помет каждому значению. Первый подход предполагает поиск адекватного объединяющего термина, покрывающего все случаи, второй требует контекстного различения значений. Одной из перспективных задач ТТ является поиск решений для каждого подобного случая» [4]. С этой точки зрения, мы будем придерживаться пути, объединяющего оба подхода, с одной стороны, сохраним общность омонимичных аффиксов, аннотировав пометой HAB и аффикс причастия, и аффикс модальности, с другой стороны, дополнительно пометим причастие разметкой PCP. Таким образом, надеемся, что такое решение окажет некоторую помощь в последующем разработчикам корпусной грамматической аннотации языка саха.

Таблица 7

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ ХАБИТУАЛЬНОГО

Сокращения Tags	**Расшифровка сокращений Full term**	**Название категории Name of the category**	**Алломорфы Allomorphs**	**Морфемы Morphemes**
PCP_HAB	participle habitualis	причастие хабитуальное	-ааччы/-ээччи/ -өөччү/ -ооччу/	-ААччЫ
PCP_HAB+ PART_NEG	participle habitualis negative	причастие хабитуальное отрицательная форма	-ааччыта/ -ээччитэ/ -өөччүтэ/ -ооччута суох	-ААччЫтА суох

Для аннотирования отрицательной частицы *суох* предлагаем использование разметки PTCL_NEG от английского negative particle – отрицательная частица.

Причастие на –а илик.

Форма на –а илик также была рассмотрена при аннотировании наклонений якутского языка. Для обозначения наклонения несовершившегося (неосуществленного) действия мы будем придерживаться варианта А.В.Дыбо CUNC (Cuncative) «Cunc — кункатив, еще не совершившееся действие» [7], так как форма еще не совершившегося действия в хакасском языке соответствует наклонению несовершившегося (неосуществленного) действия в якутском языке.

Для формы причастия нами добавлена дополнительная разметка PCP – PCP_CUNC.

Таблица 8

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ КУНКАТИВНОГО

Сокращения Tags	**Расшифровка сокращений Full term**	**Название категории Name of the category**	**Алломорфы Allomorphs**	**Морфемы Morphemes**
PCP_CUNC	participle cuncative	причастие кункативное	-а/ -о/ -э/ -ө илик -ыы/ -ии/ -уу/ -үү илик	-А илик -ЫЫ илик

ТРЕТИЧНЫЕ ПРИЧАСТИЯ

Присвоение грамматических помет третичным причастиям, так называемым собственно якутским причастиям, образованных от первичных путем присоединения морфемы *–лаах*, затруднено из-за отсутствия аналогичных вариантов в рассматриваемых тюркских языках.

«В предикативной функции причастия на *–ыахтаах, -ыа суохтаах, -ымыахтаах, -ардаах, -баттаах, -быттаах, -батахтаах* имеют время-модальное и видо-временное значенияю»… Роль аффикса на *–лаах* в образовании этих причастий является определяющей [6, с. 296].

Причастия на –ардаах/-баттаах, на –ыахтаах/-ымыахтаах/-ыа суохтаах

«Причастие на *–ардаах/-баттаах* выражает действие, которое субъект обязан, призван совершить в будущем, или обозначает субъект, который обязался совершить действие» [6, с. 125-126].

«Причастие на *–ыахтаах/-ымыахтаах/-ыа суохтаах* обозначает действие, которое субъект должен совершить в будущем, или самого субъекта, который назначен совершить действие в будущем. Причастие на *–ыах/-ымыах* в определительном употреблении имеет значение возможности (барыах дьон ‘люди, которые могут уйти’), а причастие на *–ыахтаах/-ымыахтаах/-ыа суохтаах* – значение долженствовательности (барыахтаах дьон ‘люди, которые должны уйти’)» [6, с. 126].

«При этом … причастие на *–ар/-бат*, актуализируя значение будущего времени, воспринимает аффикс на *–лаах* как форму, выражающую обязательство; причастие на *–ыах/-мыах/-ыа* суох к своему значению будущего времени придает долженствовательное значение» [6, с. 126].

Причастия на *–ардаах, –ыахтаах* образуют формы долженствовательного наклонения – настояще-будущее и будущее время. Нами долженствовательному наклонению якутского языка была присвоена грамматическая помета OBL (Obligative) (см. раздел **Наклонения**) — облигативами называют модальные значения, а также формы и конструкции со значением обязательности. В связи с чем, для третичным причастиям на –ардаах, -ыахтаах предлагаем присвоить разметки PCP_OBL (participle obligative).

Таблица 9

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ ОБЛИГАТИВНОГО

Сокращения **Tags**	**Расшифровка сокращений** **Full term**	**Название категории** **Name of the**	**Алломорфы** **Allomorphs**	**Морфемы** **Morphemes**

		category		
PCP_ OBL	participle obligative	причастие облигативное	-ардаах/-эрдээх/-ордоох/-өрдөөх -ыахтаах/-иэхтээх/-уохтаах/-үөхтээх	-АрдААх -ЫАхтААх
PCP_OBL_NEG	participle obligative negative	причастие облигативное отрицательная форма	-баттаах/-бэттээх/-боттоох/-бөттөөх -ымыахтаах/-имиэхтээх/-умуохтаах/-үмүөхтээх -ыа/-иэ/-уо/-үө суохтаах	-бАттААх -ЫмЫАхтААх - ЫА суохтаах

Причастие на –быттаах/-батахтаах.

Причастие на –быттаах/-батахтаах выражает результат действия, которое субъект давно совершил **один раз** (*выделено нами*), то есть результат обладания совершенным действием, также субъект действия с тем же значением [6, с. 126].

Семантику однократности повторения ситуации можно рассмотреть на примерах с ГСЯЛЯ:

1 л. барбыттаахпын
2 л. барбыттааххын — ‘(мне, тебе, ему) пришлось однажды уйти’
3 л. барбыттаах

Исходя из семантики причастия на *–быттаах/-батахтаах* предлагаем использование следующей грамматической пометы – PCP_REFACT (participle refactive).

«Рефактивным принято называть значение однократного повторения ситуации (повторного осуществления действия)» [8, с.67]. «Значение однократного повторения ситуации (‘Р снова / еще раз’) выражается граммемой рефактива; показатели рефактива по своим семантическим и морфологическим свойствам, как правило, отличаются от других показателей глагольной множественности, апеллирующих к неограниченной кратности ситуаций» [3, с. 159].

Таблица 10

МОРФОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ПРИЧАСТИЯ РЕФАКТИВНОГО

Сокращения Tags	**Расшифровка сокращений Full term**	**Название категори Name of the category**	**Алломорфы Allomorphs**	**Морфемы Morphemes**
PCP_ REFACT	participle refactive	причастие рефактивное	- быттаах/-биттээх /-бүттээх/-буттаах -пыттаах/-питтээх /-пүттээх/-путтаах -мыттаах/-миттээх/ -мүттээх/-муттаах	-ББІттААх
PCP_ REFACT_ NEG	participle refactive negative	причастие рефактивное отрицательная форма	-батахтаах/-бэтэхтээх/ -ботохтоох/-бөтөхтөөх -патахтаах/-пэтэхтээх/ -потохтоох/-пөтөхтөөх -матахтаах/-мэтэхтээх/ -мотохтоох/-мөтөхтөөх	-БАтАхтААх

ЛИТЕРАТУРА:

1. Грамматика современного якутского литературного языка: Фонетика и морфология. Т.1/Л. Н. Харитонов, Н. Д. Дьячковский, С. А. Иванов и др.; Отв. ред. Е. И. Убрятова. – М.: Наука, 1982. – 496 с.
2. Коркина Е.И. Наклонения глагола в якутском языке. М.: Наука, 1970. – 308 с.
3. Плунгян В. А. Общая морфология. — Москва: Едиториал УРРС, 2003. – 374 с.
4. Попова Н.И. Атрибутивная функция причастных форм якутского языка. // Исследования по грамматике якутского языка: Сборник научных трудов. — Якутск: Изд-во ЯФ СО АН СССР, 1983. – С.104-109.
5. Тумашева Д.Г. Татарский глагол (Опыт функционально-семантического исследования грамматических категорий). – Казань: изд-ва Казанского университета, 1986.
6. Филиппов Г.Г. Причастия якутского языка: комплексное типологическое функционально-семантическое исследование. – Якутск: Издательский дом Северо-Восточного федерального ун-та, 2014. – 606 с.
7. Козлова Н.В. Лингвистические корпуса: определение основных понятий и типология//Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2013. – Т.11, № 1.- с. 79-99.
8. Заморщикова Л.С. Ассоциативно-вербальная сеть и системность образа мира // Гуманитарные научные исследования. 2014. № 3 [Электронный ресурс]. URL: http://human.snauka.ru/2014/03/6130 (дата обращения: 20.11.2016).

СТАТИСТИЧЕСКО-СИНЕРГЕТИЧЕСКОЕ ИССЛЕДОВАНИЕ УЗБЕКСКИХ ФОЛЬКЛОРНЫХ ЖАНРОВ

Д. Б. Уринбаева[1], Самаркандский государственный университет, Самарканд, Узбекистан, dilbarxon@inbox.ru

В статье рассматривается лексико-статистическая структура узбекского языка, фольклорные тексты на основе количественного сравнительного исследования: средняя частота словоформ, покрываемость текста различными частями лексики и соотношение редких (случайных) лексических единиц.

Ключевые слова: *средняя частота слов; коэффициент заполнения; эпос; фольклор; частота; количественный; синергетика; синтетизм; статистика.*

STATISTICAL-SYNERGETIC STUDY OF THE UZBEK FOLKLORE GENRES

D. B. Urinbaeva[1], 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.

Key words: *the average frequency of words; fill factor; epos; folklore; frequency; quantitative; synergitic; rate synthetism; statistics.*

Предлагаемая статья выросла на материале двух авторских исследований. Первое — ведущееся в течение многих лет изучение устной народной традиции, в основном книги народного творчества как дастан, сказки, пословицы, загадки и песни. Второе — работа по созданию частотного словаря языка узбекского фольклора. Казалось бы, что обе работы самостоятельны и нет смысла соединять их в одну статью, однако с самого начала между ними существовала связь. В исследованиях ставилась цель описать традиционные тексты устно-письменной природы, главным образом дастана, сказки, пословицы, загадки и песни. Эта же область народного творчества стала главным материалом при разработке частотного словаря языка узбекского фольклора.

Фольклорный жанр – многогранный объект исследования, входящий в разные лингвистические концепции: как известно, единицы, большие по объёму, чем предложение, изучаются в лингвистике текста, теории текста, стилистике текста. Многие термины и понятия используются сразу несколькими областями науки. Так, например, все указанные области науки в своём терминологическом аппарате содержат термин жанр, который трактуется в данных областях лингвистики по-разному. Определение значения термина и круга концептуально связанных с ним понятий, необходимых каждый раз при анализе конкретного жанра, зависит от того, в проблематике какой из указанных областей науки находится ответ на вопрос о построении его типологии.

В.Я.Пропп установил в основе исследования волшебных сказок, что «единство структуры соответствует единству всей поэтики волшебной сказки и единству выраженного в ней мира идей, эмоций, образов героев и языковых средств»[1,47]. Значит, эти слова стоит расценить как исключительно плодотворное методологическое положение, завещанное нам великим ученым для дальнейшего приложения его к богатейшему разнообразию фольклорных жанров. В конечном счете категория жанра нужна нам не для упорядочения материала и внешней его характеристики, а для проникновения в «мир идей, эмоций, образов», созданных в рамках жанровой системы. П.Г.Богатырев пишет: «в основе языка литературы лежит литературный язык, в основе языка фольклора лежит диалектный язык литературы – это литературный язык в его эстетической функции, язык фольклора – это диалект в его эстетической функции. Невозможно выявить специфику языка фольклорного произведения того или иного жанра, не зная диалекта, на котором исполняется это произведение, диалекта в его коммуникативной функции» [2,106-116]. Вместе с тем ученый отмечает, что, обнаруживаются значительные отличия между языком фольклора и диалектом на уровне фонетики, морфологии и синтаксиса, и поэтому народно-поэтический язык нельзя отождествлять с языком разговорной речи.

Уместно отметить, несмотря на то, что в сфере изучения фольклорных произведений, в частности, с точки зрения лексикографики, удалось достичь определённых результатов, ещё не создан толковый словарь, включающий в себя все произведения устного народного творчества. В решении данных задач неоценима роль статистического метода.

Узбекская грамматическая система, как и вообще система тюркских языков, способна породить практически бесконечное количество словоформ. В связи с этим особо важным при изучении лексико-статистической структуры узбекского фольклорного текста является сопоставительное исследование средней повторяемости словоформы, покрываемости текста различными участками частотного словаря, а также

соотношения редких лексических единиц и достаточно часто и устойчиво употребляющихся словоформ и слов.

При решении ряда теоретических и прикладных задач необходимо выделить ту лексику, которая с достаточно большой вероятностью будет встречаться наугад взятых текстах данного языка, стиля или подъязыка [3,67]. Опираясь на данную выборку текста, следует отделить редкие словоформы и слова от лексических единиц, имеющих среднюю и высокую употребительность, а затем определить, какой процент текста покрывают эти средне- и высоко употребительные слова. Исходя из опыта предшественников [3,67;4,368;5,56], определяется граница между редкими и употребительными лексическими единицами, эту границу можно провести между единицами, употребленными один и два раза.

Рассмотрим оценки веса редких словоформ с точки зрения квантитативной типологии, предполагая, что они могут служить характеристиками, отличающими тексты фольклора [6;7;8;9;10]. Для этого при сравнении редкоупотребляемых словоформ мы будем опираться на выборки одинакового объема для текстов фольклора, приведенных в табл. 1-2-3-4, где при сравнении редкоупотребляемые словоформы в 1000, 2000, 6000 объемах выборки в различных жанрах фольклора.

1-таблица

КВАНТИТАТИВНЫЕ ДАННЫЕ ПО ВЫБОРКАМ В 1000 С/У

Частотные словари и словоформы	**N**	**$L_{с/ф}$**	**F_1**	**$F_{1\%}$**	**F_2**	**$F_{2\%}$**	**$F_{1,2}$**	**$F_{1,2\%}$**	ξ
дастан	1000	553	F_{385}	38	F_{73}	7,3	F_{458}	45,3	55
сказки	1000	500	F_{346}	34,6	F_{67}	6,7	F_{413}	41,3	59
пословицы	1000	582	F_{435}	43,5	F_{72}	7,2	F_{507}	50	50
нар.песни	1000	559	F_{459}	46	F_{51}	5,1	F_{510}	51	49
загадки	1000	627	F_{371}	37	F_{83}	8,3	F_{454}	45,4	55

2-таблица

КВАНТИТАТИВНЫЕ ДАННЫЕ ПО ВЫБОРКАМ В 2000 С/У

Частотные словари и словоформы	**N**	**$L_{с/ф}$**	**F_1**	**$F_{1\%}$**	**F_2**	**$F_{2\%}$**	**$F_{1,2}$**	**$F_{1,2\%}$**	ξ
дастан	2000	1090	F_{774}	38,7	F_{158}	7,9	F_{932}	46,6	54

сказки	2000	856	F_{539}	26,9	F_{144}	7,2	F_{683}	34,1	66
пословицы	2000	1208	F_{898}	44,9	F_{167}	8,3	F_{1065}	53,2	47
нар.песни	2000	1197	F_{828}	41,4	F_{202}	10	F_{1030}	51,5	49
загадки	2000	1036	F_{697}	34,8	F_{170}	8,5	F_{867}	43,3	57

3-таблица

КВАНТИТАТИВНЫЕ ДАННЫЕ ПО ВЫБОРКАМ В 6000 С/У

Частотные словари и словоформы	**N**	**$L_{с/ф}$**	**F_1**	**$F_{1\%}$**	**F_2**	**$F_{2\%}$**	**$F_{1,2}$**	**$F_{1,2\%}$**	ξ
дастан	6000	2499	F_{1607}	26,7	F_{396}	6,6	F_{2003}	33	67
сказки	6000	2615	F_{1744}	29	F_{358}	5,9	F_{2102}	35	65
пословицы	6000	2725	F_{1875}	31,2	F_{419}	6,9	F_{1456}	25	76
нар.песни	6000	2693	F_{1923}	32	F_{408}	6,8	F_{2331}	38	62
загадки	6000	2685	F_{1775}	29	F_{419}	6,9	F_{2194}	36	64

При выборке в 1000 словоупотреблений редкоупотребляемые словоформы составляют в текстах сказки ξ=59, в народных песнях ξ=49. При увеличении объема в 2000 с/у доля редкоупотребляемых словоформ в сказках увеличилась ξ=66, а в пословицах ξ=47.

При увеличении объема нашей выборки в 6000 с/у доля редкоупотребляемых словоформ в пословицах ξ=76, в народных песнях ξ=62. С увеличением объема выборки редкоупотребляемые словоформы тоже увеличились.

4-таблица

КВАНТИТАТИВНЫЕ ДАННЫЕ ПО ОБЩЕМУ ОБЪЕМУ

Частотные словари и словоформы	**N**	**$L_{с/ф}$**	**F_1**	**$F_{1\%}$**	**F_2**	**$F_{2\%}$**	**$F_{1,2}$**	**$F_{1,2\%}$**	ξ
дастан	96011	14029	F_{7404}	52,7	F_{4342}	30,9	F_{11746}	83,6	88
сказки	77304	14837	F_{8265}	55,7	F_{4580}	30,8	F_{12845}	86,5	83

пословицы	45132	12231	F_{7029}	57,4	F_{4054}	33,1	F_{11083}	90,5	75
нар.песни	31858	10692	F_{6809}	63,6	F_{2538}	23,7	F_{9347}	87,3	71
загадки	27334	8569	F_{5275}	61,5	F_{2636}	39,7	F_{7911}	92,2	72

В общем объеме выборки самую высокую частоту показывают тексты дастанов ξ=88, самую низкую частотность показывают народные песни ξ=71.

5-таблица

ДИНАМИКА РОСТА РЕДКОУПОТРЕБЛЯЕМЫХ СЛОВОФОРМ

Частотные словари и словоформы	**N (словоупотреблений)**			
	Зоны			**Общий объем**
	1-1000	**1-2000**	**1-6000**	
дастан	55	54	67	88
сказки	59	66	65	83
пословицы	50	47	76	75
нар.песни	49	49	62	71
загадки	55	57	64	72

Из всего сказанного выше следует, что описанный статистический эксперимент, проведенный в одинаковых условиях относительно фольклорных жанров дает разные по достоверности и качеству результаты. Динамику роста этих показателей можно наблюдать, что в сказках при увеличении объема выборки доля редкоупотребляемых словоформ почти не изменилась.

В условиях общего объема выборки дадим краткий квантитативно-типологический комментарий между жанрами фольклора. Поведение значения коэффициента заполнения ξ между жанрами фольклора при общем объеме выглядит следующим образом:

Дастан-сказки – значение ξ в дастане на 5 % больше, чем в сказке ($\xi_{сказки} < \xi_{дастан}$).

Дастан – пословицы – значение ξ в дастане на 13% больше ($\xi_{пословицы} < \xi_{дастан}$).

Дастан – народные песней – значение ξ преобладает в дастане на 17% по сравнению с народной песни ($\xi_{народныепесни} < \xi_{дастан}$).

Дастан – загадкыми – значениями ξ в дастане на 16 % по сравнению с загадками ($\xi_{загадки} < \xi_{дастан}$).

Значение ξ ниже в народных песнях на 17% по сравнению с другими жанрами, что свидетельствует о доминировании редкоупотребляемых словоформ.

Сопоставляя значения коэффициента заполнения ξ текстов, редкоупотребительными словоформами в испытуемых жанрах узбекского фольклорного текста, получим следующее неравенство:

$$\xi_{народныепесни} < \xi_{загадки} < \xi_{пословицы} < \xi_{сказки} < \xi_{дастан}$$

Это неравенство является показателем различия в разнообразии сопоставляемых жанров фольклора. Таким образом, чем ниже значения коэффициента заполнения ξ, тем разнообразнее является рассматриваемый жанр. В нашем случае таким жанром являются народные песни.

ВЫВОДЫ

Нами были исследованы частотные словари – наиболее элементарный и в то же время имеющий большое практическое значение способ описания статистики словарного состава. Общее рассмотрение разнообразных по структуре и задачам частотных словарей показало, что помимо ряда специфических вопросов, связанных с составлением словарей определенного типа, существенно решить более общую задачу – дать методику составления частотного словаря, обеспечивающего получение сведений требуемой точности о заданном проценте слов текста.

Квантитативное исследование узбекского фольклорного текста и сопоставление полученных данных позволило выявить количественно типологические расхождения между изучаемыми текстами. Эти различия последовательно и недвусмысленно обнаруживаются в таких величинах, как средняя частота словоформы, рост статистической покрываемости текста, заполняемость текста наиболее употребительными и редко употребительными словоформами.

В нашем лингвостатистическом эксперименте значение степени коэффициента синтетичности в испытуемых жанрах узбекского фольклорного текста имеет место следующее неравенство:

$$\mathrm{Sint}_{дастан} < \mathrm{Sint}_{сказка} < \mathrm{Sint}_{пословицы} < \mathrm{Sint}_{загадки} < \mathrm{Sint}_{народныепесни}$$

Как видно из этого неравенства наиболее высокую степень имеют народные песни, а самая низкая свойственна текстам дастана. Это свидетельствует о том, что народные песни обладают доминирующим статистическим весом с точки зрения разнообразия лексем и их грамматических форм по отношению к другим жанрам фольклорного жанра.

Помимо этого в качестве показателя различна между жанрами фольклора, был использован коэффициент заполнения текста редкоупотребительными лексическими единицами. Сопоставление этих коэффициентов между жанрами фольклора, представим в виде следующего неравенства:

$$\xi_{народныепесни} < \xi_{загадки} < \xi_{пословицы} < \xi_{сказки} < \xi_{дастан}$$

Это неравенство также свидетельствует о разнообразии текстов народных песен.

Информационно-статистический эксперимент позволяет различить типологические различия и среди жанров фольклора. В народных песнях дает высокий рост темпа покрываемости, что свидетельствует, очевидно, об их большей аналитичности по сравнению с другими жанрами. Последнее наблюдение вполне согласуется с конкретными историческими судьбами жанров, например, в народных песнях (которые мы исследовали) много детского фольклора. Детский фольклор — одно из самых живых и богатых явлений узбекской культуры. В нем одновременно существуют и очень старинные произведения, и только что рожденные. И те, и другие непрерывно обновляются, переделываются. Как и фольклор взрослых, детский фольклор отражает историю, идет с ней в ногу современем. Веселые стишки, смешные песенки, забавные дразнилки, скороговорки знают, передают друг другу дети всей страны. Это общенациональное детское искусство слова. В словаре народных песен много слов непонятных, придуманных детьми. За счет этого, лексика народных песен относительно других жанров разнообразна.

ЛИТЕРАТУРА

1. Пропп В.Я. Морфология волшебной сказки. – Москва: Лабиринт, 2001. С.47.
2. Богатырев П.Г. Язык фольклора. Вопросы языкознание. №5, 1973. С. 106-116.
3. Айимбетов М.К. Проблемы и методы квантитативно-типологического измерения близости тюркских языков. Автореф.дисс..док.филол.наук. — Ташкент, 1997. С.21.
4. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика: учебное пособие для пединститутов. – М.: Высшая школа, 1977. С.368.
5. Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики. — Таллин: Валгус, 1987. Материалы для обработки
6. Dastan. Alpomish. (1998) Fozil Yo‘ldosh o‘g‘li. Toshkent: «Sharq» nashriyoti-matbaa konserni bosh tahririyati: Ravshan. (1954) Toshkent: Fan.
7. Сказки. O‘zbek xalq ertaklari. (2007) I tom. Toshkent: «O‘qituvchi» nashriyot-matbaa ijodiy uyi.
8. Пословицы. O‘zbek xalq maqollari. (2005) Toshkent: «Sharq» nashriyot-matbaa aksiyadorlik kompaniyasi bosh tahririyati.
9. Загадки. Topishmoqlar. (1981) Toshkent: G‘.G‘ulom nomidagi Adabiyot va san’at nashriyoti.
10. Песни. Бойчечак. (1984) Тошкент: Ғ.Ғулом номидаги Адабиёт ва санъат нашриёти.

CORPORA — THE WHYS AND WHEREFORES

Belinda Mary Harper Sousa Maia,
Universidade do Porto, Portugal, bhsmaia@gmail.com

This paper hopes to contribute to the discussion of how the Central Asian languages can further develop the resources they already have, or still need, to make fast progress in language technologies. It is not intended as a research paper of the kind normally accepted for the Turk Lang conference. Instead it is a reflection on the subject of 'corpora', based on over my 30 years of experience of making and using corpora of written language.

Corpora projects are started for a variety of reasons; some to respond to research in the humanities, others to develop language technology. A corpus can be constructed carefully to act as a representative repository of a national language, or it can be bootstrapped by a web crawler for more immediate use. Between these two extremes are smaller corpora that are compiled for a wide variety of specialized objectives. The focus here will be on the interdisciplinary cooperation needed in the planning of corpora and the subsequent collection and annotation of texts. I shall also refer to some of the many areas that can benefit from the research that provides these resources and tools.

Key words: *Central Asian languages, corpora.*

КОРПУС — ПОЧЕМУ И ГДЕ

Белинда Мари Харпер Сьюзи Майя,
Университет Порто, Порто, Португалия, bhsmaia@gmail.com

Данная статья нацелена на внесение вклада в обсуждение того, каким образом языки Центральной Азии могут создавать необходимые ресурсы, а также продолжать развивать уже имеющиеся, для активного развития языковых технологий. Статья не носит исследовательский характер, но в ней представлен взгляд о вопросе «корпусов», основанный на более чем 30-летнем личном опыте автора по созданию и использованию корпусов письменного языка.

Корпусные проекты создаются по ряду причин причинам; одни — в ответ на новые исследования в области гуманитарных наук, другие — для развития языковых технологий. Корпус может быть специально разработан для того, чтобы выступать в качестве репрезентативного хранилища языка, либо может быть введен путем сканирования веб-страниц для более оперативного применения. Между этими двумя

крайностями можно отметить малые корпуса, составленные для решения широкого круга специализированных задач. Основное внимание уделяется междисциплинарному взаимодействию, необходимому при проектировании корпусов и последующем сборе и аннотации текстов. Статья также обращает внимание на некоторые из множества сфер, которые могут извлечь пользу из исследований, представленных этими ресурсами и инструментами.

Ключевые слова: *языки Центральной Азии, корпус.*

Introduction

Our languages are an essential part of our personal, cultural, and political identities, and any study of them is based on a variety of beliefs, attitudes, and traditions. For example, Lithuanian claims to be the nearest existing relation of Proto-Indo-European; Icelandic, with its long literary tradition, has always striven to maintain the 'purity' of its language; and Slovenian identified its speakers long before Slovenia became an independent country in 1991. Other languages are still in the process of self-identification and stabilization, like Tetum in East Timor, derived from the local languages and the colonial influence of Portuguese. In many cases, attempts to provide a standard form of a language in a particular territory is disputed for political and cultural reasons. (Ex)colonial languages often continue as linguas francas, influencing the local languages, particularly the lexicon that reflects contemporary reality. These examples are taken from personal experience, but no doubt there will be those at TurkLang who recognize such situations in relation to their own languages.

These factors mean that those interested in compiling corpora will approach the task from a variety of positions. Historical linguists will need texts from past centuries; literary experts will want literary corpora; modern language experts need empirical evidence from contemporary texts for designing new dictionaries and grammars; others try to examine how discourse functions in all its many forms. Computational approaches tend to focus on the 'how' and 'what' rather than the 'why' and 'wherefore'.

Previous TurkLang Proceedings contain several references to corpora building in the various languages that take part, with a special section given to the subject in the 2015 Proceedings. From what I can discover, there is a large Turkish National Corpu s, and there are two corpora in Kazakh, one annotated by philology students, the other automatically. Tatar boasts a fairly large corpus and several of the other languages either have corpora of various sizes or are planning them.[5] The Proceedings also discuss the various problems of annotation.

Computers and language

The relationship between computers and language is complex and has led to much discussion for some time. At one extreme we have those who believe that

[5] I am indebted to Google Translate from Russian to English in my search of previous TurkLang Proceedings.

there is an underlying universality to all languages that can be expressed in terms similar to logic and mathematics; at the other side of the spectrum are those who believe that human languages can never be understood by machines. Somewhere in between exists a large body of research dedicated to providing useful language technologies.

Theoretical linguists still develop their work around simplified samples of language, but the results prove less than useful for dealing with real world language use and the challenges it presents to those working in natural language processing (NLP). Fortunately, over the last few decades, the computerization of language data has opened up many possibilities for research, allowing for extensive use of empirical evidence of language used in real life situations. This evidence can be found in corpora and other language resources, and today it is essential for both the humanities and the information technology (IT) research communities. However, the research has often proceeded along parallel paths and led in different directions. One of the objectives of this paper is to show how this disconnection between disciplines is a negative factor, and to encourage mutual understanding and cooperation.

3. Compiling corpora – monolingual, parallel and comparable

The objectives for developing any corpus should be defined from the outset, whether they are for very large national corpora, or for small corpora for specialized research. Although the focus is often on monolingual corpora, parallel and comparable corpora are of great interest to those involved in cross-language research and machine translation.

The collection of texts for corpora has evolved over the last few decades from manual typing, to scanning with optical character recognition (OCR) software, to today, when enormous quantities of text can be harvested off the Internet or other digital resources very quickly. The advance of technology in this respect has allowed developments beyond the aspirations of the early corpora makers, but it has also brought problems.

One of these problems is the selection of relevant data; the other is the thorny issue of copyright. While corpora compilers debate the quality and quantity of texts they need, and struggle to convince suppliers of text that they will respect copyright, many in the IT world, (in)famously Google, Facebook, and others, use ‘big data’ or large quantities of text that have been harvested with little regard for its possible relevance, and with no consent from the authors. One might wonder why such issues will affect corpora compilation and computational linguistics, but if one considers the end results of such research, whether it is to promote advertising through effective opinion mining or contribute to an understanding of the slang of the mafia, the connections are easier to make.

The best known, early monolingual corpora, like the Brown Corpus, the Bank of English, and others, provided empirical data from which to develop dictionaries and grammars based on contemporary usage, rather than the

intuitions of (often pedantic) academics. As technology improved, the objectives became more ambitious and diversified, as can be seen from Mark Davies' corpus collection in English and other languages[1] or Diana Santos' Linguateca corpora for Portuguese[2]. Projects like these, however, are the results of many years of work by interdisciplinary teams of computer scientists and linguists.

While the compilers of Very Large Corpora argue about how to provide a 'balanced' corpus and on how to provide metadata on the texts[3], there has been a wide variety of research using small, specialized monolingual corpora for well-defined purposes. Linguists can analyze the discourse of an individual or specific type of text, using technology like Sketch Engine[4]. Computer scientists train tools for data, text, and opinion mining using only texts in the discipline and language in which they are interested.

Parallel corpora, or originals in one language aligned with their translations into other languages are essential for researchers in statistical machine translation, and organizations like the European Commission supply material for this research. Large translation technology companies also use their translation memories for similar ends, but smaller translation agencies and individual translators are rarely interested in ceding their material to help what they regard as the competition. Some researchers turn to 'comparable corpora', such as news reports and Wikipedia, where the texts in both languages are similar enough to facilitate the automatic retrieval of named entities, key terms and expressions, and similar texts in both monolingual and multilingual situations. The Web itself and big MT projects, like Google Translate, use all this material, together with all the data at their disposal to further both linguistic and language technology research. This has allowed enormous advances for the most widely used languages; the situation for languages with few resources is different.

4. Corpora and linguistic information

The term 'Corpus' and its plural Latin form 'corpora' is not always used in exactly the same way in NLP; the linguists add linguistic information to the corpus; the computer scientists may dispense with it, or attempt to categorize the information using big data and algorithms. If a 'corpus' does not include linguistic information, we call it a 'raw' corpus or simply a database of texts that can be useful for certain types of projects, like lexicography or terminology, where observing and recording the use of words in context is the objective.

NLP can clearly both contribute and benefit from adding linguistic information to a collection of texts. Although there are linguistic categories that are recognized to exist across all languages, like nouns or verbs, and Subject or Object, each language expresses them differently. It is clearly helpful to recognize those categories that languages have in common for cross-language research, but the academic influence of English on theoretical linguistics and computational methods sometimes distorts the outcomes. Categorization is difficult enough for trained linguists; computerizing this knowledge to create

(even semi-) automatic morphological analyzers and parsers is even more complicated, but not impossible. However, technologies developed for languages with simple morphology, like English, or the Romance languages with their complex systems of gender, complex verb inflections, and noun adjective agreement, are not easily adaptable to the agglutinative Turkic languages.

Annotation of corpora creates discussions between linguists anxious to apply their favorite theories on language to morphological or syntactic analysis but, since manual annotation is too time-consuming, they usually have to compromise over what is computationally possible. The compilation and annotation of corpora is not, nor should it be, a linear process – i.e. one makes the corpus and then annotates it. The linguistic and computational approaches benefit from discussion and coordination, and theories are often modified and refined when confronted with empirical evidence. The evidence for this can be found when comparing grammars written before and after consulting evidence from corpora.

5. Using corpora

A corpus should be planned with certain objectives in mind, whatever the size or the project. It is generally agreed that even Very Large Corpora are only a sample of the language they represent, and the texts that are included should be classified properly into sub-corpora, so that researchers can select the specific area they wish to investigate.

5.1 Corpora and linguists

Linguists traditionally build corpora in order to provide empirical evidence for their dictionaries and grammars. Etymological dictionaries, like the Oxford English Dictionary, will need texts taken from several centuries. They will show, however, that the lexicon and grammar of any language changes over time, so dictionaries and grammars for teaching the contemporary use of languages need to be based on large corpora of texts from recent decades. Corpora assist the development of tools like wordnets, framenets, and treebanks. As this information is then interconnected, machine earning can use it to train the analysis of much bigger quantities of text.

Since there are now tools for creating corpora that require less input from their computer colleagues, many linguists, particularly those working in English, have (re)turned to focusing the perspectives of the humanities. For example, the Corpus Linguistics conferences show more interest in political, social, educational and academic discourse, or the psychology of language teaching than in computational developments. There is, of course, a need and space for this kind of research, but interdisciplinary cooperation suffers as a result.

5.2 Corpora and computational linguistics

A much wider perspective of corpora and related work in computational linguistics is reflected in the Proceedings of the bi-annual LREC — Language Resources and Evaluation conferences, where one can search the proceedings

since 2000[4] online. An analysis of the titles of the articles mirror developments in corpus making and annotation over this period. In the latest version, apart from references to more traditional problems of corpus compilation and annotation, one can see that the tools and resources are increasingly sophisticated and varied. Technology now provides possibilities for spoken and multimodal corpora that were not available in 2000; tagging now also covers gesture, gaze, key strokes, sign language, and images; annotation can be of concepts, arguments or sentiments; and there are claims of automatic detection of these items.

Luckily, languages that still need to build sizeable corpora can take full advantage of some of the more advanced technology for finding and processing texts. However, although a language with advanced speech recognition technology can use it to transcribe spoken language in real time, the same technology will not work for a language for which it has not been trained. Therefore it is in the interests of all concerned for computational linguists working with less resourced languages to build large reliable resources and effective tools before attempting more ambitious projects. It is not enough to create 'toy' resources or prototype tools; they must be proven to work in a wide variety of circumstances.

5.3 The need for interdisciplinary cooperation

Over the years, NLP has brought linguistics and computer technology together so successfully that today we forget how each discipline has affected the other. How many people stop to think of where the spelling and grammar checkers in programs like Microsoft Word come from, or how our cell phones use predictive writing to accelerate our messages on social media? Google uses language to help us find what we want, and Facebook uses our 'likes' and comments to target us with information it 'thinks' we will like. Our computer scientists may claim it is all done by algorithms, but others would argue that the more linguistic information that is added, the better the algorithms function.

The media love to tell stories of academic articles written by computers, but analysis by an educated person will detect what has happened, – and a linguist should be able to explain why. Question and answer technology and summarization programs will only work well with considerable linguistic input. Sentiment analysis and opinion mining will function better with insights from sociolinguists and psychologists who, in turn, can gain insights from quantitative analysis found in suitable corpora. Forensic linguistics uses the linguist's qualitative judgment supported by a computerized quantitative analysis to detect situations like plagiarism, false suicide notes, or faked SMS messages.

Machine translation is the ultimate example of coordination between linguistics and computer technology. For years researchers struggled to solve MT with complex linguistic analysis for Rule-Based MT – and some still do. Statistical MT (SMT) advanced beyond many people's expectations by matching material from parallel and comparable texts. Google Translate now claims to use

Neural MT, which would suggest it has progressed from SMT by adding information acquired by machine learning of linguistic data, acquired, I would dare to suggest, from the annotated corpora and the related resources and tools discussed here.

Conclusions

As stated at the beginning, this paper is meant only as a starting point for discussion, and the focus is to ask questions about why anyone would want to compile a corpus and with what objectives in mind. Research needs a focus and, importantly, funding. It is clearly easier to persuade the powers-that-be to fund speech technology for forensic purposes like detecting which criminal or terrorist is talking on the phone, than to argue for the need for an etymological dictionary. However, there are more mundane, but more generally useful reasons for building corpora, and academic transparency and cooperation are essential for progress.

All of what is said here could be backed up by an extensive list of references that it would be impossible to include in the space allowed. Should anyone be interested in a particular point, please do not hesitate to contact me.

REFERENCES:

https:// corpus.byu.edu
https://www.linguateca.pt/ACDC/
http://www.tei-c.org
https://www.sketchengine.eu
http://www.lrec-conf.org/proceedings/

MILLIY KORPUSGA ASOSLANGAN TARJIMA

F. Bakiyev, Samarqand davlat chet tillar instituti,
Samarqand, O‘zbekiston, bj.fakhriddin@gmail.com

Maqolada tilshunoslik va tarjima ishlarida korpusning ahamiyati, o‘zbek tilida korpus yaratish, ushbu sohaning rivojlanishida korpusga asoslangan tarjima, avtomatlashtirilgan va statistik tarjima texnologiyalarining ahamiyati haqida fikrlar bildirilgan.

Tayanch so‘zlar: *korpus, monolingual korpus, mos yozuvlar korpusi, qiyoslangan ikki tilli korpus, parallel korpus, tadqiqotchi tomonidan yaratilgan parallel korpus, matn uslubi, kontrast tahlil.*

CORPUS BASED TRANSLATION STUDIES

F. Bakiyev, Samarkand State Institute of Foreign Languages,
Samarkand, Uzbekistan, bj.fakhriddin@gmail.com

The importance of corpus in linguistics and translation studies, necessity to create corpus in the Uzbek language, issues to develop this sphere, corpus based translation, automatic and statistic translation are discussed in the article.

Key words: *corpus, corpora, annotation, monolingual corpora, reference corpora, comparable bilingual corpora, parallel corpora, researcher-constructed parallel corpora, the style of a text, contrastive analysis.*

ПЕРЕВОД НА БАЗЕ НАЦИОНАЛЬНОГО КОРПУСА

Ф. Бакиев, Самаркандский государственный институт
иностранных языков,
Самарканд, Узбекистан, bj.fakhriddin@gmail.com

В статье описываются значение корпуса в лингвистике и переводческих исследованиях, необходимость создания корпуса на узбекском языке, вопросы развития этой сферы, перевод на основе корпуса, технологии автоматического и статистического перевода.

Ключевые слова: *корпус, аннотация, одноязычные корпуса, эталонные корпуса, сопоставимые двуязычные корпуса, параллельные корпуса, построенные исследователем параллельные корпуса, стиль текста, контрастивный анализ.*

Hozirgi kunda nafaqat tilshunoslikda, balki tarjima nazariyasi va amaliyotida ham korpus lingvistikasi, milliy korpus terminlari keng qo‘llanilib, bu sohada ingliz, ispan, fransuz, nemis, rus va boshqa ko‘plab tillarda axborot texnologiyalariga asoslangan samarali nazariy va amaliy ishlar, tadqiqotlar, loyihalar bajarildi va bajarilmoqda. O‘zbek tilida esa bu borada ko‘zga ko‘rinarli hech qanday ish qilinmagani achinarli, bu haligacha mamlakatda kompyuter

lingvistikasi mutaxasislari tayyorlanmasligi bilan izohlanishi mumkin. Korpus bu li matn yoki nutq bo‘lib, unda turli uslubdagi matnlar (tadqiqotchilar tomonidan suniy yaratilmagan balki muloqot maqsadida yuzaga kelgan), masalan badiiy, publisistik, ilmiy, internet sahifalari matnlari va bq. kabi ko‘plab matnlar qamrab olinadi. Matn kompyuter tanishi, matnga ishlov berish uchun kodlashtiriladi va anotatsiyali matn turli maqsadlar (lingvistik va tarjima tadqiqotlarini olib borish, inson resursisiz katta hajmli ma’lumotlarga ishlov berish, kompyuterning inson nutqini tushunishi va berilgan topshiriqlarni bajarishi) va bq.)da foydalaniladi. Quyida g‘arbda korpusga asoslangan tarjima yo‘nalishining paydo bo‘lishi va rivojlanishi haqida to‘xtalamiz.

2001-yilda korpusga asoslangan yo‘nalish mashhurlikka erishib u tarjimashunoslikda yangi paradigma sifatida e’tirof etildi. Aslida bu yo‘nalish dastlab 1980-yillarning boshlarida Buyuk Britaniyada John Sinclair va uning jamoasi tomonidan COBUILD ingliz tili loyihasi doirasida ishlab chiqilgan bir tilli korpus lingvistikasi texnikasi va usullaridan foydalana boshlagan edi. Kompyuter tiziming tezkor tarraqiyoti tabiiy matnlar (tadqiqotchilar tomonidan suniy yaratilmagan balki muloqot maqsadida yuzaga kelgan)dan iborat elektron korpus yaratish imkonini berdi. Kompyuter korpusidan foydalanishning asosiy sababi o‘sha vaqtgacha asosan tahlilchining intuitsiyasiga bog‘liq bo‘lib kelgan lingvistik tahlillar sifati va leksik birliklar, so‘z birikmalarning tipik qo‘llanilishini tog‘ri aniqlash bilan bog‘liq.

M.Beyker o‘zining «Korpus lingvistikasi va tarjimashunoslik» asarida tarjimashunoslik tadqiqotlarida kompyuter korpusidan foydalanishga urg‘u berib tipiklik (typicality) konsepsiyasini norma, qoida va unversallik konsepsiyalari bilan bog‘laydi. M.Beyker asosiy e’tiborini tarjima matnlaridan iborat korpus tili tipik xususiyatlarini aniqlash hamda buni tarjima qilinmagan matnlar bilan taqqoslashga qaratadi. Natijada tarjima jarayoni va ishda belgilangan normalar sababli farqlarni yuzaga keltiruvchi elementlarni aniqlaydi. M.Beyker tarjimaning xarakter xususiyatlari sifatida tushunarlilik, grammatik standartlashtirish va «say» (demoq) kabi odatiy so‘zlarning ko‘p qo‘llanilishini ko‘rsatib o‘tadi [1, 49]. Bunga o‘xshash gipoteza axborot texnologiyalariga asoslangan davrdan oldingi paytlarda ham ilgari surilgan. Masalan, J.Levyning fikricha, tarjima grammatik to‘g‘ri, lekin ko‘p qo‘llaniladigan so‘zlardan iborat deb xarakterlanadi [5, 148]. Blum-Kulka va Levenston leksik soddalashtirish tarjimaning tipik xususiyati deb hisoblashsa, Vinay va Darbelnet tarjima jarayonini umumiylashtirib, tarjima matni asliyat matnidan uzun bo‘lishini ta’kidlashadi [6, 94]. Katta hajmli kompyuterlashtirilgan ma’lumotlar bazasining yaratilishi bunday gipotezalarni ko‘plab matnlarda tekshirib ko‘rish imkonini berdi.

Laviosaning «Korpusga asoslangan yo‘nalish: tarjimashunoslikdagi yangi paradigama» asari ikkiga: nazariy metodologik masalalar va amaliy tadqiqotlarda yangi korpusga asoslangan dasturlardan foydalanish masalalariga bo‘lingan. Bu ilmiy ish nashr qilinganidan buyon texnologiyaning tezkor taraqiy etishi hamda

katta hajmda elekton matnlarning yuzaga kelishi yuqorida tilga olingan ikki masalani rivojlantirdi, ammo bu masalalar umumiy qabul qilingan tadqiqot metodologiyasiga aylanmadi. Chunki, metodologiya o‘z-o‘zidan tadqiqotning obyektiga bog‘liq bo‘lib, tarjima nazariyasi ham sof leksikografik loyihalardan maqsadiga ko‘ra farqlanadi. Bu yerda asosiy masala korpusning turiga bog‘liq bo‘lishi mumkin. S.Bernadini va bq.larning «Tarjimani o‘rgatishda korpus: Kirish» (2003) asarining «Tarjimonlarni tayyorlashda korpusdan foydalanish» bo‘limida «bu sohada terminlar barqaror emas»ligini ta’kidlab, korpus tipologiyasi va uning har bir turidan foydalanish haqida o‘z xulosalarini berishgan. Asarda korpusning 3 turi muhokama qilingan:

1. **Bir tilli korpus.** Bu korpusdan tarjimon o‘zi tarjima qilgan so‘z birikmalari va boshqa til birliklarini to‘g‘ri va tabiiy chiqayotganligini tekshirishda foydalanadi. Ingliz tiliga tarjima qilayotganlar uchun Britaniya milliy korpusi (British National Corpus), Ingliz tili Kobild banki (Cobuild Bank of English) kabilar ma’lumotlar zahirasi bo‘lib xizmat qiladi.

2. **Qiyosiy ikki tilli korpus**da asliyat tillarining bir biriga o‘xshash matnlari to‘planadi. Bunday korpus ikki tildagi bir sohaga oid matnlar ichidan tegishli termin va ekvivalentlarni aniqlashda juda samaralidir.

3. **Parallel korpus** asliyat va tarjima matnlaridan iborat bo‘lib, tarjimon tomonidan qo‘llanilgan metodlarni aniqlash va turli tarjimashunoslik tadqiqotlarini olib borish imkonini beradi.

Muhim jihati S.Bernadini «parallel korpus til o‘rganuvchilar va tadqiqotchilarga ikki tilda ham tarjima ta’siri ostida yaratilgan matnlarning xususiyatlarini taqqoslash imkonini berishi»ni ta’kidlab o‘tadi [2, 6], ya’ni tarjima matnida ko‘p qo‘llanilgan leksik yoki grammatik birliklar belgilanadi, so‘ngra asliyat tilida bunday xususiyatlar bor yoki yo‘qligi aniqlanadi. Masalan, Olohan va Beyker (2000) Ingliz tili tarjima korpusi (Translational English Corpus — TEC)da inglizcha «that» nisbiy olmoshi qo‘llanilishini tekshirib ko‘rishadi va uni Britaniya milliy korpusi bilan taqqoslashadi. Tadqiqot natijasiga ko‘ra og‘zaki uslub ko‘rsakichi sifatida qo‘llanilganda nisbiy olmosh odatiy holatda tushirib qoldirilgan, boshqa tomondan tarjima korpusida nisbiy olmosh ko‘p qo‘llanilgan, bu holat tarjima tilining ta’siri va xususiyati deb izohlangan.

Maeve Olohanning «Tarjimashunoslikdagi korpusga kirish» (Introducing Corpus in Translation Studies, 2004) asarida tadqiqotlarning bu sohasi haqida so‘ngi qarashlar berilgan va sintaktik keys tahlilar hamda boshqa jihatlar ham qamrab olgan. Olohanning tadqiqotlarida asosan Ingliz tili tarjima korpusiga e’tibor qaratilgan bo‘lib, u parallel korpus tahlillarini bera oladigan «Wordsmith Tools» (2007) tijoriy dasturidan foydalanadi. Bunda asliyat va tarjima matnlari elektron formatda bo‘lib, nusxa ko‘chirish ruxsati berilgan. Dastur yordamida miqdor (asliyat va tarjima matnlarida so‘zlarning qo‘llanilish darajasi, gaplarning uzunligi, kalit so‘zlarni aniqlash va bq.) va sifat (matnda aynan o‘zaro mos keladigan gaplar qatorini aniqlash) tahlillari olib borilgan. Bunday metodlar

korpusga asoslangan yo‘nalishni boshqa metodologiya va yondashuvlar bilan bog‘lab, tarjima mahsulotini o‘rganish, tarjimaning tipik xususiyatlarini aniqlashga qiziqishni oshiradi.

Yuqoridagi tahlillar g‘arbda korpusga asoslangan tarjima yo‘nalishi tarjimashunoslik sohasida allaqachon o‘z o‘rnini topganini va undan samarali foydalanilayotganligini ko‘rsatib turibdi. Shundan kelib chiqib o‘zbek tilshunosligi va tarjimashunosligi oldida zudlik bilan hal etilishi lozim vazifalar sifatida quyidagilarni sanab o‘tish maqsadga muvofiq:

1. Kompyuter lingvistikasi keng qamrovli soha bo‘lib, bu soha mutaxasisi ayni vaqtda tilshunos hamda axborot texnologiyalari mutaxasisi bo‘lishi talab etiladi. Shu sababli yurtimizning tegishli oliy ta’lim muassasalarida kompyuter lingvistikasi magistratura mutaxasisligini o‘qitishni tashkil etish va kadrlar tayyorlash lozim (Yevropa Ittifoqining Erasmus+ dasturi asosida O‘zbekiston va Qozog‘iston Respublikalari Oliy ta’lim muassasalarida kompyuter lingvistikasi bo‘yicha magistratura dasturini ishlab chiqish va sohani rivojlantirish (CLASS project) loyihasining Urganch davlat universiteti, Samarqand davlat chet tillar instituti, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti va O‘zbekiston milliy universitetida olib borilayotganligi quvonarli);

2. O‘zbek tili milliy korpusini yaratish (o‘zbek tilidagi turli uslubdagi matnlar, masalan badiiy, publisistik, ilmiy, internet sahifalari matnlari va bq. kabi ko‘plab matnlarni to‘plash va ularni kompyuter tanishi uchun kodlashtirish talab etiladi);

3. O‘zbek tilida avtomatik dasturlarini yaratish (Stanford CoreNLP, FreeLing, LinguaKit, UDPipe kabi tabiiy tilni kodlashtirish dasturlarini yaratish).

Xulosa sifatida shuni aytish mumkinki, milliy korpus ega bo‘lish faqatgina tarjimon va tilshunoslarga foydali bo‘lmasdan, balki o‘zbek tilining xalqaro miqyosda obro‘sini oshiradi, o‘zbek tilining davlat tilidan xalqaro til sifatida o‘z o‘rniga ega bo‘lishini kafolatlaydi.

ADABIYOTLAR:

1. Baker, M. (1993) ‘Corpus linguistics and translation studies: implications and applications’ // Korpus lingvistikasi va tarjimashunoslik.
2. Bernadini, S., D. Stewart and F. Zanettin ‘Corpora in translation education: an introduction’// Tarjimani o‘rgatishda korpus, 1–14 b.
3. J.Munday Introducing translation studies: theories and applications. Routlidge Publishing, New York, 2008 // Tarjimashunsolikka kirish: nazariya va amaliyot.
4. Laviosa, S. (1998a) ‘The corpus-based approach: a new paradigm in translation studies’, Meta 13.4: 474–9 // Korpusga asoslangan yo‘nalish: tarjimashunoslikdagi yangi paradigma.
5. Levý, J. (1967/2000) ‘Translation as a decision process’, in L. Venuti (ed.) (2000): 148–59 // Tarjima qaror qilish jarayoni sifatida.
6. Vinay, J.-P. and J. Darbelnet (1958, 2nd edition 1977) Comparative Stylistics of French and English // Fransuz va ingliz tillarining qiyosiy stilistikasi.
7. http://nlp.lsi.upc.edu/freeling/demo/demo.php

O‘ZBEK BOLALAR SHOIRLARINING IJODI BO‘YICHA KORPUS LINGVISTIKASINI YARATISH AHAMIYATI

B. Jamilova, Buxoro davlat universiteti
jamilova_11@mail.ru

Maqolada o‘zbek bolalar shoirlarining ijodi, bolalar she’riyati namunalari bo‘yicha o‘zbek lingvistik korpusini yaratish haqida fikr yuritilgan. Ayniqsa, bolalar adabiyoti namunalarining manbalari, o‘zbek tilining yoshlar ishlatadigan asosiy lug‘atlari, qisqa akademik grammatikasi kompyuterga kiritilib, elektron shaklga keltirilishining dolzarb ahamiyati ta’kidlangan.

Tayanch so‘zlar: *konkordansiya, lemmatizatsiyasi, lingvistika, korpus, tashbih, poetik ko‘chim, akademik grammatika, badiiy matn, elektron lug‘at, mashina tarjima, chastotali lug‘at.*

ВАЖНОСТЬ СОЗДАНИЯ КОРПУСА С ТВОРЧЕСТВОМ УЗБЕКСКИХ ДЕТСКИХ ПОЭТОВ

Б. Жамилова, Бухарский государственный университет,
Бухара, Узбекистан, e-mail: jamilova_11@mail.ru

В статье рассматривается идея создания узбекского лингвистического корпуса на основе творчества узбекских детских поэтов и образцов детской поэзии. Особенно подчеркивается важное значение источников образцов детской литературы, часто используемых молодежью узбекских слов, а также введение краткой академической грамматики в компьютер и превращение ее в электронную форму.

Ключевые слова: *конкорданция, лемматизация, лингвистика, корпус, академическая грамматика, художесвенный текст, электронный словарь, автоматический перевод, частотный словарь.*

IMPORTANCE OF CREATING CORPUS LINGUISTICS IN THE FRAME OF CREATIVITY UZBEK CHILDREN’ POETS

B. Jamilova, Bukhara State University,
Buhara, Uzbekistan, jamilova_11@mail.ru

The article considers the idea of creating an uzbek linguistic corpus based on the works of children’s poets and children’s poetry samples. The special

emphasis is put on the important role of the sources of children's literature samples, uzbek vocabulary frequently used by youngsters and insertion of the concise academic grammar into the computers and so transforming them into electronic forms.

Key words: *concordance, lemmatization, linguistics, corpus, academic grammar, literary text, electronic dictionary, automatic translation, frequency dictionary.*

O‘tgan asr boshlari xalqimiz, ma`naviyatimiz, tilimiz va qadriyatlarimiz uchun muhim davr hisoblanadi. Jumladan, yurt kelajagi sanalmish yosh avlod tarbiyasi, ularning qunt bilan bilim olishi, til o‘rganishiga alohida e’tibor qaratilgan. Binobarin, yoshlarning o‘z ona tillarini asrash bilan birga boshqa tillarni ham puxta o‘rganishi istiqlol fidoyilarining yetakchi maqsadlaridan edi. Bugun farzandlarimiz erishayotgan baxtni roppa-rosa bir asr ilgari ma`rifatparvarlik harakati boshida turgan siymolar naqadar orzu qilgan, deyish mumkin. Ular ham maktab, ham madrasalardagi ta’lim tizimini, ularda o‘qitiladigan fanlarni, darsliklarni sharqona an’analarni saqlagan holda isloh qilish yo‘lidan borib, "usuli jadid" maktablarining milliy-madaniy taraqqiyotimizda muhim omil bo‘lishini asoslab bergandi. Chunki, o‘sha vaqtlardagi boshlang‘ich sinf o‘quvchilari o‘qish darsligidagi matnlar fors-tojikcha bo‘lib, u turkiy tilda so‘zlashuvchi bolalarning savod chiqarishi, mazmunini tushunishlarini ancha qiyinlashtirgan. Shu sababli, Mahmudxo‘ja Behbudiy «Oyna» jurnalining 1913-yil, 1-sonidagi «Ikki emas, to‘rt til lozim» maqolasida: «Biz turkistoniylarga turkiy, forsiy, arabiy va rusiy bilmoq lozimdur», deb yozgan edi. Ya’ni, o‘z ona tilisi hisoblangan turkiy tilni bilishi zarurligi holda, adabiyot va san`at tili sanalgan arab, fors tillari yoniga dunyo madaniyati, ilm-fani, taraqqiyoti bilan bo‘ylasha olish uchun rus tilini ham egallash joizligi ta`kidlangan. Demak, Behbudiy o‘z vaqtidayoq yoshlarni ko‘proq til o‘rganishga da`vat etgan edi.

XXI asr boshlariga kelib til o‘rganish tag‘in dolzarb muammoga aylanayotgani bejiz emas. Bugun nafaqat o‘zga tillarni o‘rganishimiz, balki o‘z tilimizni dunyoga tanitishimiz uchun ham qayg‘urishimiz kerak. Ayniqsa, ona tilimizning qonun-qoidalari, rang-barang tovlanishlari-yu badiiy jilolari asosida korpus yaratish nihoyatda muhim vazifa. Chunki butun dunyoda o‘zbek tilidagi so‘zlarning jarangi yayrashi uchun, uning bazasini ham tilimizning o‘zi kabi jozibali yaratish maqsadga muvofiq, deb o‘ylaymiz. Negaki, o‘zbek tilining badiiy- tasviriy vositalari asosida yaratilgan tilni modellashtirish, indekslashda matndagi barcha ma’no qirralari namoyon bo‘lishiga erishish lozim. Bu ham hozirgi texnik globallashuv jarayonida tilni asrashdek bir gap.

Matbuot va maorif ravnaqiga sezilarli hissa qo‘shgan taniqli ma’rifatparvar adib Abdulla Avloniy ona tilini asrash va uni har xil qorishiqlikdan tozalashga bag‘ishlangan «Hifzi lison» maqolasida ham ona tilining sofligini saqlash haqida

fikr yuritgan edi. Jadal rivojlanayotgan texnika asrida ayniqsa yoshlar har bir yangilikka chanqoq. Ular ongiga milliy til xususiyatlarini singdirish bir muncha qulay. Shu ma'noda, bolalar adabiyoti namunalarining manbalari, o'zbek tilining yoshlar ishlatadigan asosiy lug'atlari, qisqa akademik grammatikasi kompyuterga kiritilib, elektron shaklga keltirilsa, jumladan, bolalar she'riyati, badiiy adabiyotiga oid matnlar korpusi yaratilishi bu sohaning yetakchi muammolaridan biridir.

Oldimizda turgan muhim vazifani shundan ham anglash mumkinki, bugun deyarli ko'pgina dunyo tillarining lingvistik korpusi mavjudligi bu tildagi elektron lug'at va mashina tarjimalar xizmatini ommalashtirmoqda.

Binobarin, endi ona tilimizning texnika vositalari orqali ham badiiy sayqallash haqida qayg'urish zarur. Aytaylik, bolalar she'riyatida M.A'zamning ona tili haqidagi qator she'rlari bor. Shoirning «Til darsi» she'rini falsafiy, ta'lim-tarbiyaviy jihatdan jadidlar orzu qilgan o'zlikka qaytish, deyish mumkin. 1975-yilda yozilgan bu she'r go'yo, dastlabki qadam edi. She'rdan anglashilishicha, lirik qahramon gruzin bolasidan katta saboq oladi.

– Gruzin tilini bilasizmi? – dedi,
– Gruzin tilini bilmayman, – dedim.
– Gruzin tilini o'rgataymi? – dedi.
– O'rgat, – dedim.

Fiala ismli bola «mehmon»ga o'z ismi, qishlog'ining nomini faxr bilan aytib, darrov o'z tilini o'rgata ketadi. Suhbatdosh bir pasda boladan «aka», «opa», «qo'l», «oyoq», «men», «sen», «kitob» kabi so'zlarning gruzinchasini bilib oladi; ayni paytda o'zi ham Fialaga «uy», «xalq», «ko'cha» degan o'zbekcha so'zlarni o'rgatayotganda manziliga yetib keladi. Ammo, ana shu qisqa muddatda lirik qahramon — kichkintoy boladan bir jahon saboq oladi. Uning o'z tiliga munosabati, iftixori ongu shuurini sergaklantiradi:

Lekin qulog'imda qoldi, so'zlarining qo'ng'irog'i,
Ko'zlarimda, «bildingizmi»,– deganday,
Ko'zlarining jovdirashi…

She'r rosmana dars ifodasi bo'la olganidan tashqari, savol-topshiriq ham beriladi:

– Bolalar! Siz shunday qilasizmi hech?
Sizning shahringizga, qishlog'ingizga,
Mehmon kelsa, uzoq ellardan,
Siz o'zbek tilini o'rgatasizmi?
O'zbek so'zlarining jarangini,
Mehmon qulog'ida yangratasizmi?

Afsuski, bu savolga she’r yozilgan vaqtda «ha» deya javob bera ololmasdik. Shu sababli, «dars»ning ta`sirchanligi naqadar yuqoriligini mustaqillik yillarida his qilayotirmiz. Unda qo‘yilgan muammoning dolzarbligi hamon pasaygani yo‘q.

M. A’zam mustaqillik yillarida bu mavzuga tag‘in alohida e’tibor qaratdi. Uning yosh lirik qahramoni endilikda ona tilisini boshqalardan kam sevmaydi:

Ona tilim, turkiy tilim,
Ardoqligim o‘zbek tilim,
Avaylayman seni doim
Qaboq ila ko‘zdek, tilim.

To‘rt bandli she’rning har birida ona tili turli sifatlar bilan qo‘shib takrorlanadi. Bunda o‘zbek tilining **qadimliligi, ko‘rki, donoligi**, onaga qiyoslanadi hamda uni almashtirib bo‘lmasligi ta`kidlanadi. Shoir ona tilini avaylashni **ko‘zu qaboqqa** mengzashi ham bejiz emas. Ko‘z insonga berilgan ilohiy ne’matlar ichida eng ardoqlisi. U yorug‘ olam ma’nosini bildirishdan tashqari, inson botinini o‘zida namoyon etadi. Donishmandlar odamni ko‘ziga qarab, kimligini ayta olishgan. Ko‘zimizni asrash bilan nafaqat, hayotimiz mazmunini, demakki, o‘zligimizni, ko‘nglimizni ham asragan bo‘lamiz. Xuddi shu xususiyat ona tiliga ham xosligini shoir:

Turkiy tilim, ko‘rkli tilim,
O‘zbek tilim, ko‘zdek tilim,

misralarida jo etgan. Muhimi, bolalarga yoshligidanoq o‘z ona tillarini sevishga da’vat etish ana shunday satrlar ta’sirida ommalashadi.

Ko‘rinadiki, o‘zbek tili poetik ko‘chimlarga boy. Bu she’rni tushunmoqchi bo‘lgan o‘zga tilli o‘quvchi ana shu sifatlarning asl ma’nosini hamda matndagi ma’nosini ilg‘ab olish kerak. Demakki, o‘zbek tilining milliy korpusini yaratishda ana shu ma’no qirralari to‘liq aks etsagina uning muqobil variant tanlanadi.

Bu she’rning ma’nosini teran anglash uchun shoir she’ridagi so‘zlarni konkordansiyalash usulidan foydalanish, ya’ni so‘zlarning chastotali lug‘atini tuzish zarur.

Ko‘rinadiki, M. A’zam ijodining chastotali lug‘atini yaratish va uni kompyuter korpusiga joylashtirish orqali birgina ona tili mavzusidagi she’rlarining g‘oyaviy dolzarbligini teran anglash mumkin. Masalan, ona tilim, turkiy tilim, o‘zbek tilim, ko‘zdek tilim, o‘z tilim, qadim tilim, gruzin tili kabi so‘zlarning morfologik tahlili – lemmatizatsiyasi orqali tashbehlarning ohorligi va ma’noviy o‘tkirligi, shoir aytmoqchi asl muddao yaxlit uyg‘unlashadi. U ijodida shunga o‘xshash so‘zlarni 100 marta, yoki 1000 marta ishlatganiga qarab,

mavzuning shoir ijodidagi o‘rnini, dolzarbligini, hatto, 1975–2015-yillarda yozilgan she’rlarida shu mavzuga, so‘zga necha marta murojaat qilgani yaqqol oydinlashadi. So‘ngra, bu umuman bolalar shoirlari ijodi misolida qiyoslanib, xulosa qilishga asos bo‘ladi. Negaki, Istiqlol yillarida bu mavzu o‘zbek bolalar she’riyatida turli ohang va mazmunda jilolangani ma’lum. Jumladan, T.Adashboyev she’rining jajji lirik qahramoni, ona tilini «jahon bo‘ylab dovrug‘ solgan», Buyuk Temur, Mir Alisher, Bobur Mirzo she’rlaridan rang va qiyos olgan»ligini ta’kidlab:

Shunday tildan tonar bo‘lsam,
Qiyma-qiyma bo‘lsin tilim
Alla bo‘lib jarang olgan,
Ona tilim – jon-u dilim,

– deya qasamyod qiladi.

Yoki, Kavsar Turdiyevaning shu turkumdagi she’rlari «O‘zbek tili — tilimiz, unda so‘zlar elimiz» deb nomlanadi:

Onajonim allasi,
Opamlarning yallasi,
Singlim yozgan husnixat,
Akam olgan tabrik xat

Bizlarni yig‘ib boya,
Bobom aytgan hikoya,
Men yodlagan go‘zal she’r,
O‘zbek tili go‘zal der.

Darhaqiqat, o‘zbek tili o‘zining boy tarixiga ega, ulug‘ mutafakkir Alisher Navoiy davrida u yanada kamol topdi. Uning rang-barang qirralarini o‘rganishga e’tibor istiqlol yillarida kuchaydi. Zotan, K.Turdiyeva ijodidan o‘rin olgan bu turkumdagi she’rlarning paydo bo‘lishi ham davr taqozosi va ehtiyojidir, albatta.

Binobarin, bu o‘zbek korpus lingvistikasining maxsus turi (janr, uslub, davr) hisoblanib, sinxron – ayni jarayonda qo‘llanilayotgan nutqiy birliklarning tahlil qilishda bolalar adabiyoti, she’riyati korpusini jadal shakllantirishga e’tibor zarurligini unutmasligimiz, bugun yaratilayotgan har qanday yangilik eng avvalo kelajak avlod uchun xizmat qilishini, demakki bolalar uchun ijod qiladigan yetakchi adib va shoirlar ijodini qamrab olish muhim omil hisoblanadi.

ADABIYOTLAR:

1. Zaxarov V.P., Bogdanova S.Yu. Korpusnaya lingvistika. Uchebnik. Irkutsk. IGLU, 2011
2. Po‘latov A., Muhammedova S. Kompyuter lingvistikasi. T., 2001.
3. N. Abdurahmonova N. Kompyuter lingvistikasi. T., 2015
4. Miraziz A’zam. Saylanma. T., 2012
5. K.Turdiyeva. Dunyoni saqlar bolalar.Toshkent, 2012. –B.132.

A METRICAL ANALYSIS OF MEDIEVAL GERMAN POETRY THROUGH CORPUS LINGUISTICS

M. Qochqorova, Uzbek state world languages university
Tashkent, Uzbekistan, maftuna9008@gmail.com

This paper has displayed a new approach to deal with an extremely old issue for medieval German poetry. Examining the meter of this custom postures one of a kind difficulties to abstract researchers, philologists, and computational etymologists alike. By developing a directed model of the meter, this paper exhibits the advantages of a quantitative vast examination empowering us to describe its eccentricities and recommend enhancements to the current academic methodology.

Key words: *computational linguisitics, literary data, machine learning, machine translation, post-editing, rhetorical figures, themes and motifs in poetry, Middle High German epic poetry.*

СТАТИСТИЧЕСКИЙ АНАЛИЗ СРЕДНЕВЕКОВОЙ НЕМЕЦКОЙ ПОЭЗИИ С ПОМОЩЬЮ КОРПУСНОЙ ЛИНГВИСТИКИ

М. Кочкорова, Узбекский государственный университет мировых языков,
Ташкент, Узбекистан, maftuna9008@gmail.com

В статье описан новый подход к исследованию средневековой немецкой поэзии. Изучение метрики этой литературы создает проблемы для литературоведов, филологов и компьютерных лингвистов. Развивая модель прямой метрики, эта статья демонстрирует преимущества квантитативного анализа всего контента, позволяя нам охарактеризовать его особенности и предложить улучшения текущей академической методологии.

Ключевые слова: *вычислительная лингвистика, литературные данные, машинное обучение, машинный перевод, пост-редактирование, риторические фигуры, темы и мотивы в поэзии, средневекавая эпическая поэзия.*

KORPUSGA ASOSLANGAN O‘RTA ASR NEMIS SHE’RIYATINING METRIK ANALIZI

M. Qo‘chqorova, O‘zbekiston Davlat jahon tillari universiteti
Tashkent, Uzbekistan, maftuna9008@gmail.com

Ushbu maqola o‘rta asr nemis sheriyatidagi muammolarga yangi yondashuvni taqdim etadi. Ushbu an’anani o‘rganish adabiyotshunoslarga, filologlarga va hisoblash lingvistlariga ayrim muammolarni keltirib chiqaradi. Nazorat modelini o‘rnatish orqali ushbu maqolada o‘zida mavjud bo‘lgan pedagogik yondashuvni takomillashtirishni taklif qiladigan va uning xususiyatlarini tavsiflashga imkon beradigan, miqdoriy korpusga asoslangan tahlilning afzalliklari ko‘rsatilgan.

Kalit so‘zlar: *hisoblash lingvistikasi, adabiy ma’lumotlar, mashinani o‘rganish, mashina tarjimasi, tahrirlash, retorik figuralar, she'riyatdagi mavzular va motivlar, o‘rta asrlar nemis epik she'riyati.*

Middle High German (MHG) epic verse introduces a one of a kind answer for the etymological changes supporting the progress from traditional Latin verse, in light of syllable length, into later vernacular cadenced verse, in view of phonological pressure. The prevailing example in MHG refrain is the rotation among pushed and unstressed syllables, yet syllable length additionally assumes a pivotal job. There are an aggregate of eight conceivable metrical qualities. Single or half more syllables can convey any of three sorts of pressure, bringing about six blends. The seventh esteem is a twofold more, i.e., a since a long time ago focused on syllable. The eighth esteem is an omitted syllable. We develop a regulated Contingent Irregular Field (CRF) model to foresee the metrical estimation of syllables, and in this way research medieval German artists' utilization of semantic and resonating accentuation through meter. The highlights utilized are: (1) the syllable's situation inside the line, (2) the syllable's length in characters, (3) the syllable's characters, (4) elision (last two characters of past syllable and initial two characters of central syllable), (5) syllable weight, and (6) word limits. Extra metrical standards are upheld and peripheral probabilities are ascertained to yield the undoubtedly lawful scansion of a line. The model accomplishes a weighted normal F-score of 0.925 on inner cross-approval and 0.909 on held-out testing information. We establish that trochaic variation with a one syllable anacrusis and words conveying clear pressure task are the least demanding for the model to check. Lines with numerous twofold morae of syllables with few characters are the most troublesome. We at that point rank all the epic verse in the Mittelhochdeutsche Begriffsdatenbank (MHDBDB) by the trouble of the meter. At last, we examine the twofold mora, which MHG artists

used to attract regard for picked ideas. We infer that artists by and large utilized the twofold mora to underline exceedingly sonorant words.

Lovely meter in the Middle High German (MHG) convention has dependably been a quarrelsome and complex subject, as it requires a nuanced information of MHG writing, a solid comprehension of MHG phonetics, especially phonology, and learning of the melodic practices of the period1. Most work so far has not possessed the capacity to ace these areas2. While this paper does not endeavor to completely join these different fields, it seeks to take cautious thought of each in building up a computational model to all the more likely see how medieval German writers created their words into meter, and thus help us in our own perusing of the content. The expanded fame of machine learning calculations and their application to literary information shows an especially productive open door in an area that has tormented MHG grant for quite a long time. Rather than a deductive methodology, i.e., starting with the supposition of trochaic rotation as essential, administered learning takes into consideration a substantial scale inductive methodology, providing the calculation with an abundance of particular precedents from which general standards can be observed. Urgently, the objective of any such model isn't to set up a flat out truth about a recorded dialect; the objective is to consequently imitate the comment choices of researchers on a huge scale. Clarifying the whole MHG epic corpus would enable us to all the more likely see any standards that do exist and in addition the difficulties a specific content stances. Programmed explanation would likewise bolster an expansive scale examination on how particular metrical qualities and meter composes are summoned in various settings. Researchers frequently talk about how changes in meter, metrical qualities, or particular rhythms are activated in particular scenes, however would we be able to quantify this intricacy? MHG meter accommodates interesting adaptability in accentuation, however did creators have inclinations for various metrical qualities? Are sure messages or entries deliberately made to be more hard to filter? This paper looks to answer these inquiries and others through a vast scale examination of naturally checked verse.

While late twentieth century grant dismissed meter principally because of hypothetical differences and an absence of composition confirm, Christoph März as of late re-confined MHG grant on meter in his article "Metrik, eine Wissenschaft zwischen Zählen und Schwärmen," in which he endeavors to resuscitate a meter-based formal point of view (März, 1999). As per März, shape has two vital capacities and openings: it reminds us, and it takes into consideration correlation (März, 1999, p. 325). Both of these perceptions give inspiration to the accompanying investigation. Beautiful meter acts not exclusively to help the memory of an entertainer or author, yet in addition influences the group of onlookers, provoking this relative gathering. März composes:

I recall the experience that when you try to remember a poem, you often only remember the pattern—a few words may come along with that pattern or not.

Also, if you forget parts of the text, the threads can be found again in certain passages by humming the rhythm of the verse (März, 1999, p. 325).

This demonstration of recollecting fills in as a chance to distinguish associations among tunes and messages (both formally and semantically), and look at writings, as März would have it. This examination, when perceived by an entertainer or crowd, can produce and add significance to a sonnet or tune. Particularly in the MHG custom, an association among frame and substance has dependably been assumed. However März is additionally keen on bring down level associations and references inside kinds. März asks whether these preoccupied metrical schemata "transport" particular thoughts, and assuming this is the case, how they are made (März, 1999, p. 325). Klaus Kohrs made a comparative inquiry decades sooner. Kohrs clarifies in Saussurian terms how meter itself can add implication to dialect, which it doesn't innately convey: "With the metrical, that is even "semi melodic" development of dialect as a representative and sonoric wonder the side of the signifié is semi sublimated, i.e., sensical and semantic references wind up harmful, which the "normal" dialect does not have and does not need" (Kohrs, 1969, p. 605). Hugo Kuhn displays the thought comparably in connection to music and tune, however accentuates its "Gebrauchsfunktion" (utilize work), i.e., the utilization cases for these fine arts, as folksongs, religious uses, for the court, knights, and so forth. (Kuhn, 1969, p. 38). This point is taken up by Thomas Cramer, addressing what the genuine Gebrauchsfunktion for these fine arts was, and whether our thoughts of them are right as indicated by the sources (Kuhn, 1969, p. 39). In any case, März significantly reshapes this inquiry, rather than asking what importance or capacity lovely meter may contain, he takes note of that meter is constantly decided moderately (März, 1999, p. 325). As Paul Zumthor and Ferdinand de Saussure have guaranteed about words and sound, there is no significance in the base component itself, just in setting and example. Be that as it may, for the two words and lovely meter, this setting must be stretched out past the contained question of a line of verse to the assemblage of referential articles.

The point of this paper is to disambiguate these relative connections. This venture does not expect to contend that a specific metrical hypothesis is without blame, nor that particular metrical qualities even exist thusly, but instead that executing any system unavoidably coaxes out relative contrasts inside a corpus. Heusler denounces the nineteenth century philologists for adjusting the content and making a measurable examination of MHG measurements inconceivable, and consequently he gives no insights in his MHG investigation of meter (Heusler, 1956, p. 4). However the concentration here isn't concrete, as far as outright numbers or measurements, yet rather in setting up relative connections between

writings, which, when accumulated over an entire content or corpus, won't overwhelm clear characteristics.

The conveyance of Latin into unmistakable provincial tongues had significant semantic and abstract ramifications for all of Europe. One striking outcome was on verse with quantitative meter. Indeed, even before the Middle Ages, the syllable length of established Latin had been almost overlooked in the vernacular4. Latin verse had utilized quantitative meter, in which syllable length is the arranging guideline. Syllable length was a phonologically unmistakable element to Latin speakers. Be that as it may, the rising lingos contrasted from Latin in that pressure turned into a phonologically essential element, and in this manner called subjective meter ("musical verse") prevailed in the Romance dialects. Accommodating these phonetic contrasts, MHG meter depended on both pressure and syllable length. This half and half metrical shape presents interesting difficulties to checking verse and took into consideration an assorted advancement in sort and style (Heusler, 1956, pp. 74– 75). However this flexibility brings up one of the primary issues and hypothetical issues in MHG explore on meter: not really Heusler's inquiry of "How am I to quantify it?," but instead what: in an arrangement of "estimated syllable refrain with free syllable tallies"— would could it be that we can tally, or shouldcount? (Heusler, 1956, pp. 9, 13) "What is countable in the section?" (März, 1999, pp. 323– 324) We could tally syllables, yet it isn't clear if the artists did this normally themselves, notwithstanding what the Meistersänger5 might want us to accept. Herbert Bögl portrays MHG stanza in his Abriss der mittelhochdeutschen Metrik: mit einem Übungsteil: MHG "exhibits in a unique dialect of images the succession of syllables in a refrain and measures them considering their length and stress" (Bögl, 2006, p. 9). It is this "considering" presents a troublesome computational issue for investigation MHG meter, in that strict guidelines for length and stress can't simply be utilized. A widely cited poem displaying the shift from quantitative to qualitative rhythmic poetry in the Latin tradition is Bishop Auspicius of Toul's late fifth century letter to Arbogast, the Count of Trier, imitating the iambic dimeter already made famous by Ambrose. The letter begins:

Praecelso exspectabili his Arbogasti comiti
Auspicius qui diligo salutem dico plurimam.

The first hemistich —◡—◡—◡ shows that a quantitative scansion would be ill-fitted to the rest of the verse, and that a strictly iambic scansion is preferred with a paroxytone in the cadence. Much Latin poetry followed suit, and the medieval Codex buranusfamously bears witness to the intermingling of Latin and MHG rhythmic verse, clearly demonstrating that they were drawing from the same rhythmical schemata. Germanic verse, on the other hand, did not originally follow the quantitative meter of antiquity, preferring organization according to alliteration and stress. In fact, Heusler calls alliteration the «Hausmarke» (house brand) of the Germanic language family (Heusler, 1956, pp. 92–93). In addition

to alliteration, a further marker of Germanic verse is the Langzeile (long line), traditionally consisting of two Kurzzeilen (short lines), an Anvers (first half of the line) and Abvers (second half of the line) (Heusler, 1956, p. 100). While this tradition began earlier, a classic example of Germanic alliterative verse is the ninth century Bavarian Muspilli:

…sin tac piqueme, daz er touuan scal.
uuanta sar so sih diu sela in den sind arheuit,
enti si den lihhamun likkan lazzit,
so quimit ein heri fona himilzungalon,
daz andar fona pehhe: dar pagant siu umpi.
sorgen mac diu sela, unzi diu suona arget,
za uuederemo herie si gihalot uuerde.

This alliterative stanza ruled all through a large portion of OHG and proceeded with solid in the Nordic conventions. Around a similar time that the Muspilli was composed in the southeast, in the west Otfrid von Weißenburg in Alsace was starting to fuse attributes of Old French verse into his ninth exceptionally old High German (OHG) refrain. Otfrid's choice to consolidate end rhyme (alluded to as an entirely Romance dialect impact by Heusler) is the primary bore witness to occurrence of Germanic verse's break from the alliterative convention. Along these lines Otfrid is for the most part thought about the beginning stage for an investigation of current German refrain. Otfrid's Evangelienbuch turned into the model for this new Germanic section, however he held the Langzeile from the more established Germanic custom. Otfrid set up a significant number of the new metrical conceivable outcomes in rhythm (monosyllabic full, bisyllabic ringing, and trisyllabic ringing) saw in the MHG time frame (Heusler, 1956, p. 13). A great part of the effect on Otfrid's style originated from different works on religion, chivalrous stories, and charms recorded at the time15. Heusler contends that this opportunity in refrain came fundamentally from the congregation, particularly church tunes. Heusler states: "tune all the more effectively exploits the prosodic flexibility" (Heusler, 1956, p. 32). Concerning rhyme, for almost 300 years there was just match rhyme in the AABB shape, once in a while AAA, until around 1150 (Heusler, 1956, p. 12). Otfrid's rhyme started as unadulterated monosyllabic rhyme, and later formed into multi-syllable sound similarity and different composes (Heusler, 1956, p. 20). As the significance of rhyme developed, it ended up fundamental for the rhyming syllable to likewise convey complement (Heusler, 1956, p. 24,3 1). This new rhyme and highlight gave an elective intends to integrate stanzas, yet in addition introduced new flexibilities of estimating section, as rhyme expected syllables to identify with each other, something underscored by the contemporary musicologists (Heusler, 1956, p. 9). The type of the Ambrosian psalm is the nearest metrically to Otfrid. The best distinction exists in the development of the

line, where the syllable tally isn't sure, and partitioned lifts16 are inexhaustible (Heusler, 1956, p. 35).

Otfrid's establishing of the Germanic musical stanza was what Heusler calls a "Germanicizing" of the Romance rhyming refrain: free filling of sections with syllables, anacrusis, and more shifted rhythms (Heusler, 1956, p. 36). Heusler graphs the improvement of Germanic section and its impact from the Romance convention, especially in that the blending of alliterative and match rhyme stanza prompted the early Germanic free filling of feet. However metrical traditions existed in Otfrid's section. The last foot was still entirely monosyllabic and stanzas could go from four to ten syllables, yet were all the more frequently some place in the middle of (Heusler, 1956, p. 43). OHG section frequently had feet with a larger number of syllables than MHG in light of the fact that OHG words just had more average syllables17. Rather than MHG stanza, OHG refrain was more reliable with syllable length and term (Heusler, 1956, p. 56). In this sense, OHG stanza was a "go between" among Latin and alliterative refrain (Heusler, 1956, p. 63).

The most exhaustive and still referenced investigation of German meter is Andreas Heusler's three volume Deutsche Versgeschichte. Heusler's hypothesis has been scrutinized unendingly throughout the years, yet holds on as the acknowledged hypothesis for MHG meter today. März claims that as hesitant as we are to utilize Heusler's hypotheses, we utilize them in light of the fact that there is basically no better option (März, 1999, p. 318). While endeavors have been made to supplement or study Heusler's work, particularly the presence of the basic "Takt" (measure, as in music), it has demonstrated troublesome for elective hypotheses to escape worldly limitations. On the off chance that there is no "Takt," is there no foot, or stress variation? (März, 1999, p. 319) As März watches, huge numbers of the elective speculations don't vary altogether from Heusler's, just Franz Saran's "Schallanalyse" (acoustic investigation) is recommended by März as a conceivable choice to all the more likely join the genuine voice of the stanza (März, 1999, pp. 321– 322).

Also, the prototypical MHG epic section foot is two syllables long, a focused on syllable taken after by an unstressed syllable. Be that as it may, feet can likewise be filled by one or three syllables (Domanowski et al., 2009). In the event that a foot is filled by one syllable, the syllable must be phonologically substantial (containing a long vowel or completion in a consonant). On the off chance that the foot is filled by three syllables, either the initial two or the last two syllables are regularly phonologically light22.

It is in these a typical feet that the impact of quantitative meter, where syllable length is a key factor, ends up obvious in MHG refrain. The foot in a Vierheber must be somewhat re-imagined to represent this. Phonologically, syllable length is estimated in morae, a unit of time with the end goal that a short syllable has one mora and a long syllable has two morae (Fox, 2000)23. A foot in

MHG meter is all the more definitely characterized as having two morae, not really two syllables24. In reality the mora, not the syllable, has been known as the crucial unit of MHG stanza, despite the fact that the mora capacities diversely in this wonderful convention than in its phonological definition (Tervooren, 1979, p. 1). In the event that a foot has just a single syllable, the syllable must be substantial on the grounds that an overwhelming syllable is two morae and the MHG foot requires two morae. A light syllable can't be the main syllable in a foot, since it can't be two morae. On the off chance that a foot has three syllables, two are frequently light since half morae are regularly light syllables (the primary half mora of a couple should dependably be light), together shaping one mora25. The other syllable is examined as one mora, yielding the required two morae in the foot. To outline, a syllable can have one of three length esteems: mora, half mora, or twofold mora. A half mora must be phonologically light, and a twofold mora must be phonologically overwhelming. Phonological length is generally insignificant and any syllable can be one mora (Heusler, 1956, p. 111).

An entirely lead based way to deal with examining MHG epic meter was attempted by Friedrich Dimpel in 2004 (Dimpel, 2004a). As Dimpel's work is the main of its kind in this field, it merits extraordinary thought here. As a major aspect of his thesis and proceeding with work at the University of Erlangen, Dimpel built up an arrangement of instruments named ErMaStat (Erlanger-Mittelalter-Statistik), created particularly for MHG epic verse (Dimpel, 2004b). Albeit beyond any doubt to concede the inadequacies of such a methodology, the opening pages of first experience with ErMaStat uncover his stylometric expectations in making such a suite of apparatuses:

At whatever point one endeavors to approach abstract, insightful inquiries with quantitative procedures, at that point one must expect that writings from various writers (or diverse times of a writer's work) exhibit certain unmistakable attributes on a phonological, morphological, lexical, and linguistic level, which enable themselves to be caught quantitatively (Dimpel, 2004b).

Dimpel's rundown of factors include: (1) syllable, word, and line check, (2) vowel and consonant tallies, (3) work words (particular parts of discourse), (4) similar sounding word usage, sound similarity, and enjambment, (5) postfixes, (6) word frequencies, (7) prefixes, (8) regular words (a better estimation than word recurrence), (9) word mixes (guileless bigrams), and (10) a metrical investigation. He will likely model style, or qualities of style, keeping in mind the end goal to look at writings and gauge probabilities of works being composed by a similar writer.

Dimpel proceeds with three precedents. In the primary model, he takes four of the better known MHG legends: Parzival, Tristan, Wigalois, and Willehalm. Utilizing the factors above, he ascertains and midpoints importance esteems, demonstrating that Parzival and Willehalm, both composed by Wolfram von Eschenbach, do in reality have a lower level of measured expressive distinction in

respect to each other than to the works by different creators. Dimpel is likewise ready to decide the commitments from singular factors. Dimpel's second examination concerns the gathering of Wolfram's Parzival into parts and the postulation proposed by Elisabeth Karg-Gasterstädt of four diverse sound composes, following crafted by Eduard Sievers (Karg-Gasterstädt, 1925). Dimpel's ErMaStat bolsters Karg-Gasterstädt's speculation as a probability. His last precedent considers the date of origin of Hartmann von Aue's Iwein regarding Hartmann's Erec.

Dimpel approaches MHG meter by first programming for variation and after that progressively making principles to represent pressure. In spite of the fact that his work must be praised for its exactness and semantic commitment, it is a difficult errand, unyielding, and amazingly dialect particular. Our goal here isn't to copy his work, nor reject it. Or maybe, through directed learning we offer another way to deal with an old issue for MHG. It additionally gives a chance to the "drei-stufige" (three-level, i.e., representing optional pressure and twofold morae) scansion Dimpel has not yet endeavored, but rather notes is an extraordinary test to demonstrating MHG meter. There are additionally focal points quite compelling to humanists. A directed strategy will figure out how to examine more in the way of a human than an entirely govern based methodology would, maybe staying more genuine to the wonderful custom, and giving understanding into what presents challenges for human scanners. It additionally takes into account more prominent adaptability, and an opportunity to break down the prosody past the epic meter, and maybe even exposition (Dimpel, 2004b, 2015).

REFERENCES:

1. Giegerich, H. J. (1985). Metrical Phonology and Phonological Structure: German and English. Cambridge: Cambridge University Press, 1-301
2. Hartman, C. O. (1996). Virtual Muse: Experiments in Computer Poetry. Hanover, NH: Wesleyan University Press.
3. Hayes, B. (1989). Compensatory lengthening in moraic phonology. Linguist. Inq. 20, 253–306.
4. Hench, C. (2017). «Phonological soundscapes in medieval poetry,» in Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, (Berkeley, CA), 46–56.
5. Heusler, A. (1956). Deutsche Versgeschichte: Mit Einschluss Des Altenglischen Und Altnordischen Stabreimverses. Grundriss Der Germanischen Philologie 8. Berlin: W. De Gruyter.
6. Karg-Gasterstädt, E. (1925). Zur Entstehungsgeschichte des Parzival Vol. 2. M. Niemeyer.

7. Kohrs, K. H. (1969). Zum verhältnis von sprache und musik in den liedern neidharts von reuental. Deutsche Vierteljahrsschrift Literaturwissenschaft Geistesgeschichte 43:604.
8. Kuhn, H. (1969). Text Und Theorie. His Kleine Schriften, Bd. 2. Metzler: Stuttgart.
9. Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). «Conditional random fields: probabilistic models for segmenting and labeling sequence data,» in ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, (Morgan Kaufmann Publishers Inc. San Francisco, CA).
10. Longfellow, H. W. (1932). Poems, Including Evangeline, The Song of Hiawatha, The Courtship of Miles Standish, Tales of a Wayside Inn. No. 56. Modern library.

СЕКЦИЯ 4.
ОНТОЛОГИИ

UZBEK ONTOLOGY OF UZBEK LANGUAGE AS EXAMPLE OF ADJECTIVE

Nilufar Abdurakhmonova[1], Mirsaid Aripov[2],
[1] Tashkent State University of Uzbek language and literature,
[2] National university of Uzbekistan, Tashkent, Uzbekistan
abdurahmonova.1987@mail.ru, mirsaidaripov@mail.ruv

This paper is devoted to implemention of the work on adjective of Uzbek as grammatical category of protégé program. It is laso discussed the classification of morphological categories according to build ontology as example of Uzbek language.

Key words: *ontology, protégé program, the Uzbek language, grammatical categories.*

УЗБЕКСКАЯ ОНТОЛОГИЯ УЗБЕКСКОГО ЯЗЫКА НА ПРИМЕРЕ ПРИЛАГАТЕЛЬНОГО

Nilufar Abdurakhmonova[1], Mirsaid Aripov[2],
[1] Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, [2] Национальный университет Узбекистана имени Мирзо Улугбека,
Ташкент, Узбекистан, abdurahmonova.1987@mail.ru,
mirsaidaripov@mail.ruv

Данная статья посвящена описанию работы по созданию онтологии грамматической категории прилагательного узбекского языка в программе Protege. В статье описывается классификация морфологических категорий в рамках онтологии на примере узбекского языка.

Ключевые слова*: онтология, протеже, узбекский язык, грамматические категории.*

I. Introduction

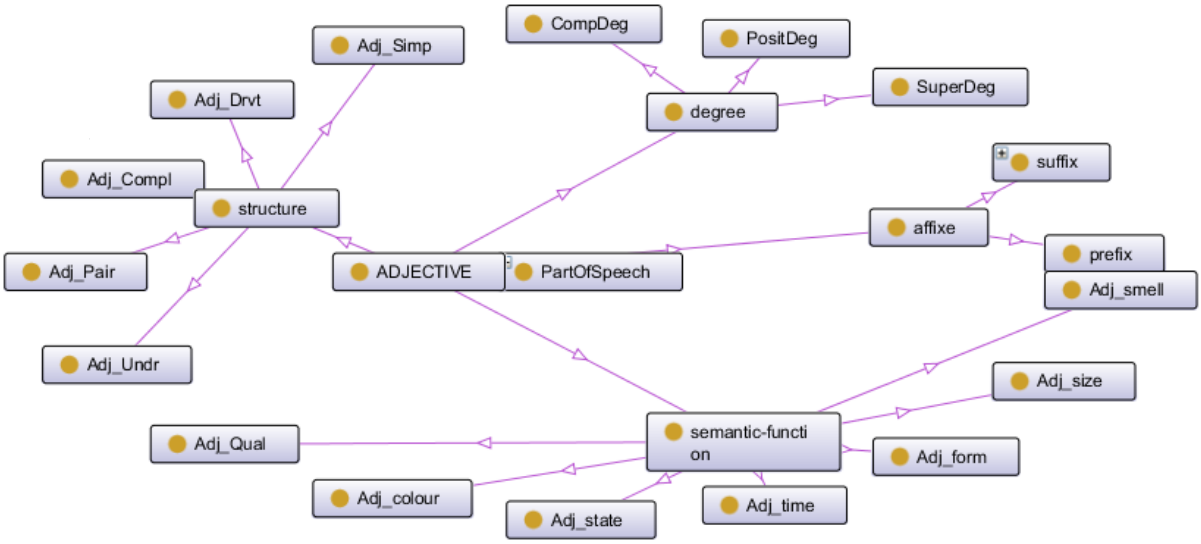
The paper is based on in frame of project «AP05132249 — Development electron thesaurus of Turkic languages for creating system of multilingual retrieval and extracting knowledge» (according to contract №132 at «12 » march 2018). This project focus to create ontology of grammar and thesaurus of Turkic languages like Uzbek, Kazakh, Turkish, Tatar and Kirgiz. Based on this project each part of speech of languages are modeled by protégé program that is used for Java and query can be easily got other application as well. Still there has not any unique model of Meta language of Turkic languages for machine purposes. Hence, it has some problems that seems a bit difficult for being very rich morfological comonents and different categories. Due to the growing digital contexts as web pages required to generate enormous information in database. On the other hand to clarify the text for any specific purpose for NLP, directed grammar rules and models are used because of richness affixes and exception of natural languages (even today there is corpora is used as object of language). «In the light of this growing context of digitalisation, diverse information representation and retrieval tools exist, which must be studied in addition to diverse fields of knowledge in which these tools have originated: Linguistics, Artificial Intelligence, Documentation, Linguistic Engineering... Hence, in specialised literature, analyses are performed on information representation and retrieval tools, taxonomies, classification systems, computational lexicons, lexical databases, thesauruses, titles lists, knowledge bases, conceptual maps, ontologies, synonym rings and semantic networks, among others».

Ontology is as the tool of IR and machine learning give us to clear information about the words and related meanings on a sphere.

The conceptual relations input in an ontology are particularly different of partonomy and taxonomy that belong to the field of knowledge to be structured as ontology and hierarchical clasisifican of any type of subject can ease to find information. Each property of components of grammar represented as sole set in order to parse or analyze a text correctly. «Ontology is about organizing, collecting and managing knowledge in an IRS. They have many similarities that they may both describe domain specific knowledge and contain terms and relations, however, building ontology for large domains is a costly affair and in many domains thesauri have been built».it can be seen that ontology is tool between macheni and user to share and reuse information to get resources.

2. Ontological model of adjective in Uzbek language.

Here the ontological model of Uzbek languages showed in the picture.



Adjective is reasonably easy than other part of speech in the stage of morphological analysis. Comparing other Turkic languages, it may reach conclusion that there are a lot of variety the types and classification of the adjective, however that is not very different.

Most of scholars agree that ontological model is rather comfortable to use that thesaurus. Why? For the reason that «While in thesauri only terms are related, RDFS introduces the notions of concepts and instances. A concept describes an entity on an abstract level with generic properties, whereas an instance is an actual representation of this concept with specific values of these properties».

Here we will offer to focus more semantic aspects of adjective as well. Peculiarly, some type of adjective in Uzbek may function as adverb even they have description of quality of noun. If taking one more examples, *«yaxshi qiz-yaxshi o‘qimoq», «to‘g‘ri yo‘l-to‘g‘ri gapirmoq», «chiroyli manzara-chiroyli raqs tusmoq»*. In English and Russian it differs, both of them like *правильный ответ-сказать правильно, true answer-tell truly*. Consequently, in ontological model should embrace all aspects of parts of speech for semantic analysis.

Tag	
Adj	Adjective
Adj_Simp	Simple
Adj_Undr	Underivation
Adj_Drvt	Derivation
Adj_Compl	Complex
Adj_Fus	Fusion
Adj_Comp	Compound

Adj_Qual	Qualitive
Adj_state	State
Adj_colour	Colour
Adj_form	Form
Adj_size	Size
Adj_taste	Taste
Adj_smell	Smell
Adj_time	Time

Conclusion

«Development electron thesaurus of Turkic languages for creating system of multilingual retrieval and extracting knowledge» as directed in our project is organized and presented, accounts for its usefulness or adequacy for a specific purpose. As globalization becomes more pervasive, people all over the world need to make use of information in different languages in their specialty.

REFERENCES:

1. Silvia Arano. Thesauruses and ontologies Citación recomendada: Silvia Arano. Thesauruses and ontologies [en linea]. "Hipertext.net", num. 3, 2005. <http://www.hipertext.net> [Consulted: 12 feb. 2007].
2. WU Dan1, WANG Hui-lin2Role of Ontology in Information Retrieval Jun. 2006 Journal of Electronic Science and Technology of China Vol.4 No.2
3. Boris Lauser From thesauri to Ontologies: A short case study in the food safety area in how ontologies are more powerful than thesauri From thesauri to RDFS to OWL http://www.fao.org/agrovoc.

ОНТОЛОГИЧЕСКОЕ МОДЕЛИРОВАНИЕ ИМЕНИ ПРИЛАГАТЕЛЬНОГО КАЗАХСКОГО ЯЗЫКА

А. А. Шарипбай, Г. К. Елибаева, А. С. Муканова, Л. Жеткенбай,
Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан,
sharalt@mail.ru, gaziza_y@mail.ru, Asel_ms@bk.ru, jetlen_7@mail.ru

Настоящая работа посвящена онтологическому моделированию морфологических правил имени прилагательного казахского языка. Онтологические модели позволяют сравнить сходства и различия тюркских языков. Результаты работы будут применяться для разработки многоязычного тезауруса и портала для тюркских языков, содержащих учебные и научные материалы, программные приложения семантического поиска и извлечения знаний, а также других сервисов.

Ключевые слова: *тезаурус, обработка естественного языка, онтологическое моделирование, морфологические правила, имя прилагательное, казахский язык.*

ONTOLOGICAL MODELING OF THE ADJECTIVE OF KAZAKH LANGUAGE

A. Sharipbay, G. Yelibayeva, A. Mukanova, L. Zhetkenbay,
L. N. Gumilyov Eurasian National University, Astana, Kazakhstan,
sharalt@mail.ru, gaziza_y@mail.ru, Asel_ms@bk.ru, jetlen_7@mail.ru

The present work is devoted to the ontological modeling of morphological rules of the adjective of Kazakh language. Ontological models allow comparing the similarities and differences of Turkic languages. The results can be used to develop a multilingual thesaurus and a portal for the Turkic languages, containing training and scientific materials, software applications of semantic search and extraction of knowledge and other services.

Key words: *thesaurus, natural language processing, ontological modeling, morphological rules, adjective, Kazakh language.*

1. Введение

Онтологическая модель морфологических признаков имени прилагательного казахского языка была создана в соответствии с целью проекта, которая является разработкой единого многоязычного электронного тезауруса тюркских (казахского, татарского, киргизского, узбекского и турецкого) языков для многоязычного поиска и извлечения знаний.

Прикладная онтология «имени прилагательного» казахского языка основана на принципах общей онтологии и построена с помощью программного обеспечения Protégé, которое позволяет описывать не только понятия, но и конкретные объекты, а также имеет богатый набор операторов — например, пересечение, объединение и отрицание. Protégé основан на логической модели, которая позволяет создавать определения, соответствующие неформальному описанию. Логическая модель позволяет использовать рассуждения, которые могут проверить все ли утверждения и определения в онтологии взаимно согласуются и могут также выяснить, какие концепции соответствуют заданным определениям [1].

2. Онтологическая модель имени прилагательного казахского языка

Онтологическая модель прилагательного казахского языка состоит из отдельных индивидов, свойств и классов. Индивиды, представляют собой конкретные объекты интересующей предметной области, это основные, нижне-уровневые компоненты онтологии [2].

Рисунок 1 показывает представление индивидов в степенных сравнениях (Degrees of Comparison) имени прилагательного казахского языка, мы представляем отдельных индивидов как ромбики в диаграммах.

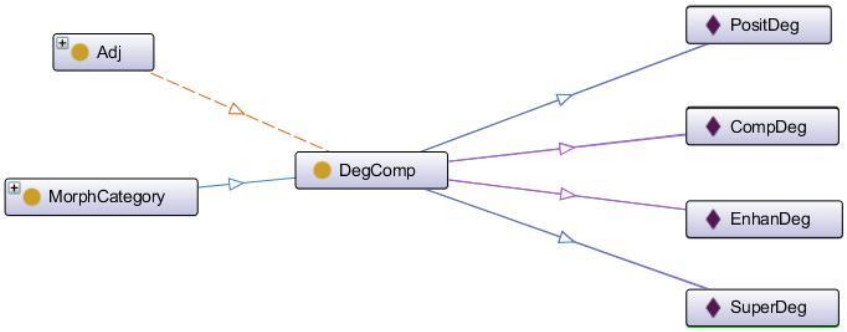


Рисунок 1. Изображение индивидов степенных сравнений имени прилагательного казахского

Свойства – это бинарные отношения на индивидах. Другими словами, свойства соединяют двух индивидов. Например, в нашем примере свойство hasSemanticType (имеет семантическое значение) связывает «имя прилагательного» с индивидами «качественные прилагательные» или «относительные прилагательные». Рисунок 2 показывает представление некоторых свойств, соединяющих некоторых индивидов.

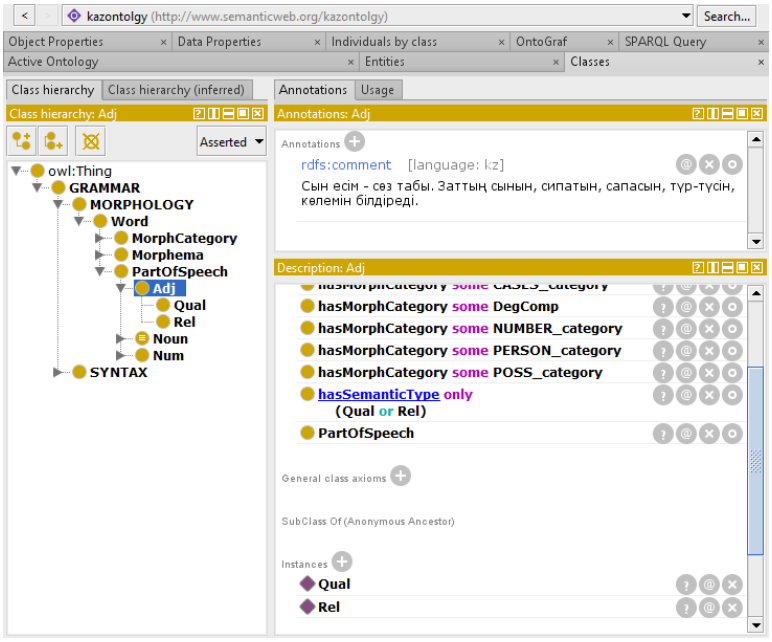


Рисунок 2. Изображение свойств имени прилагательного казахского языка

Классы интерпретируются как множества, элементами которых являются индивиды. Они описываются, используя формальные (математические) конструкции, которые декларируют требования для членства в классе. Классы могут быть организованы в иерархию отношений вида «подкласс-суперкласс», которая также известна как таксономия. Подклассы специализируют (т.е. являются подмножествами) своего суперкласса. Например, рассмотрим классы «Грамматика» и «Морфология». «Морфология» определяется как раздел Грамматики, и поэтому можеть быть подклассом Грамматики (таким образом, Грамматика – суперкласс класса «Морфология»). Это означает, что все «морфологические признаки» – эта грамматика, все члены класса «Морфология» – члены класса «Грамматика». Все спроектированные классы и подклассы прилагательного казахского языка отображены в окне Class hierarchy (Рисунок 3).

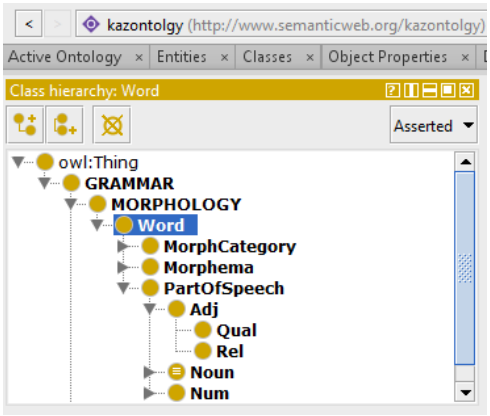


Рисунок 3. Классы и подклассы имени прилагательного казахского языка

После создания классов прописывались в них поля – свойства. К примеру, у класса «CompDeg» (Сравнительная степень) будет свойство «hasSuffixes» (имеет суффикс), которое будет содержать суффиксы сравнительного характера -рақ, -рек, -ырақ, -ірек,-лау, -леу, -дау, -деу, -тау, -теу, -қыл, -ғыл, -қылт, -ғылт, -тым, -шыл, -ші̇л, -қай -аң в котором изменяются смысловые значения прилагательных (Рисунок 4).

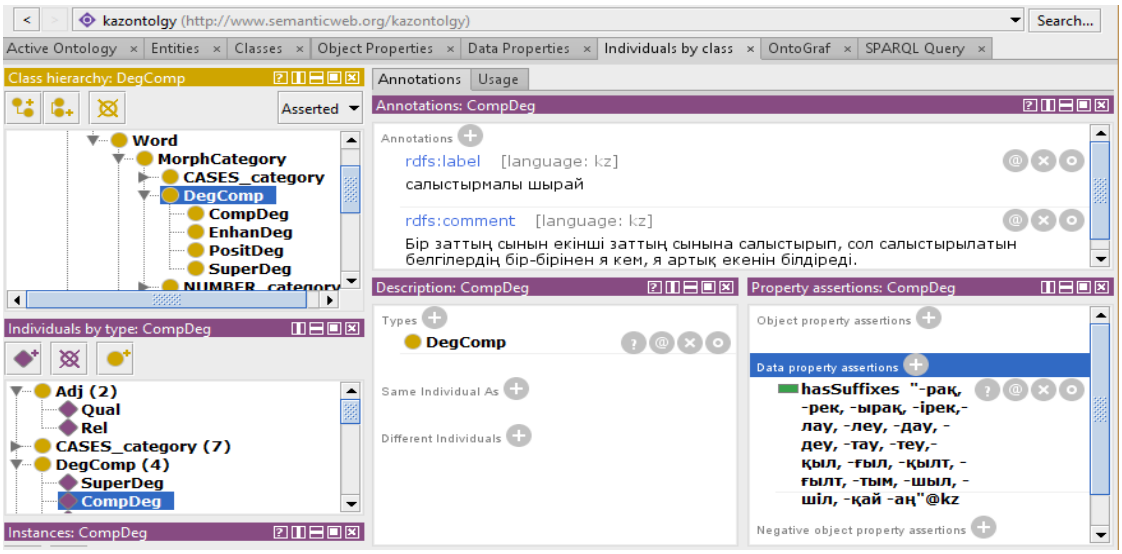


Рисунок 4. Пример свойств суффиксов имени прилагательного казахского языка

После обработки всех классов, свойств и индивидов была построена онтологическая модель имени прилагательного казахского языка (Рисунок 5).

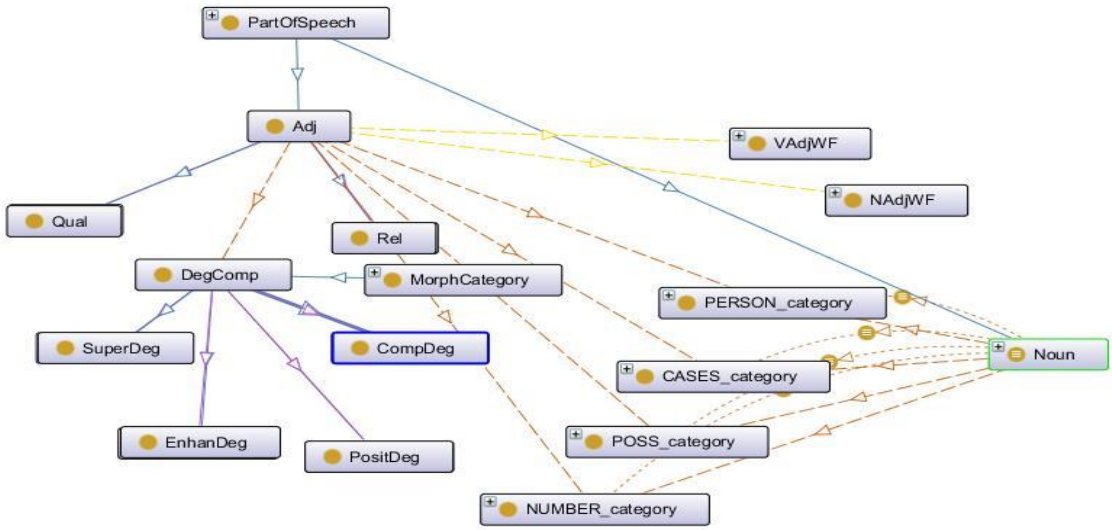


Рисунок 5. Онтологическая модель имени прилагательного казахского языка

Таким образом, онтологическая модель включает в себя все компоненты и совокупности связанные с морфологическими признаками имени прилагательного казахского языка [3-5].

3. Заключение

В результате работы была построена онтологическая модель имени прилагательного казахского языка, которая поможет пользователю получить обширную информацию о морфологических признаках имени прилагательного казахского языка, а также имеет большое влияние для разработки тезауруса и программы машинного перевода тюркских языков.

Статья подготовлена в рамках проекта: ИРН «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» по договору №132 от «12 » марта 2018 г.

ЛИТЕРАТУРА:

1. https://protege.stanford.edu/
2. Gruber, T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing // International Journal Human-Computer Studies. – 1995, vol. 43, pp.907-928
3. Ысқақов А. Қазіргі қазақ тілі (2 басылымы). – Алматы: Ана тілі, 1991. – 384 б.
4. Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис. – Астана, 2002. – 784 б.
5. Қазақ тілі (Қысқаша грамматикалық анықтағыш). – Алматы: Мемлекеттік тілді дамыту институты, 2010. – 92 бет.

СРАВНЕНИЕ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ СУЩЕСТВИТЕЛЬНЫХ КАЗАХСКОГО И КЫРГЫЗСКОГО ЯЗЫКОВ

А. А. Шарипбай[1], Б. Ж. Ергеш[1], Г. К. Елибаева[1], Л. Жеткенбай[1],
[1]Евразийский национальный университет имени Л. Н. Гумилева,
Астана, Казахстан, sharalt@mail.ru, b.yergesh@gmail.com,
gaziza_y@mail.ru, jetlen_7@mail.ru

Н. Исраилова[2], П. Бакасова[2], [2]Кыргызский государственный технический
университет им. И. Раззакова, Бишкек, Кыргызстан,
inela.kstu@gmail.com, bakasovap@mail.ru

Онтология является одним из способов представления знания, которая используется в системах обработки естественного языка (NLP). В этой статье приводится сравнение онтологических моделей морфологии, на примере существительного казахского и кыргызского языков. Для построения онтологии был использован метаязык, построенный в рамках проекта. Результат этих работ могут быть использованы в приложениях NLP.

Ключевые слова: *казахский язык, кыргызский язык, метаязык, онтология, существительное, тюркские языки.*

ONTOLOGICAL MODELS MATCHING OF NOUNS OF KAZAKH AND KYRGYZ LANGUAGES

A. Sharipbay[1], B. Yergesh[1], G. Yelibayeva[1], L. Zhetkenbay[1]
[1]L. N. Gumilyov Eurasian National University, Astana, Kazakhstan,
sharalt@mail.ru, b.yergesh@gmail.com, gaziza_y@mail.ru, jetlen_7@mail.ru

N. Israilova[2], P. Bakasova[2], [2]Kyrgyz State Technical University after I. Razzakov
Bishkek, Kyrgyzstan, inela.kstu@gmail.com, bakasovap@mail.ru

Ontology is one way of representing knowledge that is used in natural language processing systems (NLP). This article compares the ontological models of morphology, on the example of the noun Kazakh and Kyrgyz languages. We used a metalanguage for constructing ontologies, it was developed as part of the project. The result of these works can be used in NLP applications.

Key words: *Kazakh language, Kyrgyz language, metalanguage, ontology, noun, Turkic languages*

1. Введение

Статья подготовлена в рамках проекта: ИРН «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» по договору №132 от «12 » марта 2018 г.

Известно, что онтологии используются как источники данных для многих приложений обработки естественных языков, они позволяют более эффективно обрабатывать сложную и разнообразную информацию. Сегодня онтология широко применяется в системах обработки естественного языка, как способ представления знаний. Этот способ представления знаний позволяет приложениям распознавать семантические отличия, которые не известны компьютеру.

Одно из самых известных определений онтологии дал Том Грубер «Онтология – это спецификация концептуализации» [1]. В общих чертах под онтологией понимается система понятий некоторой предметной области, которая представляется как набор сущностей, соединенных различными отношениями. Онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную предметную область [2]. С помощью онтологии можно обрабатывать знания различных предметных областей[3-5].

Создание онтологических моделей даст толчок для разработок систем извлечения знаний, систем информационного поиска, машинного перевода и других приложений тюркских языков. Онтологические модели приводятся в соответствие к друг другу посредством унифицированного метаязыка, которая разрабатывается в рамках проекта.

С 2014 года на семинарах Uniturk обсуждаются проблемы разработки унифицированной морфологической разметки текстов на тюркских языках для использования в корпусах и других системах автоматической обработки текста.

В данной работе в качестве примера показываются только онтологические модели морфологических правил и таблицы сравнений понятий существительных в казахском и кыргызском языках. Онтологические модели морфологии казахского и кыргызского языков строятся в среде Protege, это свободно распространяемый открытый редактор онтологий и фреймворк для создания интеллектуальных систем. [6].

2. Онтологическое моделирование существительных в казахском и кыргызском языках

На рисунке 1 показана часть онтологической модели морфологических правил образования существительных казахского языка, а на рисунке 2 – часть онтологической модели морфологических правил образования существительных кыргызского языка.

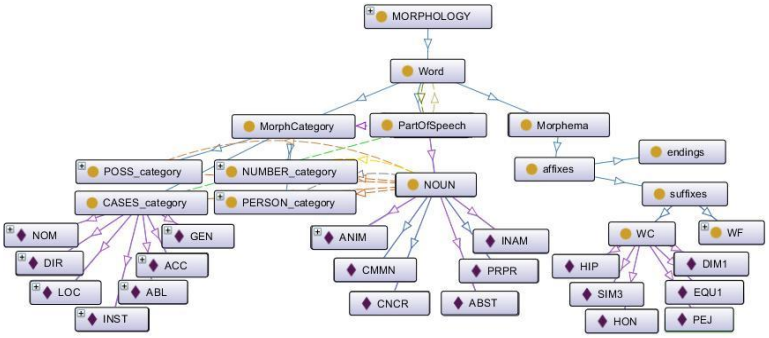


Рисунок 1 – Часть онтологии морфологических правил образования существительных казахского языка

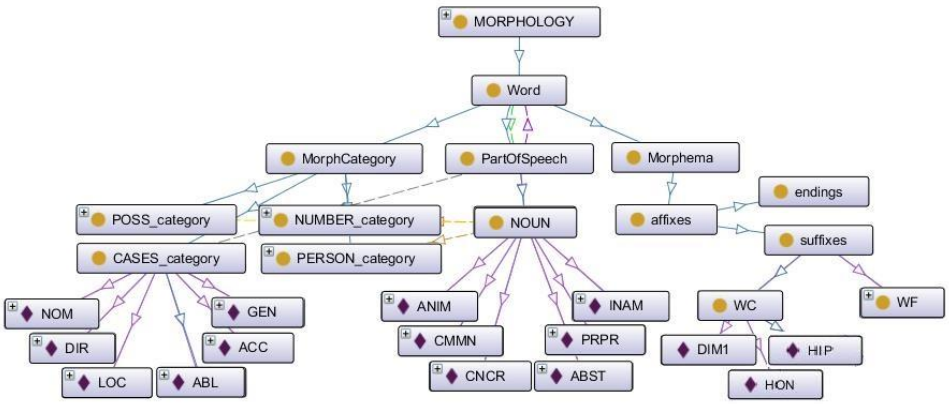


Рисунок 2 – Часть онтологии морфологических правил образования существительных кыргызского языка

Построенные таким образом сравнительные онтологии правил образования существительных охватывают все компоненты и совокупности, которые касаются морфологии.

Наименования и морфологические признаки, примененные в построении онтологий морфологических правил образования существительных казахского (KZ) и кыргызского (KG) языков приведены в таблице 1.

Таблица 1

СОПОСТАВЛЕНИЕ МОРФОЛОГИЧЕСКИХ ПРИЗНАКОВ, НА ПРИМЕРЕ СУЩЕСТВИТЕЛЬНЫХ КАЗАХСКОГО И КЫРГЫЗСКОГО ЯЗЫКОВ

№	Tag	Name_English	KZ	KG
1	N	Noun	1	1
2	SIMP	Simple Noun	1	1
3	CMPL	complex Noun	1	1
4	FUSW	Fused words	1	1
5	PAIR	Pair Noun	1	1
6	CMPN	compound Noun	1	1
7	ABBR	Abbreviations	1	1
8	UNDR	underivatives Noun	1	1
9	DRVT	derivatives Noun	1	1
10	COMP	compound	0	0
11	ANIM	animate Noun	1	1
12	INAM	inanimate Noun	1	1
13	CMMN	common Noun	1	1
14	PRPR	proper Noun	1	1
15	CNCR	Concrete nouns	1	1
16	ABST	Abstract	1	1
17	Number	Number category	1	1
18	SG	Singular	1	1
19	PL	Plural	1	1
20	COL	Collective	1	1
21	ENDS	Endings	1	1
22	Cases	Case endings	1	1
23	NOM	Nominative case	1	1
24	GEN	Genitive case	1	1

25	DIR	Direction- dative case (directive)	1	1
26	DIR_LIM	Limited directive	0	0
27	ACC	Accusative (initial) case	1	1
28	LOC	Locative case	1	1
29	ABL	Ablative case	1	1
30	INST	Instrumental case	1	0
31	Plural	Plural endings	1	1
32	SG	Singular	1	1
33	PL	Plural	1	1
34	Person	Personal endings	1	1
35	P1SG1	1 personal singular	1	1
36	P1SG2	1st person singular	1	1
37	P2SG1	2 personal singular	1	1
38	P2SG.P1	2 personal singular formal	1	1
39	P2SG2	2st person singular	0	0
40	P2SG.P2	2 personal singular formal	0	0
41	P3SG	3 personal singular	1	1
42	P1PL1	1 personal plural	1	1
43	P1PL2	1st person plural	0	0
44	P2PL1	2 personal plural	1	1
45	P2PL.P1	2 personal plural formal	1	1
46	P2PL2	2st person plural	0	0
47	P2PL.P2	2 personal plural formal	0	0
48	P3PL	3 personal plural	1	1
49	POSS	Possesive endings	1	1
50	POSS.1SG	1st person singular possessive	1	1
51	POSS.2SG	2st person singular possessive	1	1
52	POSS.2SG.P	2 Possesive singular formal	1	1

53	POSS.3SG	3 Possesive singular	1	1
54	POSS.1PL	1 Possesive plural	1	1
55	POSS.2PL	2 Possesive plural	1	1
56	POSS.2PL.P	2 Possesive plural formal	1	1
57	POSS.3PL	3 Possesive plural	1	1
58	SUF	Suffixes	1	1
59	WF	word-formative	1	1
60	NWF	Suffixes that form the noun from noun, adjective, numeral	1	1
61	VWF	Suffixes that form the noun from verbs	1	1
62	AWF	Suffixes that form the noun from adjectives	0	1
63	NumWF	Suffixes that form the noun from numerals	1	1
64	WC	word inflection	1	0
65	NWC	Suffixes that change the meaning of nouns	1	0
66	HIP	Hipocoristic	1	1
67	DIM[1]	Diminutive	1	1
68	DIM[2]	Diminutive	0	0
69	SIM[1]	Similative 1	0	0
70	SIM[2]	Similative 2	0	0
71	SIM[3]	Similative 3	1	0
72	EQU[1]	Equative	1	0
73	EQU[2]	Equative	0	0
74	HON	Honorific	1	1
75	PEJ	Pejorative	1	0

Из этой таблицы видно, что имеются некоторые отличия между морфологическими признаками казахского и кыргызского языков.

Например, в строках 62 и 69 в колонке казахского языка признаки отсутствуют, а в колонке кыргызского языка они присутствуют, и наоборот, в строках 30, 64, 65,71, 72,75 в колонке кыргызского языка признаки отсутствуют, а в колонке казахского языка они присутствуют. Это означает,

что в первом случае признаки в строках 62 и 69 свойственны только для кыргызского языка, а во втором случае признаки в строках 30, 64, 65,71, 72,75 свойственны только для казахского языка. Из 75 признаков 8 признаков различны, из этого можно сделать вывод, что морфологические признаки существительных казахского и кыргызского языков аналогичны на 89%. Возможно, что в данном сопоставлении не учтены все признаки, присущие существительным исследуемых языков. Но в данной статье была попытка полного охвата согласно грамматике языков.

Примеры образования множественного числа существительного «ребёнок», «дом», «книга» и «очередь» в казахском и кыргызском языках — «бала — балдар», «үй – үйлөр», «кітап — китептер» и «кезек — кезектер» показаны в таблице 2.

Таблица 2

ОБРАЗОВАНИЯ СУЩЕСТВИТЕЛЬНЫХ МНОЖЕСТВЕННОГО ЧИСЛА

На казахском языке	На кыргызском языке
бала + лар = балалар	бала+дар = балдар
үй + лер = үйлер	үй+лөр = үйлөр
кітап + тар = кітаптар	китеп +тер = китептер
кезек + тер = кезектер	кезек +тер = кезектер

Примеры склонения существительного «школа» в казахском и кыргызском языках «мектеп — мектеп» показаны в таблице 3.

Таблица 3

СКЛОНЕНИЕ СУЩЕСТВИТЕЛЬНОГО «МЕКТЕП — МЕКТЕП — ШКОЛА»

На казахском языке	На кыргызском языке
мектеп: мектеп+N +NOM	мектеп: мектеп + N +NOM
мектептің: мектеп+N +GEN	мектептин: мектеп+ N +GEN
мектепке: мектеп+N+ DIR	мектепке: мектеп+ N +DIR
мектепті: мектеп+N+ ACC	мектепти: мектеп+ N +ACC
мектепте: мектеп+N+ LOC	мектепте: мектеп + N +LOC
мектептен: мектеп+N+ ABL	мектептен: мектеп + N+ABL
мектеппен: мектеп+N+ INST	-
мектептер: мектеп+N+PL+ NOM	мектептер: мектеп + N+ PL+NOM
мектептердің:мектеп+N+ PL+GEN	мектептердин:мектеп+ N+ PL+GEN
мектептерге:мектеп+N+ PL+ DIR	мектептерге:мектеп + N+ PL+DIR
мектептерді:мектеп+N+ PL+ACC	мектептерди:мектеп + N+ PL+ACC
мектептерде:мектеп+N+ PL+LOC	мектептерде:мектеп+ N+ PL+LOC
мектептерден:мектеп+N+ PL+ABL	мектептерден: мектеп+ N+ PL+ABL
мектептермен:мектеп+N+ PL+ INST	-

3. Заключение

В данной работе проведен сравнительный анализ существительных казахского и кыргызского языков, выполненный на основе разработанных в рамках проекта онтологических моделей существительных обоих языков с применением единого метаязыка. Созданные онтологические модели можно использовать в морфологических анализаторах, системах извлечения знаний, системах информационного поиска, машинного перевода и других приложений обработки казахского и кыргызского языков.

ЛИТЕРАТУРА:

1. Th.Gruber. What is an Ontology// URL: http://www-ksl.stanford.edu/kst/what- is-an-ontology.html.
2. Митрофанова О.А. Онтологии как системы хранения знаний / О.А. Митрофанова, Н.С. Константинова // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. – 54 с.
3. B. Yergesh, G. Bekmanova, A. Sharipbay, M. Yergesh. Ontology-Based Sentiment Analysis of Kazakh Sentences. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10406 LNCS, pp. 669-677.
4. L.Zhetkenbay, A.Sharipbay, G.Bekmanova, U.Kamanur. Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages. 2016. Journal of Theoretical and Applied Information Technology. Vol.91.N.2.2016.P.257-263.
5. П.Бакасова, Н.Исраилова. Онтологическая модель морфологических правил кыргызского языка. Материалы № 60 Международной научно-технической конференции молодых ученых. 2018.С.25-29
6. http://protege.stanford.edu/ (accessed 10.08.2018).

ONTOLOGICAL MODEL OF THE EDUCATIONAL PROGRAM COMPUTATIONAL LINGUISTICS

A.Sharipbay, R.Niyazova, S.Kudubayeva, R.Turebayeva, A.Aktayeva, L.Davletkireeva,
L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
sharalt@mail.ru, rozamgul@list.ru, saule.kudubayeva@gmail.com, 58stud@mail.ru,
Sh. Ualikhanov Kokshetau State University, Kokshetau, Kazakhstan,
aaktaewa@list.ru, Magnitogorsk State Techninal University, Magnitogorsk, Russia

The ontological model of the master's educational program "Computer linguistics" implemented in the framework of the Erasmus+ project is proposed. Since this specialty can be entered by both bachelors-philologists and bachelors of IT specialties, it is proposed to train masters in three trajectories: Linguistics, Speech technology, Text technology. Ontological model through logical connections allows to reveal theAly structure and content of the educational program on the proposed learning paths.

Key words: *educational programs, computational linguistics, Mathematical Foundations of Computer Linguistics, Semantic analysis of texts, Methods of speech synthesis.*

ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

А. Шарипбай, Р. Ниязова, С. Кудубаева, Р. Туребаева, А. Актаева, Л. Даулеткиреева,
Евразийский национальный университет им. Л. Н. Гумилева, Астана, Казахстан,
sharalt@mail.ru, rozamgul@list.ru, saule.kudubayeva@gmail.com, 58stud@mail.ru,
Кокшетауский государственный университет им. Ш. Валиханова, Кокшетау, Казахстан, aaktaewa@list.ru, Магнитогорский государственный технический университет, Магнитогорск, Россия

Предложена онтологическая модель магистерской образовательной программы «Компьютерная лингвистика», реализуемой в рамках проекта Erasmus+. Так как на данную специальность могут поступать как бакалавры-филологи, так и бакалавры IT специальностей, то предлагается

обучать магистров по трем траекториям: Лингвистика, Речевые технологии, Текстовые технологии. Онтологическая модель через логические связи позволяет раскрыть структуру и содержание образовательной программы по предлагаемым траекториям обучения.

Ключевые слова: *образовательные программы, вычислительная лингвистика, математические основы компьютерной лингвистики, семантический анализ текстов, методы синтеза речи.*

1. Introduction

Modern computational linguistics aims at solving problems and tasks, involving formalized processing and analysis of natural languages. This is, first, well-known machine translation, information retrieval, speech recognition and synthesis, emotional analysis of language data, etc. Second, building language resources for educational and research purposes, developing language theory. Third, the development of various functional applications that utilize language data.

Computational linguistics is necessary for startups, developing new linguistic technologies — such as developing machine-human natural language interface, to automatically recognize the emotions of user texts in social networks; in companies, processing large amounts of unstructured text data.

Today, experts in computational linguistics are very much in demand in the largest companies, working in the field of linguistics. In Central Asia, the task of training such staff is new, it became clear only recently, due to the rapid development of artificial intelligence. At the same time, there are no educational programs for targeted mastering in the field of computational linguistics, which proves the urgent relevance of the master's educational program on computational linguistics, being implemented within the Erasmus+ project. The list participating universities are as following:

1. University of Santiago de Compostela, USC
2. University of a Coruña, UDC
3. Technological Educational Institution of Athens, TEIATH
4. University of Porto, U.PORTO
5. Adam Mickiewicz University in Poznań, AMU
6. Urgench State University, UrSU
7. Samarkand State Institute of Foreign Languages, SamSIFL
8. Tashkent State University of the Uzbek language and literature, TSUUL
9. National University of Uzbekistan, NUUz
10. Republican State Enterprise operating under the right of economic management A.Baitursynov Kostanay State University of Ministry of Education and Science of the Republic of Kazakhstan, KSU
11. L.N. Gumilyov Eurasian National University, ENU
12. Al-Farabi Kazakh National University, KazNU[1]

In Kazakhstan, The classifier of specialties for higher education doesn't include the specialty "Computational Linguistics". But at the L.N. Gumilyov ENU, the trajectory of training "Computational Linguistics" in the framework of the educational programs of specialties "5B060200-Informatics" for undergraduate and "6M060200-Informatics" for magistracy has been developed and implemented since 2013. For this purpose, we studied the educational programs on the computational linguistics of foreign universities.

2. Questioning of employers of specialists in the field of computer linguistics

In order to find out the employers' opinion about the prospective competencies required for specialists in the field of computational linguistics, we conducted a questionnaire using the following application:

https://docs.google.com/forms/d/e/1FAIpQLScT6iJA5e-iYKIC4mAIzcF_sMch 77CRylhGUwmi_FQ1afiRWA/viewform?c=0&w=1&usp=mail_form_link

The purpose of the questionnaire is to decide the competencies and needs of specialists in the field of computer linguistics based on interviewing potential employers.

The result of the questionnaire: the definition of the list of competencies and the needs of specialists in the field of computational linguistics.

3. Analysis of educational programs in computational linguistics of foreign universities

The purpose of the analysis of educational programs is to decide the level of education and the list of disciplines that offer these levels.

The result of the analysis of educational programs is the definition of a list of disciplines because of an analysis of the content of international educational master's programs in the field of computational linguistics.

Training of national cadres in the field of computational linguistics should be carried out through the development and implementation of quality educational programs based on the study of international experience in the development and implementation of educational programs in computational linguistics, tuning methodology tools, as well as the questioning, diagnosis and classification of training needs.

To develop the master's educational program "Computational Linguistics", implemented by the Erasmus+ project, we analyzed the educational programs of the following universities and institutes:

Table 1.

ANALYSIS OF EDUCATIONAL PROGRAMS

№	COUNTRY, UNIVERSITY, WEBSITE	EDUCATIONAL PROGRAM
1	RGGU, Russian State University for the	Introduction to fundamental linguistics Typology, comparativistics, areal linguistics

	Humanities, www.rggu.ru	Modern syntactic theories Case and experimental methods in semantics Introduction to Computational Linguistics Computer Sociolinguistics Mathematical Foundations of Linguistics Statistical models in linguistics Methods of artificial intelligence in Computational Linguistics Programming of linguistic tasks Linguistic annotation / markup Specialized linguistic databases Methods for evaluating AOT systems Models and methods of Computational Linguistics Classification methods and machine learning Linguistic basis of machine translation Computer Parsing Analysis of oral speech
2	SPbSU, St. Petersburg University, spbu.ru	Methods and models of ontological engineering Methods of knowledge engineering in humanitarian research Text understanding systems Text analysis models and their software implementation Statistical methods in language engineering Hull methods in language engineering Linguistics of the text and theory of verbal communication Languages and standards for describing information resources Expert systems and methods of inductive generalizations Methods of decision support Methods of software implementation of intelligent information technologies Mathematical modeling in data processing technologies Methodology and technology of designing information systems Information Society and Problems of Applied Informatics Business English Philosophical problems of science and technology
3	HSE, High School of Economics Computational Linguistics	Linguistic data: quantitative analysis and visualization Introduction to Linguistics Mathematics Formal models in linguistics Functional and cognitive models in linguistics

		Computational Linguistics Programming (Python) Analysis of linguistic data: quantitative methods and visualization (taught in English) Mathematical foundations of Computational Linguistics Machine learning Experimental Linguistics Database Ontologies and semantic technologies Digital Humanitarian Technologies: Resources, Tools, Case Studies Designing of linguistic resources and systems
4	MPhTI, Moscow University of Physics and Technology, mipt.ru	Mathematical Foundations of Linguistics Statistical models in linguistics Introduction to fundamental linguistics Typology, comparativistics, areal linguistics Russian corpus grammar Introduction to Computational Linguistics Computer Sociolinguistics Modern syntactic theories Typology of grammatical categories English for professional communication Models and methods of Computational Linguistics Data structures and basic algorithms The main algorithms of linguistic analysis Analysis of oral speech. Corpus linguistics: building and using enclosures Classification methods and machine learning Computer models of discourse Linguistic basis of machine translation Formal models and resources of world languages Linguistic annotation / markup Methods for evaluating AOT systems Computer parsing. Methods of artificial intelligence in Computational Linguistics Application Packages for Linguistic Studies Specialized linguistic databases Linguistic support of the tasks of document analysis Automatic assessment of the complexity of texts
5	ITMO(St. Petersburg) St. Petersburg National	**Information Technology:** System analysis and modeling of information processes

	Research University of Information Technologies, Mechanics and Optics, ifmo.ru	and systems; Designing information systems; Organization of design and development of distributed systems software; Organization of software design and development for embedded systems; Software testing; Quality management software development. **Speech technologies:** Digital signal processing; Digital processing of speech signals; Mathematical modeling and decision theory; Pattern recognition; Recognition and synthesis of speech; Recognition of the speaker (speaking by voice); Multimodal Biometric Systems
6	University of Oxford http://www. ox.ac.uk	Analysis of functional and structural data images of the brain. Physiological neuroimaging. Brain disorders. Diffusion of the image. Speech and the brain. Visualization. Neurodegeneration. Cognition. Psychiatry
7	University of California, Los Angeles (UCLA) http://www. ucla.edu	Phonetics. Phonology. Syntax. Semantics. Psycholinguistics. Math. Linguistics. Historical Linguistics. African, Indian languages
8	Harvard University https://www. harvard.edu	Fundamental studies of the speech apparatus and speech functions. Clinical studies of human voice and speech abnormalities. Mechanics, biophysics, physiology and / or molecular biology of the middle and inner ear. Acquired or congenital abnormalities of the mechanisms of hearing. Neurophysiological or modeling approaches in the study

		of nerve cells and circuits underlying auditory processing. Neurovisual studies of the mechanisms of tinnitus. Cognitive neurobiology of language signal processing. Design, development and improvement of the hardware and software system for hearing aids, ear implants, vestibular prostheses or algorithms for automatic speech recognition.
9	Cambridge University https://www. cam.ac.uk	Acoustic modeling (statistical models). Fundamental research in machine learning. Optimize dialogue using reinforcement learning. Recognition on large dictionaries. Pattern recognition. Speech recognition on mobile devices. Dictator independence and noise cancellation. Dialog systems and VoiceXML. Statistical language modeling. Statistical machine translation. Processing and transcription of recognized speech
10	Carnegie Mellon University https://www. cmu.edu	User Interface Software Cognitive models. Speech recognition. Understanding of natural language. Computer graphics. Handwriting recognition. Visualization of data, visual design, multimedia. Computer support for teamwork. Computer music and theatrical skill. Social technologies
11	Johns Hopkins University https://www. jhu.edu	Language modeling. Natural language processing. Neural treatment. Acoustic processing. The theory of optimization. Language Entry

Hence, although the educational programs of some universities are called differently, in terms of their content, they refer to computational linguistics.

4. Ontological model of the master's educational program "Computational Linguistics", implemented under the project Erasmus+

1. Purpose and description of the master's degree program. The master's educational program "Computer linguistics" has three directions: Linguistics, Speech technologies, Text Technologies (Text processing).

In the direction of "Linguistics", there are bachelors-philologists and bachelors-psychologists.

Graduate students with a philological education are necessary for the training of specialists in the engineering of knowledge in the field of linguistics, without the need to study the mathematical foundations of computer science and programming. They will study the methods of composing the thesauri for certain subject areas, markup languages for audio and text records for creating text and audio corpora, language models for speech technologies.

In the direction of "Speech Technology" and "Text Technology". Graduate students who graduated from the bachelor's degree in computer science will be trained in two parallel trajectories: Speech technologies and word processing technologies, which will provide mathematical basis and methods for creating these technologies.

Preparation of masters in the direction of "Linguistics" requires: a deep study of the fundamental foundations of linguistics with an emphasis on the methods of creating operational formal models of the language system, the adequate complexity of such tasks of natural language processing, as recognition and synthesis of speech and text, semantic analysis and understanding of text and speech.

Preparation of masters in the direction of Speech Technology, requires in-depth study Methods of transcription of sounds and speech, Methods of Speech Recognition, Methods of speech synthesis, Methods of transcription of sounds and speech, Methods of Speech Recognition, Methods of speech synthesis

Preparation of masters in the direction of "Text Technology" requires Methods and tools for creating text corpora

Methods and tools for creating audio corpora, Syntax analysis of texts, Morphological analysis of texts, Semantic analysis of texts.

It should be noted that any educational program consists of mandatory disciplines (**Mandatory part**) and elective disciplines (**Variational part**).

1. **Mandatory part**: History and philosophy of science, Foreign language (professional), Pedagogics, Psychology, Software development technology

2. **Variational part**:

Linguistics: Ontology design tools, Tools for processing visual data, Tools for processing audio data, Tools creating thesauri, Semantic Search Tools.

Speech Technologies: Methods of transcription of sounds and speech, Methods of Speech Recognition, Methods of speech synthesis, Methods of transcription of sounds and speech, Methods of Speech Recognition, Methods of speech synthesis.

Text technology: Methods and tools for creating text corpora

Methods and tools for creating audio corpora, Syntax analysis of texts, Morphological analysis of texts, Semantic analysis of texts.

2. Ontological model of master's degree program

One need to note, that ontology is a weighted oriented graph: the weights of the vertex are the concepts, and the weights of the edges are the relation. In this case, to shortcut the concept, they are coded by a single letter with upper and lower indices: the upper index indicates the semester's number and the lower number of the discipline in the semester. Table 2 lists the concepts and their codes.

Table 2.

THE CONCEPTS AND THEIR CODES

CONCEPTS	CODE
Tools for the creation of language corpora	M_1^1
Languages for symbol processing	M_2^1
Digital Signal Processing	M_3^1
Tools for creating text corpora	B_1^1
Tools for creating audio corpora	B_2^1
Programming in Python	B_3^1
Programming in Prolog	B_4^1
Methods of digital processing of speech signals	B_5^1
Software for processing speech signals	B_6^1
Tools for processing natural languages	M_1^2
Text recognition	M_2^2

Speech recognition	M_3^2
Ontology design tools	B_1^2
Tools for processing visual data	B_2^2
Tools for processing audio data	B_3^2
Syntax analysis of texts	B_4^2
Morphological analysis of texts	B_5^2
Semantic analysis of texts	B_6^2
Methods of transcription of sounds and speech	B_7^2
Methods of Speech Recognition	B_8^2
Methods of speech synthesis	B_9^2
SearchTools	M_1^3
Creating of text corpus	M_2^3
Creating of audio corpus	M_3^3
Tools creating thesauri	B_1^3
Semantic Search Tools	B_2^3
Sentiment analysis of natural language texts	B_3^3
Method for processing of text corpora	B_4^3
Synthesis of speech analysis of natural language	B_5^3
Methods for processing of audio corpora	B_6^3
Writing and defense of Master's degree thesis	G_1^4

Figure 1 shows the core of the ontology of the educational program on computational linguistics, which includes only profile disciplines (there are no general education and practices, there are no final exams and protection).

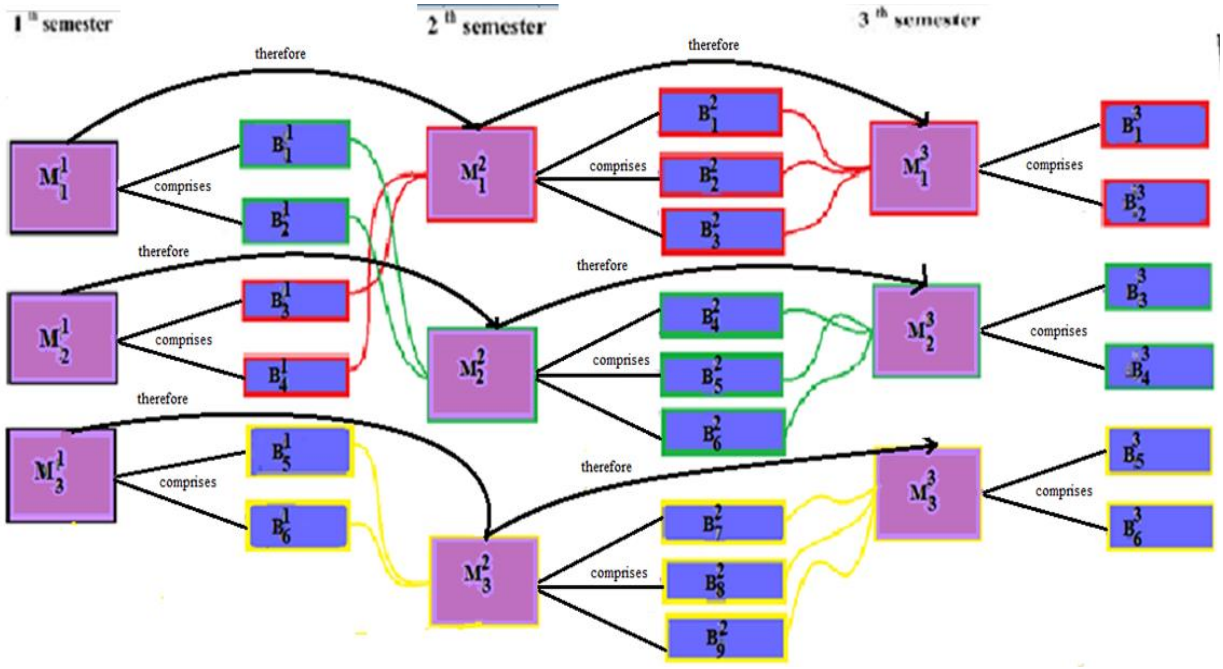


Fig. 1. Ontology of the educational program on computational linguistics

The ontological model constructed in Fig. 1 through logical connections allows us to show the structure and content of the educational program taking into account 3 possible trajectories of training, to meet the goals of the educational program and the results of the training. At the top of the ontology is the time axis (1 semester, 2 semesters, 3 semesters, 4 semesters)

C. Formalization of learning trajectories in ontologies

The ontological model allows for scripting of all possible training trajectories by mean of formal notation.

In ontology, each edge defines the relation «Preceding | It follows» between the vertices representing the elements of the educational program, which are denoted by uppercase Latin letters with upper and lower indices (upper indices — semester numbers, lower indices — numbers of disciplines in the semester):

– Modules,

– Disciplines of the first semester,

Disciplines of the second semester,

,, – Disciplines of the third semester,

In ontology, vertices and edges are colored in different colors: green color — mandatory general education disciplines and relations between them, purple color — mandatory profile modules and relations between them, blue color — elective disciplines and relations between them, beige color — research work, educational practice, research internship, and red color — a comprehensive exam in the specialty, writing and defense of the thesis.

In the ontology of the educational program, the following three areas of the learning path can be distinguished: «**Linguistics**», «**Speech technologies**» and «**Text technology**».

Now, using the ∨- disjunction operation over the vertices and the operation of • -concatenation over edges, one can formalize various learning paths for each direction in the educational program, for example, in terms of modules — a ***modular trajectory*** and in terms of disciplines — a ***disciplinary trajectory***:

Direction «**Linguistics**»:

– Modular trajectory,

∨ – Disciplinary trajectory.

Direction **«Speech technologies»:**

– Modular trajectory,

}– Disciplinary trajectory.

Direction **«Text technology»:**

– Modular trajectory,

∨ ∨ – Disciplinary trajectory.

If in the disciplinary trajectory of the direction "Linguistics" perform all operations of disjunction, then one can get 12 trajectories. Similar trajectories can be obtained for other directions, i.e. in total, 36 trajectories can be obtained.

5. Conclusion

We assume the perspectives of the development of "Computational Linguistics" in the wide use of methods of artificial intelligence and knowledge base that will create:

- Intelligent question-answer systems, which could answer questions about the meaning;

- Intelligent knowledge assessment systems that have the ability and generate questions from their knowledge base, understand and evaluate responses in the range from 0 to 100%;

- Intelligent control systems that could assess the situation or state of an object or management process and make the appropriate decision to manage it;

- An intellectually diagnosed system that can use the training sample and assess the future behavior of the research.

All these systems can be loaded into the memory of robots and create and enhance their intelligence. This means that in the future the development of information and communication technologies will be based on the achievement in the field of computational linguistics.

REFERENCES:

1. Partners [Electronic resource],-2018.-URL:http:/www.comling.enu.kz.

2. Gruber T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing / Gruber T.R. // International Journal Human-Computer Studies. –43, 907-928 (1995)

3. Yergesh B., Mukanova A., Sharipbay A., Bekmanova G., and Razakhova B. Semantic Hyper- graph Based Representation of Nouns in the Kazakh Language. Computación y Sistemas. 18(3), 627–635(2014). DOI: 10.13053/CyS-18-3-2041.
4. Zhetkenbay L., Sharipbay A., Bekmanova G., Khabylashimuly M., Kamanur U.. Thesemantical, ontological models and formalization rules Kazakh compound words. Turklang‘14 II International Conference on Computer processing of Turkic Languages, Istanbul, Turkey, 2014– Istanbul, 2014. – P.107-113.
5. Zhetkenbay L., Sharipbay A.A., Bekmanova G.T., Kamanur U. The ontological model of noun for Kazakh-Turkish machine translation system. Turklang‘15// III International Conference on Computer processing of Turkic Languages, Russia, Kazan, Tatarstan. – Kazan, 2015. –P.15-24.

СЕКЦИЯ 5.
СИСТЕМЫ МОРФОЛОГИЧЕСКОЙ И СИНТАКСИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

О РАЗРАБОТКЕ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА ТАТАРСКОГО ПРЕДЛОЖЕНИЯ

А.Р.Гатиатуллин, Р.Р.Гатауллин, А.Баширов, Академия наук Республики Татарстан, Казанский федеральный университет, Казань, Россия
ayrat.gatiatullin@gmail.com, ramil.gata@gmail.com, a.basheerov@gmail.com

В статье описывается работа по созданию семантико-синтаксического анализатора для анализа татарского предложения. Результатом работы анализатора должны стать модифицированные деревья непосредственных составляющих, дополненные семантической информацией в виде ролевой структуры предложения. Для отработки технологий и создания базы данных с правилами описания семантико-синтаксической структуры татарского предложения используется открытая библиотека NLTK. Для добавления в модифицированные деревья непосредственных составляющих ролевой структуры предложения используются базы данных с ситуационными фреймами. Кроме создания непосредственно самого семантико-синтаксического анализатора татарского предложения, данный проект должен способствовать созданию целого ряда лингвистических баз данных, содержащих информацию о структурных и семантических свойствах текстовых единиц татарского языка.

Ключевые слова*: семантико-синтаксический анализатор, татарский язык, деревья непосредственных составляющих.*

ABOUT THE DEVELOPMENT OF THE SEMANTIC AND SYNTACTIC ANALYZER OF THE TATAR SENTENSE

A.Gatiatullin, R.R.Gataullin, A.Bashirov, [2]Academy of Sciences of Tatarstan, Kazan Federal University, Kazan, Russia
ayrat.gatiatullin@gmail.com, ramil.gata@gmail.com, a.basheerov@gmail.com

The article describes work on creating a semantic-syntactic parser for analyzing the Tatar sentence. As the result, the parser should build modified

immediate constituent trees, supplemented by semantic information in the form of sentence role structure. The open-source NLTK library is used for developing and refining the technology, as well as for creating databases with semantic-syntactic sentence structure description rules for the Tatar sentence. Databases with situation frames are provided for modifying the immediate constituent trees of the sentence role structure. In addition to developing a semantic-syntactic parser for the Tatar sentence, the platform should contribute to the creation of various linguistic databases, containing the data on structural and semantic text unit properties of the Tatar language.

Key word: *semantic-sintaxic parser, Tatar language, trees of direct components.*

В настоящее время существует целый ряд методов синтаксического и семантического анализа предложения с разными моделями представления синтаксической структуры и описания семантики. Это позволяет сказать, что задача автоматического синтаксического анализа для большого количества европейских языков, а также для русского языка уже решена. Для тюркских языков также известен целый ряд разработок. Среди которых можно отметить разработки для турецкого (Eryigit G, 2008) и казахского языков (Бегимтай, 2016; Жуманов, 2012). Однако, работы по усовершенствованию программ синтаксического анализа (увеличению точности и скорости анализа), а также создание анализаторов для новых языков все еще продолжаются. Это подтвердил конкурс программ синтаксического анализа, проведенный в рамках международной конференции по компьютерной лингвистике Dialog‘2012 (http://www.dialog-21.ru/dialogue2012/results/).

Существование синтаксического и семантико-синтаксического анализаторов для предложений татарского языка нам неизвестно, поэтому создание такого анализатора — актуальная задача. Он необходим для целого ряда программных продуктов, разрабатываемых для татарского языка. Это решение задачи семантико-синтаксической разметки татарского электронного корпуса, многоязычная метапоисковая система, а также целый ряд других программ. Технологии и лингвистические ресурсы, полученные при создании этого парсера, могут быть использованы для уменьшения процента многозначности, которая получается на выходе программы морфологического анализа.

Популярные в настоящее время технологии статистического подхода и подхода с использованием нейросетей требуют наличия синтаксически размеченных корпусов большого объема. По причине отсутствия такого синтаксически размеченного электронного корпуса татарского языка нами принято решение о создании программы синтаксического анализа с использованием правил контекстно-свободных грамматик.

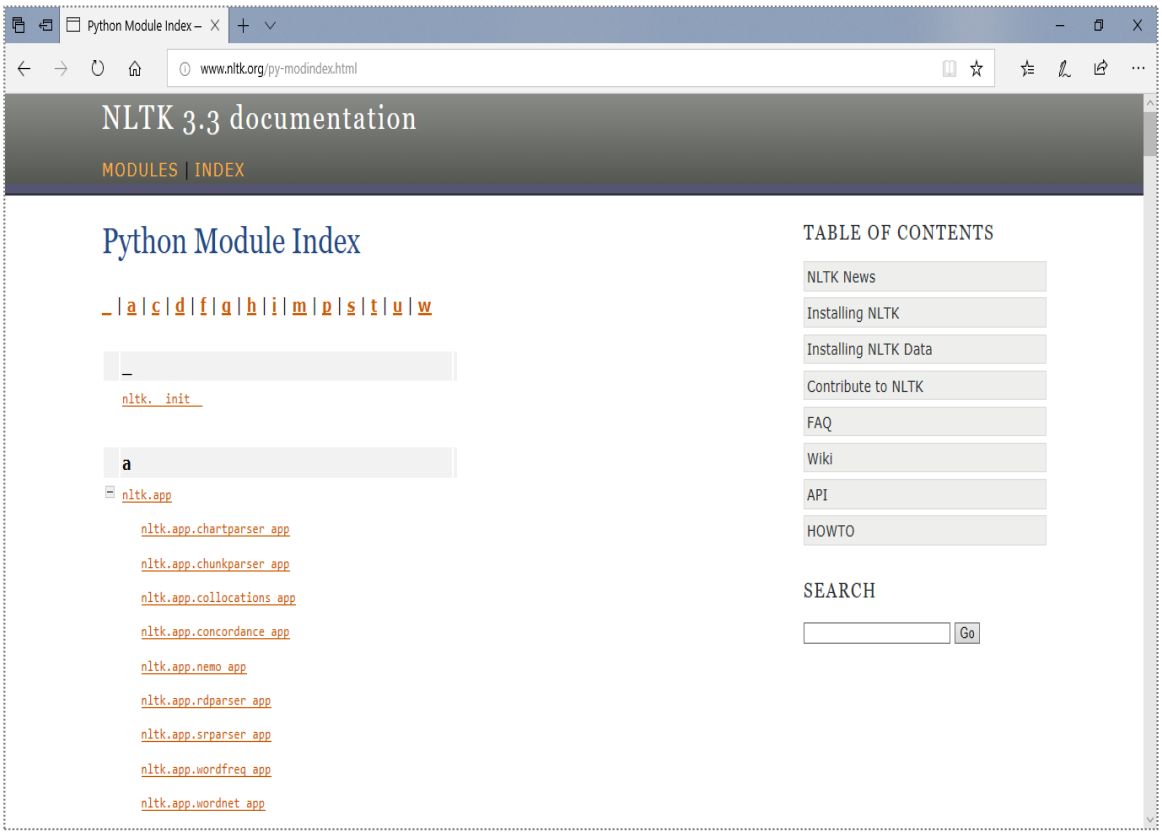
В работах (Eryigit G, 2008; Бегимтай, 2016; Жуманов, 2012) результатом работы синтаксического анализатора являются деревья зависимостей. Учитывая то, что тюркские языки в силу своей структурной особенности являются языками проективного типа, в качестве базовой модели для представления механизма синтаксического анализатора языков проективного типа возможно использовать деревья непосредственных составляющих (НС). Также выбор деревьев непосредственных составляющих связан с тем, что с помощью НС-структур в предложении можно выделить не только отдельные слова, но и некоторые аналитические конструкции, функционирующие как единое целое (например, в русском языке — «будем обязаны», в татарском — «барырга иде»). С помощью НС-структур более естественно описываются конструкции с неподчинительными отношениями, а таких конструкций в языке достаточно много. Применение классических недетерминированных анализаторов приводит к значительному проигрышу по скорости из-за необходимости возвратов (backtracking). Это означает, что нужны дополнительные приемы, которые позволят уменьшить количество альтернатив при синтаксическом анализе. В данном проекте нами выдвинута гипотеза, что уменьшению альтернативных вариантов будет способствовать использование семантических методов, так как они отсеют бесперспективные варианты и определят стратегию поведения парсера.

На следующем этапе нам необходимо было определиться: строить полностью новую программу или использовать уже существующие разработки. Для решения нашей задачи необходим инструментарий, который позволяет строить НС-деревья с использованием морфологической и семантической информации.

Проведя анализ существующих программных продуктов, мы остановили свой выбор на открытой библиотеке NLTK (Natural Language Toolkit). NLTK (www.nltk.org) — это пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python, который содержит графические представления и примеры данных.

Рис. 1.

СТРАНИЦА САЙТА NLTK



Одним из основных плюсов библиотеки NLTK является то, что она представляет из себя продукт с открытым исходным кодом (в отличие от ABBYY Compreno), а также обладает удобным инструментарием для проведения экспериментов.

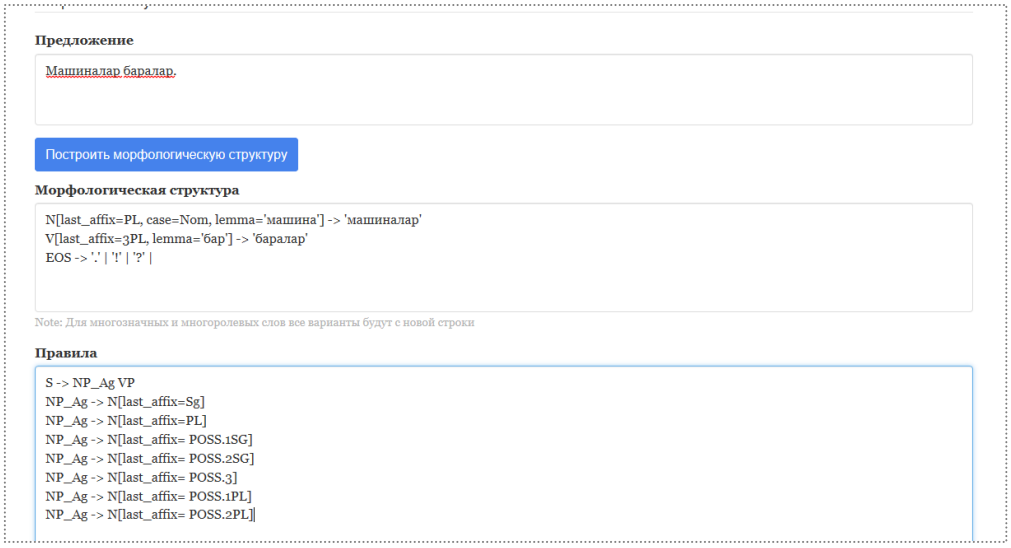
Этапы анализа. Программу семантико-синтаксического анализа мы представили в виде следующих основных этапов:

1. Морфологический анализ.
2. Определение аналитических форм.
3. Определение валентностей глаголов (из словаря валентностей).
4. Построение структуры НС-деревьев.

Для экспериментов с правилами создан сайт семантико-синтаксического анализа, который позволяет в экспериментальном режиме тестировать заполняемые правила КС-грамматики (Рис.2.).

Рис.2.

ФРАГМЕНТ ИНТЕРФЕЙСА ДЛЯ ТЕСТИРОВАНИЯ ПРАВИЛ



Используемый в программе модуль морфологического анализа татарской словоформы производит полный морфологический анализ словоформы, разбивая ее на морфемы. Однако, нами была выдвинута гипотеза о том, что для определения синтаксических связей между словоформами достаточно информации только о последней морфеме в словоформе и поэтому в окне «Морфологическая структура» (Рис.2) представлена только следующая информация: лемма, последняя морфема (last_affix), дополнительная информация. В представленном примере в качестве дополнительной информации указана грамматическая категория Номинатив, которая не определяется в явном виде посредством аффиксальной морфемы. Все эти параметры используются в дальнейшей работе программы.

Формирование аналитических форм представляет собой процедуру, которая собирает аналитические формы в единую структуру. В дереве непосредственных составляющих аналитическая форма будет представлять один цельный элемент. К собираемым аналитическим формам относятся формы, образуемые с помощью послелогов, вспомогательных глаголов и частиц.

Модуль определения валентностей производит поиск валентностей глаголов в базе данных, содержащем словарь валентностей татарских глаголов.

Правила контекстно-свободной грамматики

В окне «Правила» представлен набор правил контекстно-свободных грамматик, которые создаются для формирования деревьев непосредственных составляющих.

Рассмотрим фрагмент базы правил контекстно-свободной грамматики:

S → NP_Ag VP
NP_Ag → N[last_affix=Sg]
NP_Ag → N[last_affix=PL]
NP_Ag → N[last_affix=POSS.1SG]
NP_Ag → N[last_affix=POSS.2SG]
NP_Ag → N[last_affix=POSS.3]
NP_Ag → N[last_affix=POSS.1PL]
NP_Ag → N[last_affix=POSS.2PL]
NP[last_affix=?n] → Adj[last_affix=Adj] N[last_affix=?n]
NP[last_affix=?n] → Adj[last_affix=Adj] N[last_affix=?n]
NP[last_affix=PL] → N[last_affix=PL] CNJ[last_affix=CNJ] N[last_affix=PL]
VP → VP_1V
VP_1V → VP[sem_type=V_Sound]
VP_1V → VP[sem_type=V_Color]
VP_1V → VP[sem_type=V_Form]

В наших правилах каждой именной группе приписывается ее семантическая роль. Например, правило S → NP_Ag VP определяет, что предложение может состоять из именной группы, выступающей в роли агента (NP_Ag), и глагольной группы VP.

Правило NP_Ag → N[last_affix=PL] определяет то, что именная группа может состоять из одного имени существительного, а в квадратных скобках указываются грамматические ограничения, накладываемые на это имя существительное. В данном случае правило указывает на то, что последней аффиксальной морфемой в этом имени существительном должна быть морфема множественности -ЛАр.

Заключение

В этой статье описана работа по созданию семантико-синтаксического анализатора для анализа простого татарского предложения. Работа по созданию семантико-синтаксического анализатора находится в начальной стадии и требует создания большого объема лингвистических ресурсов для татарского языка.

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. Eryigit G., Nivre J., Oflazer K. Dependency parsing of Turkish // Computational Linguistics. – 2008. – Vol. 34, No. 3. – P. 357–389.

3. Бегимтай, У. Х. (2016) Синтаксический анализатор казахского языка на основе грамматики связей — Link grammar parser / У. Х. Бегимтай // Открытые семантические технологии проектирования интеллектуальных

систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2016) : материалы VI междунар. науч.-техн. конф. (Минск, 18 — 20 февраля 2016 года) / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2016. – С. 393 — 396.
4. Жуманов Ж. М. (2012) Разработка грамматики связи для синтаксического анализа казахского языка // Вестн. КазНУ. Серия: Математика, механика, информатика. 2012. № 2 (73). С. 71-80.
5. Шелманов, А.О. Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа: дис. ... канд. техн. наук: 05.13.17 / Шелманов Артем Олегович. – М., 2015. – 210 с.
6. 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.
7. Чапайкина, Н.Е. Семантический анализ текстов. Основные положения / Н.Е. Чапайкина // Молодой ученый. – 2012. – №5. – С. 112–115.
8. 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.
9. Москвина А.Д., Орлова Д., Паничева П.В., Митрофанова О.А. Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK// Компьютерная лингвистика и вычислительные онтологии

СЕНТИМЕНТ АНАЛИЗ КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ МЕЖДОМЕТИЯ

Шарипбай А.А., Қажымұхан Д.А., Кузенбаев Б.А.,
Евразийский национальный университет имени Л.Н. Гумилева,
Астана, Казахстан, sharalt@mail.ru

В статье определяются тональности междометия казахского языка. Показана мета таблица морфологии междометии, построенной по проекту «Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний». Разработан на их основе алгоритм сентимент анализа казахского текста.

Ключевые слова: *сентимент анализ, междометия, тональность, словарный метод.*

SENTIMENT ANALYSIS OF KAZAKH LANGUAGE BASED ON IDENTIFYING INTERJECTIONS TONALITY

A.Sharipbay, D.Kazhymukhan., B. Kuzenbayev,
L.N. Gumilyov Eurasian national university, sharalt@mail.ru

In this article, the key to the interjection of the Kazakh language is determined. A meta-table of the morphology of the interjection, constructed according to the project "Development of Turkic languages electronic thesauri for creation of multilingual search and knowledge extraction systems" is shown. The algorithm of the sentiment analysis of the Kazakh text is developed on their basis.

Key words: *sentiment analysis, interjection, tonality, dictionary method.*

1. Введение

Цифровые технологии сегодня активно развивается и применяются во всех сферах интеллектуальной деятельности человека, в том числе и в компьютерной обработке естественных языков – компьютерной лингвистике. Подготовка и обработка цифровых ресурсов на казахском языке является актуальной проблемой в нашей стране.

Одним из задач обработки текстов естественного языка является сентимент анализ на основе системы извлечения субъективных мнений из них, выраженных поисками тональностей.

В настоящее время социальные сети широко используются в казахско-язычном обществе. Стремительно увеличивается количество записей на казахском языке в социальных сетях, таких как Facebook, Instagram,

Вконтакте и других, что усложняет их анализ обычными методами без автоматизации.

Сентимент анализ позволяет автоматизировать определение тональности сколь угодно объем текстов и оценить мысли автора записей. Важность разработки и применения методов сентимент анализа состоит в том, что анализируя посты пользователей в социальных сетях можно выявить информацию о каких-либо восстаний, террористических актов или оппозиционные мнения, что значительно способствует укреплению безопасности государства.

2. Методы сентимент анализа

Существуют методы, основанные на использовании словаря эмоционально окрашенных слов[4,12] и словаря символов, обозначающих эмоции[6,10,12]. В словарных методах каждое слово обладает весом, характеризующим его эмоциональную окрашенность. Часто словари составляются с помощью сторонних инструментов, таких как WordNet[22], затем термины словаря взвешиваются, например, вручную[10].

В исследованиях, основанных на словарных методах, полярность текста определяется различными способами. В [4] текст обладает тремя независимыми оценками: позитивной, негативной и субъективной. Полярность документа складывается из полярностей его предложений. Определение полярности предложения заключается в вычислении нормализованных сумм весов терминов в тексте: каждый термин имеет положительный, отрицательный и субъективный веса по шкале целых чисел от 0 до 1.

В [12] короткие текстовые сообщения независимо оцениваются по негативной и позитивной шкале: От -5 до -1 и от 1 до 5 соответственно. Каждый термин словаря имеет два веса, определенные этим же способом. Положительный вес текстового сообщения равен весу термина с максимальной положительной оценкой. Если текст не содержит терминов входящих в словарь, ему присваивается минимальный положительный вес. Аналогичным образом определяется отрицательный вес сообщения.

Основной проблемой словарных методов считается процесс составления словаря: чтобы получить метод, классифицирующий документ с высокой точностью, термины словаря должны иметь вес, адекватный предметной области документа. Например, слово «большой» по отношению к объему памяти жесткого диска является положительной характеристикой, но отрицательной по отношению к размеру мобильного телефона.

3. Части речи казахского языка имеющий тональность

При сентимент анализе текстов на казахском языке ведется поиск прилагательных, которые имеют тональность [1]. Для этого проводится морфологический анализ, позволяющий определить части речи текущего слова. После этого формируется словарь корневых слов, имеющих

позитивные или негативные тональностей. Также рассматриваются части речи способные менять тональность других слов или влияющие на тональность рядом стоящих слов.

В казахском языке есть часть речи *междометие* (одағай), который имеет смысловые, морфологические, синтаксические особенности и тональность, влияющая на тон других частей речи в предложении. Междометия выражают различные эмоции человека, боль, тоску, радость, удивление, гнев и другие чувства, побуждают к действию или являются формулами речевого общения.

4. Морфология междометия

Опираясь созданной по проекту «Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» мета таблице морфологии междометии (Рисунок 1) можно увидеть виды междометии, имеющие позитивный или негативный тональность.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Tag	Name_Russian	Name_English		Part of speech	properties	types			types	definition	questions	example
2	1	INTRJ	Междометия	Interjection		Одағай						Одағай сөздер мағына жағынан заттың (субстанцияның) өзі туралы да, сыны, саны, қимылы туралы да, қимылдың жайы-күйі туралы да ұғым бермейді. Олай болса, одағайлар мағына жағынан зат есім, сын есім, сан есім, есімдік, етістік, үстеу сияқтанған негізгі сөз таптарына тән сөздердің бірде-біреуіне ұқсамайды. Осындай ерекшеліктерімен байланысты, одағай сөздер сөйлемде не тұрлаулы, не тұрлаусыз мүшелердің қызметтерін атқармайды, сөйлемдегі басқа сөздермен тікелей синтаксистік қарым-қатынасқа түспейді.	сұрақ қойылмайды	
3			По составу	By composition	структура	81	Құрамы							
4		SIMP	Простые	Simple			811	Дара одағай						
5		UNDR	непроизводные	underivatives			қосымшасы жоқ	8111	Түбір одағай			Бір сөзден жасалған одағай		*А!, О!, Ә!, Ой!, Пай!, Қап!, Бәсе!, Мә!*
6		DRVT	Призводные	derivatives				8112	Туынды одағай			Бұл басында мағыналы сөздерден я сөз тіркестерінен бірігіп, одағай сөздерге айналған		*Мәссаған!, Апырым-ай!, Әттеген-ай, Бәрекелді!*
7		CMPL	Сложные междометия	complex Noun			812	Күрделі одағай				Бұл топқа бірнеше сөзден біріккен, қайталанған, қосарланған немесе басқа тілдерден ауысқан сөздер енеді.		
8		FUSW	слитные(сложные)	Fused				8121	Біріккен			мағыналы екі сөздің бірігуі арқылы туындаған одағай сөз		*Мәссаған!, Бәрекелді!*
9								8122	Қайталанған			екі одағай сөздің қайталануы арқылы жасалатын		*Пай- пай, мәһ-мәһ! шөре-шөре! сарап-сарап! қос-қос/кәуіс-кәуіс! шәуім-шәуім! Әукім-әукім!*
10		PAIR	парные	Pair				8123	Қосарланған					
11			по значению	by meaning	семантика	82	Түрі					Бұл топтағы одағайлар адамның әр түрлі сезімдерін, көңіл-күйін білдіреді. Олар сан жағынан мол, семантикалық құбылу жағынан әрқилы, семантикалық реңкі өте бай топ. Бұл топ- тағы одағайлар әрі жағымды, әрі жағымсыз көңіл- күйді білдіре алады.		
12		EMOT	Эмоциональное	emotional			821	Көңіл-күй одағайлары				Бұл топтағы одағайлар адамның әр түрлі сезімдерін, көңіл-күйін білдіреді. Олар сан жағынан мол, семантикалық құбылу жағынан әрқилы, семантикалық реңкі өте бай топ. Бұл топ- тағы одағайлар әрі жағымды, әрі жағымсыз көңіл- күйді білдіре алады.		
13		POSE	позитивная эмоция	positive emotion				8211	Жағымды көңіл күйді білдіретін			жағымды мағынаға ие одағайлар		1) Алақай! Ура! *(қуаныш, шаттану),* һа! паһ- паһ! (*сүйсіну, таңдану),* Бәрекелді! (*сүйсіну, қоштау),* оho! (*таңдану*) т.б. Мысалы: *Тракторға теңеу таппай, жұрт абыржып, аузына келгенін айтты -* **Пah-пah!** *Күйше күркіреуін қарашы!* (Мұстафин). **Алақай!** *Ата! Ата!... Есбосындар келіп қалды! Соғыс бітті!* (Сланов). -Оho, жолдас болатын болдың (Әбішев) т.б.
		NEGE	негативная эмоция	negative emotion					Жағымсыз эмоцияны білдіретін			жағымсыз мағынаға ие одағайлар		1) Әттеген-ай! әттең! Қап! *(өкіну),* Тәйірі! түге! *(наразылық, ірейіш),* бай-бай-бай! (наразылық, кейістік), піш! пішту! *(жақтырмау, менсінбеу)* т.б. Мысалы: **Қап,** *Қазекеңе бақты суару керек деп айтайын деп отырып, ұмытып кеткенімді қарашы* (Шамкенов). Уай, *қайдағындар,* **түге,** *бар болғырлар!* (Сонда). **Піш,**

Рисунок 1. Мета таблица морфологии междометия

5. Эмоциональные междометия

Тональность эмоционального междометия можно определить опираясь по их смысловым видам. Они могут выражать различные положительные или отрицательные эмоции, а также то или другое

эмоциональное состояние: радость, весёлость, страх, ужас, недоумение, опасение, восхищение и т. д. Соответственно, тональность выражение радости, восторга является позитивным.

Бәрекелді! Алақай!	Браво! Ура!	Бәрекелді, керемет сурет! Браво, отличный рисунок! Алақай, емтихандар бітті! Ура, экзамены закончились!

А также, тональность выражение тревоги, угрозы, презрения, негодования, выражение досады, горя, сожаления являются негативными. Пример показан ниже.

Выражение тревоги, угрозы, презрения, негодования:		
Аттан! Бәлем! Тәйт! Масқара! Тәйірі!	Тревога! Погоди! Цыц! Ужас! О боже! (недовольство, возмущение)	Аттан, жау шапты! Тревога, враги начали атаковать! Бәлем, таяқ жейсің! Погоди, ты у меня получишь! Тәйт, ақырын сөйле! Цыц, говори тихо! Масқара, киімі қандай кір. Ужас, какая грязная одежда. Тәйірі, қызымның бағы ашылмады-ау. О боже, почему же не везет моей девочке.
Выражение досады, горя, сожаления:		
Әттеген-ай! Ойбай! Құдай-ай!	Ах, как жаль! Ой-ой! Боже мой!	Әттеген-ай, кешігіп қалдым-ау! Надо же, опоздал! Ойбай, мынау не? – Ой, что это? Құдай-ай, бұл не дегені! Боже мой, что он говорит!

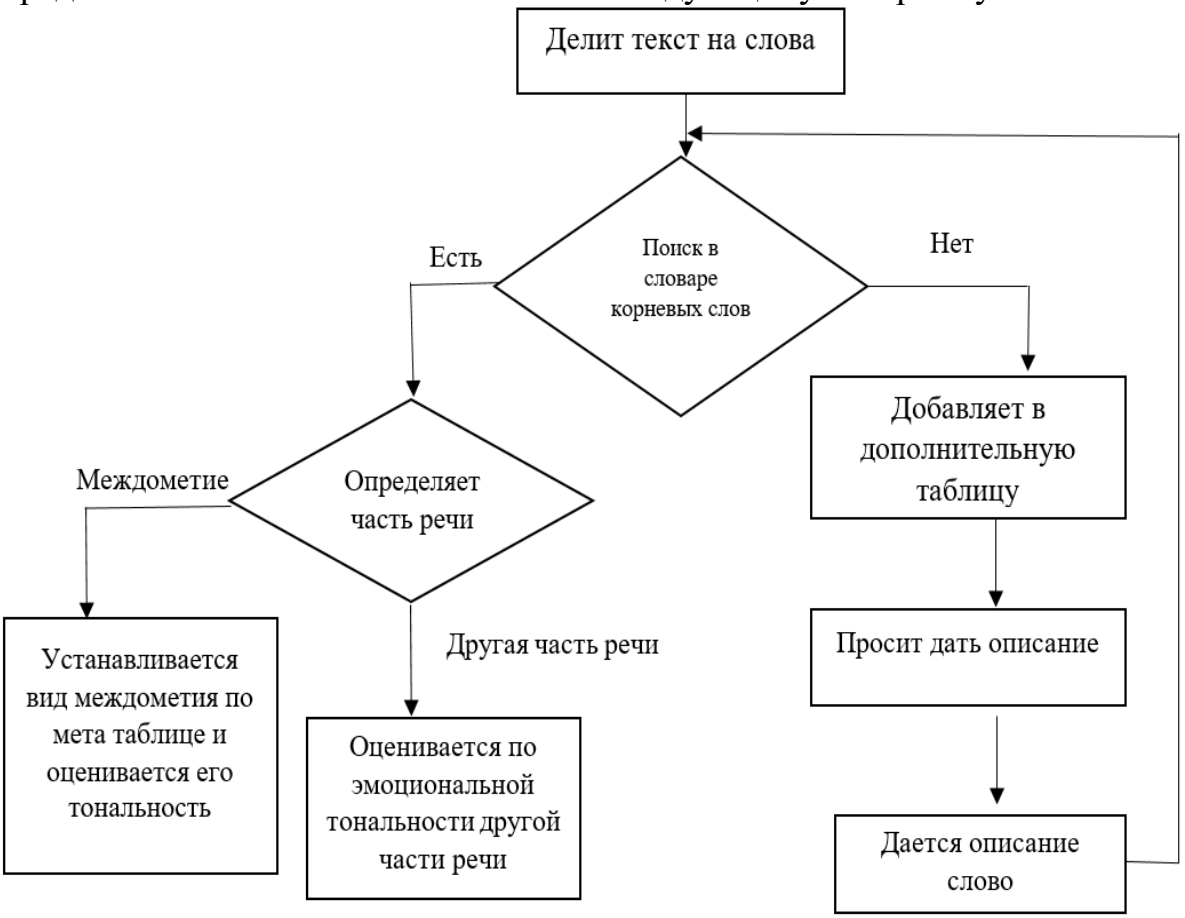
Тональность следующего вида: удивления, одобрения, сомнения зависит от контекста, значит они могут и позитивными или негативными.

Выражение удивления, одобрения, сомнения:		
Aha!	Ага! (негативный)	Aha, айтайын! Ага, скажу!

Бәсе! Мәссаған!	Вот оно как! (одобрение) Вот те на! (удивления)	Бәсе, тынысым енді ашылды. Вот оно как, прям дыхание открылось. Мәссаған, қандай керемет сурет! Вот те на, какая красивая картинка!

6. Алгоритм определения тональности текста при помощи междометия

Программа определения тональности текста опирается словарному методу, который имеет словарь слов с описаниями, части речи и тональность (позитивный, негативный, нейтральный). Также эта программа определяет тональность слов текста по следующему алгоритму.



Заключение

В данной работе показаны исследования по сентимент анализу текстов на казахском языке с использованием словарного метода и междометия, а также мета таблицу.

ЛИТЕРАТУРА:

1. Dmitry Davidov, Oren Tsur and Ari Rappoport. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys// Proceeding of the 23rd international conference on Computational Linguistics (COLING). 2010.
2. Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter Events // Journal of the American Society for Information Science and Technology archive Vol. 62. 2011.
3. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. Sentiment strength detection in short informal text.// Journal of the American Society for Information Science and Technology, Vol., 2544–2558. 2010.
4. WordNet[HTML] (http://wordnet.princeton.edu/)
5. Kerstin Denecke Using SentiWordNet for multilingual sentiment analysis// IEEE 24th International Conference on Data Engineering Workshop. 2008 pp: 507-512.
6. Yergesh, B., Bekmanova, G., Sharipbay, A., Yergesh, M. Ontology-based sentiment analysis of kazakh sentences. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). P.669-677.
7. А.Ысқақов Қазіргі қазақ тілі — 2 басылымы. Филология факультеттері студенттеріне арналған оқулық. – Алматы: Ана тілі,1991. – 384 бет.

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР ЯКУТСКОГО ЯЗЫКА

Леонтьев Н.А., Северо-Восточный федеральный университет им.М.К.Аммосова, Якутск, Россия, leonza@ysu.ru

В данной статье описывается методы и технологии, использованные при создании морфологического анализатора якутского языка. Приводится алгоритм работы анализатора. Описываются проблемы неоднозначностей. Данный морфологический анализатор может являться основой для дальнейшего развития автоматизированной обработки текстов на якутском языке.

Ключевые слова: *морфологический анализ, обработка текста, языковой корпус, якутский язык, разметка текста.*

MORPHOLOGICAL ANALYZER OF THE YAKUT LANGUAGE

Leontyev N.A.,M. K. Ammosov North-Eastern Federal University, Yakutsk, Russia, leonza@ysu.ru

This article describes the methods and technologies used to create the morphological analyzer of the Yakut language. The algorithm of the analyzer is given. The problems of ambiguities are described. This morphological analyzer can be the basis for the further development of automated text processing in the Yakut language.

Key words: *morphological analysis, text processing, language corpus, Yakut language, text markup.*

Автоматизированный морфологический анализ в лингвистике является одним из основных разделов компьютерной обработки текста. Данный вид анализа позволяет определять части слова в тексте и присваивать им соответствующую грамматическую разметку.

Развитие методов морфологического анализа позволило создать анализаторы для широкораспространенных языков. Разработка морфологических анализаторов идет для многих языков народов России, например для чувашского [1], для калмыцкого [2], татарского [3], тувинского [4], для хакасского [5]. В других странах ситуация работа с

тюркскими языками, такими как турецкий, казахский и другие, идет большим темпом.

Якутский язык (саха тыла) относится к тюркским языкам, количество носителей более 430 тыс.человек. На якутском языке имеется большое количество литературы, периодических изданий, видео- и аудио-материалов.

Для морфологического анализа была создана систему аннотирования якутского языка[6]. Данная система аннотирования была преобразована в базу данных. Были созданы база данных аффиксов, база правил согласования окончаний, база корневых слов. На основе этих баз данных был создан алгоритм и программа для морфологического анализа якутского языка. Для нахождения корня слова использовали алгоритм, описанный в работе [8]. Электронный корпус также можно аннотировать вручную с помощью многопользовательского доступа [9].

Данный морфологический анализатор работает по следующему алгоритму: Идет поиск словарного слова, в случае нахождения такого слова выводим слова, но поиск аффиксов продолжается. Производится поиск последнего аффикса, в случае нахождения аффикса, производим удаление последнего аффикса и производим поиск корневого слова с учетом изменения окончания. Данный алгоритм может корректно искать и заимствованные слова.

В результате получается цепочка аффиксов с учетом их изменений. В случае неоднозначности получается несколько цепочек аффиксов.

Таблица 1.

ПРИМЕР НЕОДНОЗНАЧНОСТИ

Словоформа	Корень	Значение
ыт	ыт	Собака
	ыт	Стрелять
ыттар	Ыт+тар	Собаки *мн.ч.*
	Ыт+тар	Дай пострелять *гл.*
	Ытын+тар	Карабкается *гл.*

Для доступа к морфологическому анализатору был написан программный код для Интернет-сайта. Программный код написан на языке PHP, это облегчает интеграция кода в оболочку веб-сервера. База данных

основана на СУБД MySQL. Это позволяет получать доступ к системе удаленным пользователям.

Для тестирования был использован корпус якутского языка [7], в ходе которого выяснили, что процент корректности составляет около 80%. На корректность влияет наличие заимствованных слов, незнакомых слов, неоднозначности.

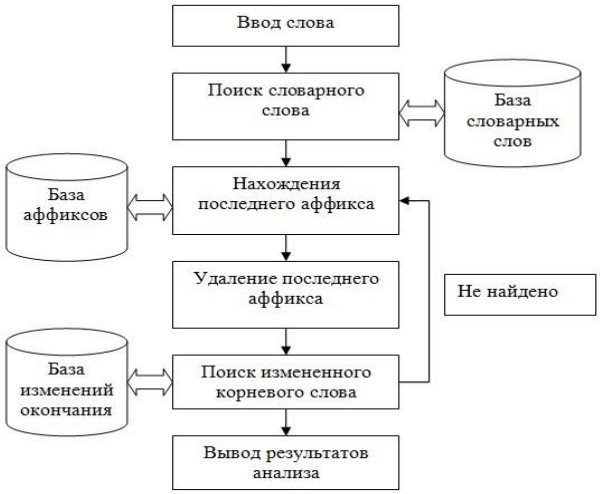


Рис.1. Блок схема алгоритма поиска аффиксов

Необходимо различать аффиксы для глаголов и остальных частей речи, имеются неизменяемые части речи, которых необходимо помечать отдельно. Встречаются неоднозначности, которые не решается текущим алгоритмом. Для решения неоднозначностей необходимо применять N-граммы и скрытые Марковские цепи.

ЛИТЕРАТУРА:

1. Желтов П.В. Морфологический анализатор национального корпуса чувашского языка // В сборнике: Совершенствование методологии познания в целях развития науки. сборник статей по итогам Международной научно-практической конференции : в 2 ч.. 2017. С. 11-13.
2. Куканова В.В., Каджиев А.Ю. Алгоритм работы морфологического парсера калмыцкого языка // В сборнике: Писменото наследство и информационните технологии. El'Manuscript–2014 Материали от V международна научной конференции. Отговорни редактори В. А. Баранов, В. Желязкова, А. М. Лаврентьев. 2014. С. 116-119.
3. Гатиатуллин А.Р., Баширов А.М., Осипов Г.С., Смирнов И.В., Шелманов А.О. Методы лингвистического анализа текстов на татарском языке и их применение в поисковой системе EXACTUS // Труды Института

системного анализа Российской академии наук. 2016. Т. 66. №1. С. 18-25.
4. Хертек А.Б., Ооржак Б.Ч.О. О морфологической разметке электронного корпуса текстов тувинского языка // Филологические науки. Вопросы теории и практики. 2012. №7-2(18). С. 214-218.
5. Дыбо А.В., Шеймович А.В. Аппарат автоматического морфологического анализа для корпуса хакасского языка // Родной язык: лингвистический журнал. 2016. № 2 (5). С. 9-39.
6. Торотоев Г.Г., Ноговицына А.Н. Лингвистическое аннотирование наклонений глагола якутского языка // Вестник Северо-Восточного федерального университета им. М.К. Аммосова. 2017. №3 (59). С. 108-120.
7. Leontiev N. The newspaper corpus of the Yakut language // В сборнике: Сборник трудов международной конференции TURKLANG-2015 Tatarstan Academy of Sciences L.N. Gumilyov Eurasian National University Ministry of Education and Science of the Republic of Kazakhstan Kazan Federal University Institute of Philology and Intercultural Communication. 2015. С. 233-235.
8. Леонтьев Н.А. Вопросы автоматизированной лемматизации якутского языка // Современная наука: актуальные проблемы теории и практики. Серия: Познание. 2015. №2 (41). С. 15-16.
9. Леонтьев Н.А., Торотоев Г.Г. Многопользовательская морфологическая разметка корпуса якутского языка // В сборнике: Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы Сборник материалов Международной научной конференции. 2017. С. 101-103.

COMPUTATIONAL ANALYSIS OF UZBEK NOUNS

M.Orhun[6],Technical University of Istanbul, Istanbul, Turkey, murat.orhun@bilgi.edu.tr

This paper explains implementation of possible morphological analyzer of Uzbek Language which is spoken in Uzbekistan. In natural language processing, solving morphological analyzer of a language is the first step. Uzbek language is an agglutinative language and it has productive inflectional and derivational suffixes. Though there are different methods for implementing an analyzer for a language, but two-level morphological analyzer is suggested in this paper. Some elementary implementation process has been explained with examples.

Key words: *Uzbek Grammar; Uzbek Nouns; Turkic Languages; Machine Translation .*

ВЫЧИСЛИТЕЛЬНЫЙ АНАЛИЗ УЗБЕКСКИХ СУЩЕСТВИТЕЛЬНЫХ

М. Орхун, Стамбульский технический университет, Стамбул, Турция, murat.orhun@bilgi.edu.tr

В этой статье рассматривается реализация возможного морфологического анализатора узбекского языка, на котором говорят в Узбекистане. Первым шагом в обработке естественного языка является решение морфологического анализатора. Узбекский язык является агглютинативным языком и имеет продуктивные инфлекционные и деривационные суффиксы. Хотя существуют разные способы реализации анализатора для языка, в данной статье предлагается двухуровневый морфологический анализатор. Некоторые элементарные процессы реализации были объяснены с помощью примеров.

Ключевые слова: *Узбекская грамматика; Узбекские существительные; Тюркские языки; Машинный перевод.*

1. Introduction

Uzbek language is spoken in Uzbekistan about by 30 million people. Except this, it is spoken in Kazakhstan, Kyrgyzstan, Turkmenistan, Afghanistan, Xin Jiang Uyghur Autonomous region in China, Turkey, Saudi Arabia and in some of western countries. It is the second language in Turkic language family which is spoken more people after the Turkey Turkish language. After the Soviet revolution, some properties of Uzbek language are changed in Uzbekistan. Therefore, there are some

difference between the Uzbekistan Uzbek language and the Uzbek language in other of part of the world. Even in Uzbekistan there are some dialects exist. In this paper the Tashkent dialect is studied mainly. The Uzbek people used Cyrillic alphabet till 1992. After Uzbekistan got its independence, Latin alphabet has been accepted its new national alphabet. The new alphabet consists of characters as shown in Figure 1.

Aa	Bb	Dd	Ee	Ff	Gg	Hh	Ii	Jj	Kk	Ll	Mm	Nn	Oo	Pp
[æ/a]	[b]	[d]	[e]	[f]	[g]	[h]	[i/ɨ]	[ʒ/dʒ]	[k]	[l]	[m]	[n]	[o/ɒ]	[p]

Qq	Rr	Ss	Tt	Uu	Vv	Xx	Yy	Zz	O'o'	G'g'	Sh sh	Ch ch	Ng ng
[q]	[r]	[s]	[t]	[y/u]	[w]	[x]	[j]	[z]	[ø/o]	[ɣ]	[ʃ]	[tʃ]	[ŋ]

There are 23 consonants and 6 vowels have been defined in the new script. Actually the /u/ and /o‘/ characters have been used to represent two different characters such as /u/, /ü/, /o‘/ and /ö/. Therefore, there are 8 vowels in modern Uzbek language. Why there are not 8 different characters used to represent Uzbek vowels is a hot topic today. But it is not the topic of this paper.

The Uzbek language is agglutinative (Usmonava, Azlorov, Sharipov, 1991). It is a language whose words are generated by adding affixes to the root words. In agglutinative languages a new word could be derived by adding an affix to the root or previously generated words.

For example:
maktab
maktablar
maktablarim
maktablarimda
maktablarimdagi
maktablarimdagilar
maktablarimdagilarni
maktablarimdagilarniki
maktablarimdagilarnikidek
maktablarimdagilarnikidekmi
.....

The root of the word is "maktab" (school) and new words have been derived with adding affixes and according to Uzbek grammar, new affixes could be added to the chain. Last word could be translated into English such as: Are they whom look like belong to our schools.

In general, in Uzbek language there is not such a long word. But this word is a valid word according to the Uzbek morphotactic rules. Therefore, it is possible to create unlimited number of words in agglutinative languages and it is not possible to write a dictionary to explain all words.

Because in this reason, to analyze a word computationally, especially a word which belong to agglutinative language, it is the first step to implement a morphological analyzer.

The first morphological analyzer for Uzbek language is has been implement with Prolog (Matlatipov, Veulani, 2009). In this analyzer some rules have been defined to solve consonant harmony problem, vowel deletion situation, doubling consonants and possessive suffixes cases etc. There are 1000 root words and 108 suffixes have been included. At the same time some irregular words also have been considers and tags defined to explain morphemes of the words.

In this paper a different approach, two-level analyzer, has been suggested to analyze Uzbek words and some initial works demonstrated.

In this paper, a short introduction has been given about Uzbek language in the first section. In the second section two-level morphological analyzer and related works have been explained. In the third section, implementation of two-level morphological for Uzbek language is explained and some two-level rules introduced. A simple evaluation and suggestion has been done in the last section.

2. Two Level Morphology

Two-level morphology is a widely used technique in morphological analysis. Especially it is very useful analysing agglutinative languages. Most of the two-level morphological analysers have been implemented with KIMMO (Karttunen, 1983) or Xerox (Karttunen, Gaal, Kempe, 1997) tools that based on final state automata. These tools completely independent of any languages. For example: Turkish (Oflazer, 1994), Crimean Tatar (Altintas, Çiçekli, 2001), Qazan Tatar (Görmez, Kurt, Kulamshaev, Kara, 2011), Turkmen (Tantuğ, Adalı, Oflazer, 2006), Kazakh(Kessikabayeva, Çiçekli, 2014, Zafer, Tilki, Kurt, Kara, 2011), Kyrgyz (Görmez, Baki, Kurt, Kulamshaev, Kara, 2011, Washington, Ipasov, Tyers, 2012), Uyghur (Orhun, Tantuğ, Adalı, 2009a, 2009b). Even all of these languages belong to the Turkic language family, there are some morphological analysers exit belong to different language family such as English (Karttunen, Wittenburg,1983), Finnish (Koskenniemi,1985), Japanese (Alam, 1983) and Korean (Kim, Lee, Choi, Kim, 1994) etc.

Two-level morphological analysers work into two directions such as lexical level to surface level, and surface level to lexical level. Because of this reason it is most dominant method of about morphological analysers.

There are two levels called lexical and surface levels. In the surface level, a word is represented in its original orthographic form. In the lexical level, a word is represented by denoting all the functional components of the word.

The phonetic modifications and restrictions are represented using four

different rule types [8]. These rules are given in Table 1.

For example: for the first rule, suppose there is a rule which defined such as: a:b=> LC_RC .

This rule explains, if there is a /a/ character appears between LC (left context) and RC (right context), then the /a/ character must be changed into the /b/ character. But it is not necessary. It means, if there is a need to change, then this change must happen only between LC and RC. Another place cannot happen. As a result, the first rule says, the /a/ character maybe changed into the /b/ character between the LC and RC, or maybe not.

Table 1.

THE PHONETIC MODIFICATIONS AND RESTRICTIONS RULE

a:b => LC_RC	/a/ is realized as /b/ only in the context LC (left context) and RC (Right context), but not necessarily.
a:b <= LC_RC	/a/ is always realized as /b/ in the context LC and RC
a:b <=>LC_RC	/a/ is always realized as /b/ in the context LC and RC and nowhere else
a:b /<=LC_RC	/a/ is never realized as /b/ in the context LC and RC

Therefore, the lexical /a/ can appears as /a/ or /b/, even both at the surface level. If we have «LbR» as a word at the lexical level, then the output at the surface level is: «LbR», it means doesn't change. User defined rules of these types are used to generate a finite state acceptor. This machine accepts a legal lexical surface matching whereas it rejects an illegal matching.

3. Two Level Rules for Uzbek Language

In order to define two-level rules, characters in an alphabet should be represent with single characters. In general, it is more practical and understandable to write rules with single and Unicode characters. There are five characters in Uzbek alphabet that consists of double characters such as /sh/, /ch/, /ng/, /o‘/ and /g‘/.

In two level rules, all characters in the alphabet are represented with lower case character and those five characters are replaced with following characters.

/sh/-> /ş/, /ch/ -> /ç/, /ng/->/N/, /g‘/->/ğ/ and the /o‘/->/ö/

After this replacement, two-level morphological analyser uses following meta alphabet that consist of following characters.

Consonants (CONS): b d f g h j k l m n p q r s t v x y z ğ ş ç N

Vowels (VOWEL): a e i o u ö

According to Uzbek constant properties, constants could be put into following sub categories.

Soft constants (CONSS): ç b t

Harsh constants (CONSH): l n r N

To harmonize the /q/ character in ablative case:

CONS2: b d f g h j l m n p r s t v x y z ş ç N

Apart from this, there are some meta character have been defined that available only at lexical level and they do not appear at surface level.

Lexical Vowel:

H = i, 0 (zero character)

Lexical consonants:

G = k, q

K = g, k

In Uzbek languagepossessive suffixes have following cases.

(i)m, (i)ng, i, (i)miz, (i)ngiz, lari

If a word terminate with vowel, then just add: "m", "ng", "si", "ngiz" and "lari", otherwise if terminate with a constant, then first add /i/ character, then add "m", "ng","si", "miz" and "ngiz" suffixes.

This rule could be written with two-level rules as below:

Rule-1 H:0 <=> VOWEL _ [m | N | l]

Rule-2 H:i <=> CONS _

Rule-3 s:0 <=> CONS _ :i

After these two rules have been defined, possessive suffix could be written at the lexical level in below, for example:

bosh+(H)m -> boshim.

In this example, because of the root word "bosh" is terminated with a consonant, which is /h/, the meta character /H/ is converted into /i/. And the suffix "m" is added correctly. In this case the second rule is applied.

baba+(H)m -> babam

In this example, the root word "baba" is terminated with a vowel the first rule is will be applied and the meta character /H/ is converted into /0/. The /0/ (zero) character is not visible at the surface level. As a result, on the surface the word "babam" will be displayed.

If a word terminated with constant and need to add the person singular person possessive, then the first rule applied directly. For example:

bosh+(H) [m | N | l] -> boshi.

In this rule, the square parenthesis holds optional parameters. It means, even there are not /m/, /N/ or /l/ characters, the meta character /H/ will be converted into the /i/ character.

In case a word terminated with a vowel and need to be followed by a third person singular suffix, the /s/ character insert first, then the /i/ character comes next. In this situation the third rule is applied. For example:

baba+s+i= babasi

To add the possessive suffix /i/, the "s" character is inserted in front of the /i/ character.

There is there suffix exist for data case in Uzbek: "ga", "ka" and "qa".

To add the them correctly, the meta character /G/ is defined and it changes

according to following rules:

Rule-4 G:k <=> k _

Rule-5 G:q <=> q _

ğ: _

Rule-6 G:g => CONS2 _

Rule-7 ğ:q => _ :q

In two-level morphology analyser, all rules are executed in parallel. Therefore, the rules should be not conflicted each other.

For example: bağ+Ga -> baqqa

choq+Ga -> choqqa

yigit+Ga -> yigitga

xotin+H+m+Ga -> xotinimga

4- Conclusion

At this level of this paper, there are seven rules have been defined. With these rules plural, possessive and some case suffixes have been analysed correctly. In order to tag these morphemes, tags that used for other Turkic languages have been used (Oflazer, 1994). For example:

maktab

maktab+Noun+A3sg+Pnon+Nom

It means, root of the word is "maktab", it is a "noun", "singular", "non-possessive", and doesn't take any "case" suffixes.

maktablar

maktab+Noun+A3pl+Pnon+Nom

It means, root of the word is "maktab", it is a "noun", "plural", "non-possessive", and doesn't take any "case "suffixes.

maktablarim

maktap+Noun+A3pl+P1sg+Nom

It means, root of the word is "maktab", it is a "noun", "plural", "first person possessive", and doesn't take any "case" suffixes.

maktablarimda

maktap+Noun+A3pl+P1sg+Loc

It means, root of the word is "maktab", it is a "noun", "plural", "first person possessive", and take "locative case "suffix.

This is an ongoing research and more rules must be defined in order to define full functional morphological analyser for the Uzbek language.

REFERENCES:

1. Oflazer, K., (1994). *Two-Level Description of Turkish Morphology. Literary and Linguistic Computing*, Vol. 9, No:2 (1994).

2. Altintas K., Çiçekli İ., (2001). *A Morphological Analyzer for Crimean Tatar*, Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, pp: 180-189.

3. Tantuğ A. C., Adalı E. and Oflazer K., (2006). *Computer Analysis of the Turkmen Language Morphology*, Fin-TAL, Lecture Notes in Computer Science, 4139, pp:186-193.
4. Orhun M., Tantuğ A. C. and Adalı E., (2009a). *Rule Based Analysis of the Uyghur Nouns*, International Journal on Asian Language Processing, 19,1, pp: 33-43.
5. Orhun M., Tantuğ A. C. and Adalı E., (2009b). *Rule Based Tagging of the Uyghur Verbs*, Fourth International Conference on Intelligent Computing and Information Systems, Faculty of Computer & Information Science, Ain Shams University Cairo, Egypt, pp: 811-816.
6. Görmez Z., Kurt A., Kulamshaev K. and Kara M., (2011). *Two-level Qazan Tatar Morphology*, The 1st International Conference on Foreign Language Teaching and Applied Linguistics, Sarajevo, Bosnia, pp: 428-432.
7. Zafer H.R., Tilki B., Kurt A. and Kara M., (2011). T*wo-level Description of Kazakh Morphology*, The 1st International Conference on Foreign Language Teaching and Applied Linguistics. Sarajevo, Bosnia, pp: 560-564.
8. Kessikabayeva, G. and Çiçekli İ., (2014). *Rule Based Morphological Analyzer of Kazakh*, Language Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. Baltimore, Maryland, USA, pp: 46–54.
9. Görmez Z., Baki Ü.S., Kurt A., Kulamshaev K. and Kara, M., (2011). *An overview of Two-level Finite State Kyrgyz Morphology*, The 2nd International Symposium on Computing in Science & Engineering. Kuşadası, Aydın, Turkey, pp:48-52.
10. Washington J.N., Ipasov M., and Tyers, F.M., (2012). *A Finite-state morphological transducer for Kyrgyz*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp: 934-940.
11. Matlatipov G. and Veulani, Z., (2009). *Representation of Uzbek Morphology in Prolog*, Aspects of Natural Language Processing, Springer-Verlag, Berlin. p:83-110.
12. Karttunen, L. Gaal, L. and Kempe, A., (1997). *Xerox Finite State Tool*. Technical Report, Xerox Research Centre, Europe.
13. Karttunen, L., (1983). KIMMO: *A General Morphological Processor*, in Texas Linguistic Forum, Texas, USA.
14. Ercilasun B.A., Aliyev A.M., Şayhulov A., Kajibek E. Z., and UULU K.K, (1991). *Karşılaştırmalı Türk Lehçeleri Sözlüğü I*, Kültür Bakanlığı Yayınları, Ankara,
15. Alam Y. S., (1983). *A Two-Level Morphological Analysis of Japanese*, Texas Linguistics Forum, Texas, USA, pp: 229-252.
16. Karttunen L. and Wittenburg K., (1983). *A Two-Level Morphological Analysis of English*, Texas Linguistics Forum, Texas, USA; pp: 217-228.
17. Koskenniemi K., (1985). *An Application of the Two-Level Model to Finnish*,

In Fred Karlsson, editor, Computational Morphosyntax, a report on research 1981-1984, University of Helsinki Department of General Linguistics, pp: 19-41.
18. Kim D.B., Lee S. J., Choi K. S. and Kim G. C., (1994). *A Two-Level Morphological Analysis of Korean*, COLING‘94 Proceedings of the 15th conference on Computational linguistics, Volume-1, Kyoto, Japan, pp: 535-539.
19. Usmonava, Z., Azlorov, E., Sharipov, Gh. (1991). *O‘zbek Tili. Toshkent. O‘qituvchi*.
20. Languagesgulper, (2018) http://www.languagesgulper.com/eng/Uzbek.html, accessed on 07/08/2018.

A PRELIMINARY WORK ON SENTENCE TOKENIZATION USING REINFORCEMENT LEARNING AND LINGUISTIC RESOURCES

Z. Yessenbayev, National Laboratory
Astana, Kazakhstan, zhyessenbayev@nu.edu.kz

In this preliminary work, we propose to develop a novel hybrid segmentation-based framework for word-level tokenization using reinforcement learning (RL) and linguistic resources (LR) for Kazakh. In the initial stage, our aim is to implement and train a neural network that can incorporate and perform at least as good as the existing LR but outperform the state-of-the-art pure deep learning-based systems. The problem of outperforming LR and generalization of the obtained knowledge can be viewed as a future work.

Key words: *linguistic resources, Kazakh, segmentation, tokenization.*

ПРЕДВАРИТЕЛЬНАЯ РАБОТА ПО ТОКЕНИЗАЦИИ ПРЕДЛОЖЕНИЯ С ИСПОЛЬЗОВАНИЕМ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ И ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

Ж. Есенбаев, Национальная лаборатория,
Астана, Казахстан, zhyessenbayev@nu.edu.kz

В этой предварительной работе мы предлагаем разработку новой гибридной системы, основанной на сегментации для токенизации на уровне слов с использованием обучения с подкреплением (ОсП) и лингвистических ресурсов (ЛР) для казахского языка. На начальном этапе наша цель состоит в том, чтобы внедрить и обучить нейронную сеть, инкорпорирующую и функционирующую, по крайней мере, так же хорошо, как и существующие ЛР, но превосходящую современные системы, основанные только лишь на глубоком обучении. Проблему опережающих ЛР и обобщения полученных знаний можно рассматривать в качестве основы для дальнейшей разработки.

Ключевые слова: *лингвистические ресурсы, казахский язык, сегментация, токенизация.*

Introduction

In the recent decade, there was a tremendous progress in the field of speech and language processing with the advent of deep learning approach and availability of computing resources. In particular, unsupervised pre-training, attention mechanism, recurrent and convolutional neural networks, word

embedding are the standard tools to tackle most of the problems in the natural language processing. However, most of the approaches ultimately exploit supervised learning, which are not suitable for low resource languages or new domains. Also, the problem of complex morphologies such Kazakh or Russian is not solved yet with the deep learning approach. The best what has been done is the character based language modelling which is still error prone. Additionally, there is no meaningful approach for out-of-vocabulary (OOV) words. On the other hand, there is a large body of work on rule-based language resources such as finite state morphological analyzers and syntactic parsers which may show better performance in the given task, but are not utilized well. One of the problems, apart from the ideological ones, is the integration of these resources in the framework of deep learning.

In this preliminary work, we propose to develop a novel hybrid segmentation-based framework for word-level tokenization using reinforcement learning (RL) and linguistic resources (LR) in order to reduce error rate caused by OOV words. In the initial stage, our aim is to implement and train a neural network that can incorporate and perform at least as good as the existing LR but outperform the state-of-the-art pure deep learning-based systems. The problem of outperforming LR and generalization of the obtained knowledge can be viewed as a future work.

Related Work

Our proposed work is related to a wide spectrum of research work such as language modeling, sequence segmentation, reinforcement learning, rule-base analyzers. Here we will just mention few of them due to space limitation. Learning morphology and syntax was always a challenging problem for morphology rich and free word order languages. In particular, there are many attempts to incorporate linguistic information into the framework of deep learning [1-3]. Most of them rely on text corpus specially prepared by existing LR. In our approach we also use existing LR to segment character or word steams into relevant linguistic units, but we don’t use labeled text corpora. In this sense, our approach is more semi-supervised. In our work, we will use existing supervised and unsupervised approached to see whichever achieves best results. Reinforcement learning became a hot topic in deep learning after its application to the game playing by Google DeepMind [4, 5].

Methodology

The proposed hybrid segmentation-based framework can be defined as follows.

Let $X = \{x_1, x_2, \ldots, x_T\}$ be an input sequence of characters which encodes some sentence to be tokenized in a given language L. Additionally, we are given a linguistic resource (LR) of this language such as a vocabulary or morphological analyzer.

Then, in the context of reinforcement learning (RL), we define a segmentation MDP (Markov Decision Process) to be a process where an agent reads the input sequence X character by character and at time t decides on the action a_t – to put a boundary or not for some observing segment s_t, i.e. a state according to some policy π (see Fig. 1). In return, the agent obtains a positive reward r_t if the segment complies with the given LR, and negative (or zero) reward otherwise. The process continues until the whole sequence is processed and the output segmentation is returned. The goal of the agent is to maximize the expected return $E[R_t]$ from each state s_t, where $R_t = \sum_{k=0}^{T} \gamma^k r_{t+k}$ – is the total accumulated return from time step t and $\gamma \in (0,1]$ is a discount factor.

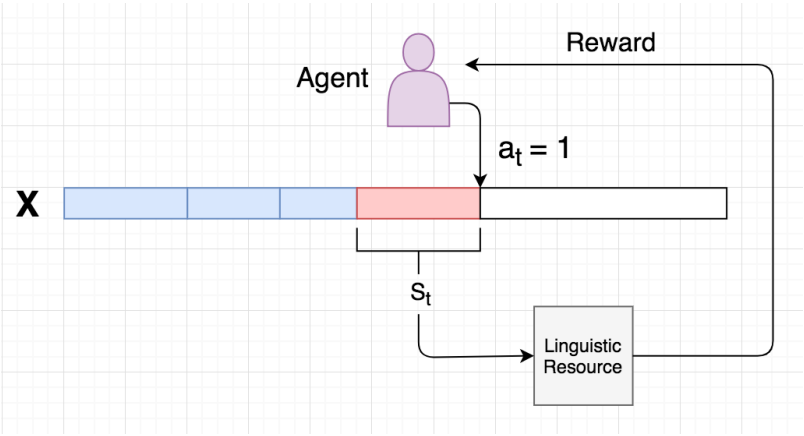


Figure 1 – A segmentation MDP

The action value function $Q^{\pi}(s,a) = E[R_t|s_t = s, a]$ is the expected return of choosing the action a in state s and following the policy π. One approach to approximate the action value function is to use some function approximator such as a neural network $Q(s,a,\theta)$, where θ are the parameters of the neural network. The parameters θ can be updated using the algorithm called Q-learning [4], which aims to directly approximate the optimal action value function $Q^*(s,a) = max_{\pi} Q^*(s,a)$. The loss function for Q-learning is defined as

$$L_i(\theta_i) = E(r + max_{a'} Q(s', a', \theta_{i-1}) + Q(s,a,\theta_i)),$$

where s' is the next state after s, i is the current iteration. For the neural network we employ Long-Short Term Memories (LSTM) as it has proven to be efficient in modelling sequential data [2, 6]. The input for the LSTM network will be a current segment s_t, and the output is the action a_t with the maximal value.

Experiments and Results

For the tokenization experiments, we prepared raw sentences without any labeling but with the spaces removed, e.g.: «Меніңүйімалыста» -> «Менің үйім алыста». Additionally, we compiled a vocabulary as a linguistic resource.

We successfully implemented Q-learning based RL algorithm using

TensorFlow toolkit to tokenize a character stream using only vocabulary. Instead of vocabulary we could use morphological analyzer what would expand words set without adding complexity to the system.

The LSTM neural network has two layers with 100 neurons in each layer. The character steam is modeled as one-hot vectors where the window of 3-7 vectors was supplied to the network as an input. The network performs actions until all tokens are in vocabulary or maximum number of iteration is reached. The weights are updated using the stochastic gradient descent algorithm.

Despite the simplicity of the model, it turns out that the training time is very slow taking about 30-40 min/sentence/iteration on a single CPU machine. That prevented us from conducting the experiments with the whole data set. Nonetheless, the model was able to segment a sentence into words within reasonable time frame. We also tried to reorder the words in the sentence to assess the generalization capability of the model. The results showed that the network is still able to segment the sentence but its performance slightly depends on the size of context window, yielding better results for the context window of smaller size (namely, 3 vectors).

Conclusion and Future work

In this preliminary work we have implemented a novel framework using reinforcement learning approach for sentence tokenization task. The result shows that the proposed method is works, but it is slow. Therefore, as a future work, we plan to implement and train asynchronous models for reinforcement learning such as policy-based RL algorithms [5]. To conduct and reduce lifecycle of the large scale experimentations on deep learning, significant computing resources are need. These include servers with several pre-install general purpose GPU cards. It is expected that the university will buy this type of resources in the next year. However, it is also possible to rent the resources from third parties like Amazon or Google.

Moreover, to fully appreciate the impact of linguistic resources, we will integrate rule-based morphological and syntactic analyzers [7].

For the experimentation, we plan to evaluate our proposed framework on the following datasets.

- TIMIT – English continuous read speech corpus [8]. English has Subject-Verb-Object (SVO) sentence structure and simple morphology.
- KLC — Kazakh read speech corpus [9]. Kazakh has SOV sentence structure and complex agglutinative morphology. This corpus will help assess the capability of the framework to model complex morphology.

As an outcome, it is expected to obtain the following results: 1) An open source Python package which implements the proposed framework; 2) Evaluation of the framework on the TIMIT and KLC corpora.

Apart from these results, the outcome of the proposed work will provide insights on the capability of deep learning approaches to model complex

structured objects such as morphology and syntax, and, consequently, have an impact on the general field of speech and language processing. On the other hand, a successful implementation of the proposed unified framework will give a second rise to the research directions of rule-based approaches which have been put aside after the recent breakthroughs in deep learning.

REFERENCES:

1. Jan A. Botha and Phil Blunsom. Compositional morphology for word representations and language modelling. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Vol. 32. JMLR.org II-1899-II-1907.
2. F. Dalvi, et al. Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder, Proceedings of the Eighth International Joint Conference on Natural Language Processing, pp. 142-151, 2017
3. Joel Legrand and Ronan Collobert, Deep Neural Networks for Syntactic Parsing of Morphologically Rich Languages, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 573–578, 2016.
4. V. Mnih, et al., Playing atari with deep reinforcement learning. In NIPS Deep Learning Workshop. 2013.
5. V. Mnih, et al., Asynchronous methods for deep reinforcement learning. In Proceedings of the 33rd International Conference on International Conference on Machine Learning — Volume 48, ICML'16, pages 1928–1937. JMLR.org, 2016.
6. J. Chorowski, et al., Attention-Based Models for Speech Recognition, CoRR, abs/1506.07503, 2015
7. A. Sundetova, et al., A free/open-source machine translation system for English to Kazakh. In Proceedings of Turklang 2015, 2015.
8. J. S. Garofolo, et al., DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993.
9. O. Makhambetov, et al., Assembling the Kazakh Language Corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1022–1031. 2013.

ОБ ОПТИМИЗАЦИИ АЛГОРИТМА МОРФОЛОГИЧЕСКОГО АНАЛИЗА

Т. Садыков[1], Б. Кочконбаева[2],[1]Бишкекский гуманитарный университет им. К.Карасаева, Бишкек, Кыргызстан, [2]Ошский технологический университет, Ош, Кыргызстан, tash_sadykov@mail.ru, buajar@mail.ru

В статье ставится задача сокращения числа обращений к дисковой памяти компьютера при автоматизированном морфологическом анализе естественного текста. Рассматриваются вопросы создания модели образования текстовых форм существительных для кыргызского языка. При этом используется словарный метод, а для управления данными создается база данных в среде Access.

Ключевые слова: *морфологический анализатор, дисковая память, аффикс, основа слова, алгоритм, фрейм-модель, естественный язык.*

ABOUT OPTIMIZATION ALGORITHM OF MORPHOLOGICAL ANALYSIS

T. Sadykov[1], B. Kochkonbaeva[2], [1] Bishkek humanities university, Bishkek, Kirgizstan, [2]Osh technological university, Osh, Kirgizstan, tash_sadykov@mail.ru, buajar@mail.ru

The article poses the problem of reducing the number of calls to the disk with an automated morphological analysis of the natural text. And also questions of creation of the model of formation of noun words are considered. When creating a morphological analyzer, a dictionary method was used and a database was created in the Access environment to manage the data.

Key words: *morphological analyzer, disk memory, affix, the basis of the word, the system,frame-model, natural language.*

Постановка задачи. Известны по крайней мере три способа создания морфологического анализатора: (1) анализатор, основанный на словарь, (2) анализатор, основанный на грамматику без словаря и (3) анализатор на базе грамматики и словаря.

В сообщении обсуждаются вопросы, связанные с созданием морфологического анализатора на базе на словаря. Для этой цели используются факты и правила кыргызского языка, где из всех языков алтайской семьи наиболее развиты законы сингармонизма.

Как известно, кыргызский язык входит в группу агглютинативных языков, где новые слова образуются с помощью словообразовательных аффиксов, а их грамматические формы — с помощью аффиксов

словоизменительных. Здесь к основе слова в соответствии с правилами сингармонизма прибавляются окончания разной огласовки. Например: тоо+Ø=тоо, тоо+нын=тоонун, үй+нын= үйдүн и т.д.

В кыргызском языке текстовые формы существительных образуются с помощью довольно строгих правил агглютинации морфем, как это показано в фрейм-модели следующего вида (рис.1). При этом узлы модели указывают на разные состояния морфологии существительного, а линии соединяющие эти узлы – на конкретные категории имен существительных. N — это основа исследуемого существительного.

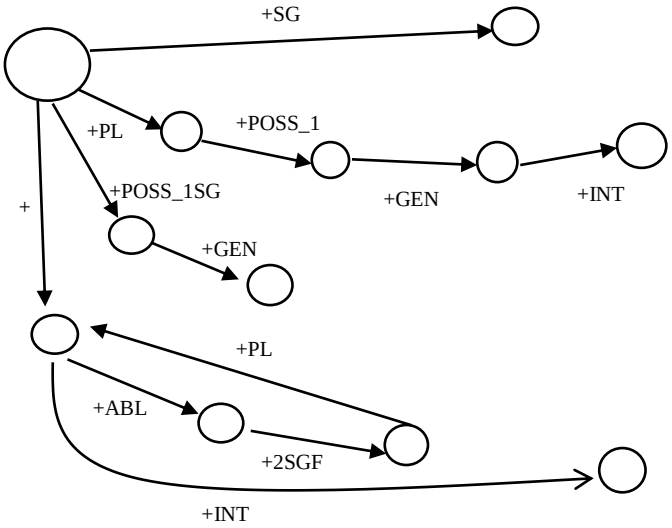


Рис. 1. Фрейм-модель формирования грамматических форм

Так, на базе этого фрейма рассмотрим схему формирования форм существительного китеп — книга:

- китеп+SG
- китеп+PL+ POSS_1SG + GEN + INT
- китеп+ POSS_1SG+1GEN
- китеп+PL+ABL+2SGF+PL+INT

Таким образом, при рассмотрении принципов работы морфологического анализатора естественного текста необходимо учесть следующие этапы:

1. Разделение входного текста на грамматические формы слов.
2. Лемматизация текстовых форм слов в их словарную форму.
3. Разделение цепочки словоизменительных аффиксов на конкретные аффиксы.
4. Выявление морфологических признаков каждого из аффиксов.

Например: при анализе текстовой формы слова балдар (дети) морфологический анализатор должен определить, что эта форма образована от основы бала (ребенок), к которой добавлен аффикс –лар. При этом

основа бала, утратив последнюю букву а, првращается в ad hoc основу бал, которая обуславливается по закону сингармонизма присоединение показателя –дар. А в случае словофрмы китебим (моя книга) морфологический анализатор должен определить, что основой является слово китеп, где после добавления окончания первого лица -ым последняя буква п, смягчаясь, переходит на букву б.

Решение проблемы

Существует два возможных подхода к решению задачи морфологического анализ словоформы, а именно: «справа налево» и «слева направо». При первом подходе делается попытка вычленить конечную часть словоформы, похожую на комплекс аффиксов, и затем проверить наличие в словаре оставшейся начальной части. При втором подходе делается попытка найти в словаре некоторую начальную часть цепочки, а затем проверить, что оставшаяся правая часть образует возможный для данной основы комплекс аффиксов. При обоих подходах в случае неудачи поиска, приходится повторять другое разбиение словоформы. Как правило, определение информации из словаря является наиболее долгой операцией при морфологическом анализе. Более точно, долгой операцией является происходящее при этом обращение к дисковой памяти. Поэтому и в первом, и втором подходах актуальной является задача сокращения числа обращений к дисковой памяти в рамках полного цикла анализа словоформы.

Поскольку средняя длина слова в научно-техническом тексте составляет примерно 7-10 букв, подход «слева направо» может потребовать до 10 обращений к словарю. Как правило, подход «справа налево» требует гораздо меньшего числа обращений к словарю, что является аргументом в пользу выбора именно этого подхода при реализации морфологического анализатора. Однако по сравнению с подходом «справа налево» подход «слева направо» имеет ряд преимуществ как по простоте реализации, так и по заложенным в нем возможностям. Таким образом, задача максимального сокращения числа обращений к диску является вдвойне актуальной, что ее решение позволит эффективно реализовать обладающий рядом преимуществ подход «слева направо» к задаче морфологического анализа.

Построение морфологической базы данных

Базы данных — это множество данных, удовлетворящих потребностям пользователя. Эти данные сортируются и хранятся в виде таблицы, которым управляет система управления базами данных.

В настоящее время существует множество систем для работы с базами данных: SQL, MySQL, Oracle, Access. Работать с большим объемом данных всегда трудно и разные системы имеют свои преимущества. В нашем случае для обработки естественного языка и проверки предложенного алгоритма была создана в среде Embarcadero RAD Studio тестирующая

программа. Здесь мы использовали среду Access для работы с базами данных.

У нас имеется три таблицы (основы, аффиксы и части речи) и связи между этими таблицами (рис.2):

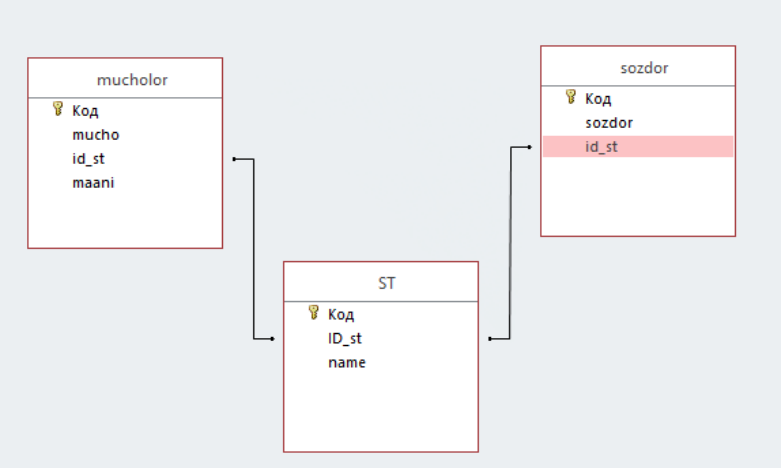


Рис. 2 Схема данных

Как видно из схемы данных, мы будем работать с тремя отдельными таблицами, которые связаны между собой ключевым полем it_st.

Оптимизация алгоритма морфологического анализа

Для оптимизации работы системы или сокращения числа обращений к диску мы использовали алгоритм следующего вида:

1. Выделяем из базы данных слова, которые начальные две буквы совпадают с начальными двумя буквами входного анализируемого слова по методу «слева направо».
2. Создаем виртуальный массив для сохранения данных.
3. Находим основу слова используя метод «справа налево».
4. Чтобы определить грамматическую категорию аффиксов будем использовать таблицу «mucholor»из базы данных. Итерация зависит от длины окончания.
5. В качесиве результата получим текстовую форму фрейм-модели, который представлен в рисунке 1.

Ниже приведем фрагмент кода программы, где мы создаем массив из отсортированных слов.

```
slovo:=copy(slovo,1,2);
text1:=slovo+'%';
text2:=quotedstr(text1);
with mucho.ADOQuery1 do
begin
close;
sql.Clear;
sql.Add('select * from sozdor where
sozdor like'+text2);
open;
end;
d:=adoquery1.RecordCount;
for i:=1 to d do
begin
```

```
  m[i]:=trim(adoquery1.Fields[1].Tex        adoquery1.Next;
t);                                         end.
```

Тестирование программы

Конечно, создать идеальный автоматизированный морфологический анализатор естественного языка невозможно. Поэтому мы провели тестирование программы. Для анализа ввели слово «ааламда». В данном случае из базы данных отсортировали более близкие 11 слов (рис.3).

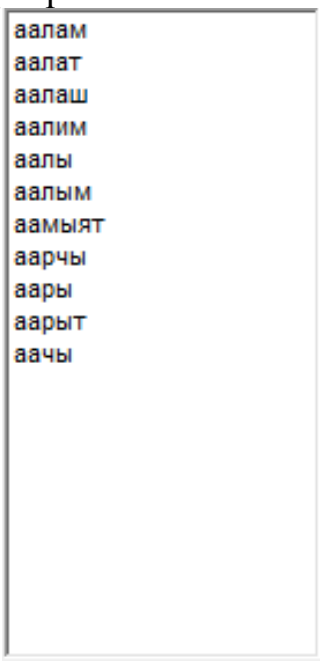


Рис. 3. Массив слов

Следующим этапом было нахождение основы слова. После выполнения модуля получили следующий результат:

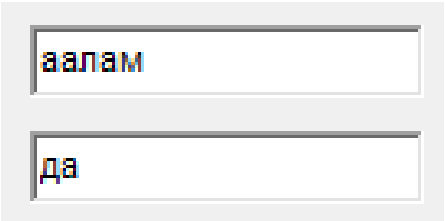


Рис. 4. Основа и окончание слова

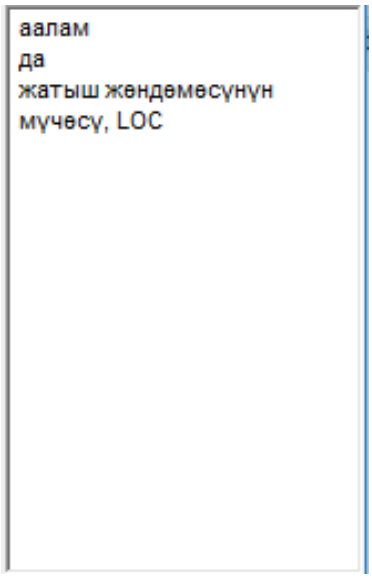


Рис. 5. Результат морфологического анализа

В данном случае система нашла основу и грамматическую категорию аффикса слова быстро, так как после добавления аффикса к основе слова не было никаких изменений.

Рассмотрим следующий пример, анализируем слово «китебим». В данном случае система находит видоизмененную основу китеб.. Для таких случаев мы создали специальные правила.

Например:

if (slovo=m[i]) or (slovo=(copy(m[i],1,length(m[i])-1))+'б')

or (slovo=(copy(m[i],1,length(m[i])-1))+'г') then writeln(m[i]);

После выполнения условия мы получим следующий результат (рис.6.):

Рис. 6. Нахождение основы и окончания слова

Выводы

В данном статье мы обсуждали проблему создания автоматизированного морфологического анализатора и об оптимизации алгоритма.

При этом:

1) была создана фрейм-модель формирования грамматических (текстовых) форм слов. Здесь для точного определения категории аффиксов была создана таблица аффиксов с указанием частей речи.

2) при создании анализатора были предложены алгоритмы сокращения числа обращений к памяти компьютера.

3) была создана тестовая программа и показаны принципы работы

Таким образом, создать идеальную систему невозможно. Ибо язык всегда имеет правила типа ad hoc. В следующих этапах путем расширения базы данных и алгоритма морфологического анализа расчитываем получить добротные результаты анализа.

ЛИТЕРАТУРА:

1. Т. Садыков, Г.Э. Жумалиева, М.Ж. Түмөнбаева, Ю. Шаршембиева Кыргыз тилинин компьютердик лигвистикасынын негиздери. Бишкек, 2015 ж.

2. Kochkonbaeva B.O., Aldosova A. Automatic processing of text in natural language. Бюллетень науки и практики 2018, Т4. №7, с.216-221.

3. Т. С. Садыков, Б.Ш. Шаршембаев, Б.О. Кочконбаева Морфологические разметки для национального корпуса. Вестник Кыргызско-Славянского Университета, КРСУ, 2018, №1. с. 91-95.

4. Segalovich I. «A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine» MLMFA, 2003. P. 273–280.

5. А. Акунова, Б. Чокошева, Г. Эшимбекова. Азыркы кыргыз тили морфология, Бишкек, 2009

6. Орузбаева Б., Турсунов А., Садыков Т. ж.б.Азыркы кыргыз адабий тили.Бишкек, 2009

РАЗРЕШЕНИЕ МНОГОЗНАЧНОСТИ ГЛАГОЛЬНЫХ МОРФЕМ В ТАТАРСКОМ ЯЗЫКЕ (НА ПРИМЕРЕ ГАн)

Б.Э. Хакимов [a,b], И.И. Фатхуллина [a],[a]Казанский федеральный университет,
[b] Академия наук Республики Татарстан, Казань, Россия,
khakeem@yandex.ru

Проблема разрешения многозначности является ключевой в современной компьютерной лингвистике. Многозначность языковых единиц проявляется на лексическом, лексико-грамматическом, функциональном, морфологическом, синтаксическом, деривационном, морфемном и других уровнях. На сегодняшний день набор основных методов устранения многозначности включает контекстные методы на основе правил, методы машинного обучения, основанные на вероятностных моделях и гибридные методы. Изучение многозначности в тюркских языках, в том числе в татарском языке, до недавнего времени было довольно фрагментарным.Таким образом, разработка формальных моделей и методов устранения многозначности весьма актуальна для татарского языка.

Предыдущие исследования показали, что наряду с использованием статистических и вероятностных методов представляет интерес также определение того, какие типы морфологической многозначности в татарском языке могут быть разрешены с использованием метода, основанного на правилах. В этом отношении поиск соответствующего многозначного контекста играет важную роль.

В данной статье предпринята попытка проанализировать многозначные контексты для многозначных морфем глагола в татарском языке на примере морфемы -ГАн как на уровне словоформ, так и предложений. Предлагается вариант классификации многозначных контекстов, а особенности контекстов анализируются на основе данных корпуса.

Очевидно, что границы многозначного контекста зависят от конкретной многозначной языковой единицы. Когда многозначная единица представлена морфемой, мы можем предположить, что минимальный многозначный контекст в некоторых обстоятельствах может быть найден в границах словоформы. Это позволяет разрабатывать контекстные правила устранения многозначности, не касаясь уровня предложения, и даже снимать определенные виды морфологической многозначности на этапе автоматического морфологического анализа путем корректировки морфологической модели. Следует отметить, что морфологические модели программ-анализаторов, как правило, являются

избыточными. Следовательно, обнаруженные типы многозначности также являются избыточными. С другой стороны, подход, основанный на уровне словоформ, лучше всего подходит для агглютинативных языков, где морфемы в большинстве случаев могут передавать одновременно только одно значение. Исходя из особенностей татарской и, в целом, тюркской морфологии, мы можем легко сделать вывод, что в словоформе по большей части правильный контекст многозначен. Это связано с тем, что агглютинативная форма слова развертывается слева направо, и в том же направлении конкретизируются ее грамматические значения. Поэтому морфемы, стоящие слева от многозначной морфемы, выражают более общие, универсальные категории. Но правильный контекст представлен морфемами с более конкретными значениями, и он может служить в качестве однозначного контекста.

Наш вывод заключается в том, что возможность нахождения многозначного контекста в словоформе ограничена следующими факторами:

• многозначная морфема не должна быть последней в цепочке аффиксов;

• морфема в правильном контексте не должна относиться к «общему»; категории;

• морфемы в альтернативных многозначных контекстах должны относиться к неперекрывающимся парадигмам или ветвям парадигм.

Ключевые слова: *морфологическая многозначность; тюркские языки; татарский язык; многозначный контекст.*

DISAMBIGUATION OF VERB MORPHES IN THE TATAR LANGUAGE (ON THE EXAMPLE OF "-GAn" MORPHEME)

Khakimov B.E., Fathullina I.I.[a],[a]Kazan Federal University,
[b] Academy of Sciences of the Republic of Tatarstan, Kazan, Russia,
khakeem@yandex.ru

The problem of disambiguation is a key in modern computational linguistics. The ambiguity of a linguistic unit appears at the lexical, lexical-grammatical, functional, morphological, syntactic, derivational, morpheme and other levels.

To date, the set of the basic methods of disambiguation includes rule-based context methods, machine learning methods based on the probabilistic models, and hybrid methods. The study of disambiguation for the Turkic languages, including Tatar, until recently were rather fragmentary. Thus, the development of formal models and methods of disambiguation is highly relevant for the Tatar language. Earlier studies showed that, along with the use of statistical and

probabilistic methods, it is also of interest to identify which types of morphological ambiguity in the Tatar language are available for applying of the rule-based method. In this aspect, finding the relevant disambiguating context plays an important role. This article attempts to analyze the disambiguating contexts for the ambiguous verb morphemes in the Tatar language on the example of the morpheme GAn both at the level of word forms, and sentences. A version of the classification of disambiguating contexts is proposed, and features of the contexts are analyzed on the corpus data.

It is obvious that the boundaries of the disambiguating context depend on the particular ambiguous linguistic unit. When the ambiguous unit is presented by a morpheme, we can assume that the minimum disambiguating context in some circumstances might be found within the boundaries of the word form. It allows developing context-based disambiguation rules without reaching the level of sentence, and even preventing certain types of morphological ambiguity at the stage of automatic morphological analysis through the adjustment of the morphological model.

It should be noted that the morphological models of the analyzer programs, are redundant as a rule. Therefore, the detected types of ambiguity are also redundant. On the other hand, the approach based on the level of word forms works best for agglutinative languages where morphemes in most cases can carry one meaning at a time. Based on the characteristics of the Tatar and, on the whole, Turkic morphology, we can easily conclude that within the word form mostly the right context is disambiguating. This is because the agglutinative word form is deployed from left to right, and in the same direction its grammatical meanings are concretized. So morphemes which stand to the left from the ambiguous morpheme express more general, universal categories. But the right context is presented by morphemes with more specific meanings and it has the potential to serve as a disambiguating context. Our conclusion is that the possibility of finding the disambiguating context within a word form is limited to the following factors:

• the ambiguous morpheme should not be the last in the affix chain;

• the morpheme in the right context should not refer to the "general" categories;

• the morphemes in the alternate disambiguating contexts must refer to non-overlapping paradigms or branches of the paradigms.

Key words: *Morphological disambiguation; Turkic languages; Tatar language; disambiguating context*

---

1. Введение

Проблема разрешения многозначности является одной из ключевых в современной компьютерной лингвистике. Многозначность (неоднозначность) языковой единицы проявляется на лексическом,

лексико-грамматическом, функциональном, морфологическом, синтаксическом, словообразовательном, морфемном и других уровнях.

К настоящему времени сформирована основная парадигма методов снятия многозначности, которая включает контекстные методы на правилах, методы машинного обучения с использованием вероятностных моделей, а также гибридные методы. Исследования проблемы снятия многозначности в системах автоматической обработки текстов для тюркских языков, в том числе и для татарского, до последнего времени проводились довольно фрагментарно. Таким образом, весьма актуальным является построение формальных моделей и методов разрешения многозначности для татарского языка.

Проведенные ранее исследования показали, что наряду с применением статистико-вероятностных методов значительный интерес представляет выявление в татарском языке типов морфологической многозначности/ омонимии, перспективных с точки зрения применимости метода контекстных правил разрешения, а также разработка контекстных правил разрешения многозначности для выявленных типов на основе выделенных разрешающих контекстов. И в этом аспекте важную роль играет определение соответствующего разрешающего контекста, который представляет собой окружение языковой единицы, снимающее ее многозначность.

В настоящей статье предпринята попытка анализа разрешающих контекстов для глагольных омоформ татарского языка на примере морфемы ГАн, как на уровне словоформы, так и на уровне предложения. Предлагается вариант классификации разрешающих контекстов, особенности разрешающих контекстов проанализированы с привлечением корпусных данных.

2. Границы разрешающего контекста для морфем татарского языка

Очевидно, что границы разрешающего контекста зависят от многозначной (омонимичной) языковой единицы. В случае, когда многозначной является морфема, можно было бы предположить, что минимальный разрешающий контекст может находиться в границах словоформы. Это позволило бы в определенных случаях строить контекстные правила разрешения многозначности без выхода на уровень предложения, что, в свою очередь, дает возможность разрешать (точнее, предупреждать) некоторые типы грамматической омонимии на этапе автоматического морфологического анализа, путем уточнения морфологической модели.

Необходимо отметить, что морфологические модели программ-анализаторов, как правило, являются избыточными, следствием этого является избыточность выделяемых типов омонимии.

С другой стороны, подобный подход на уровне словоформы наиболее эффективно работает для агглютинативных языков, в которых морфемы в большинстве являются категориально однозначными (т.е. не выражают одновременно более одного грамматического значения).

Исходя из особенностей татарской и тюркской морфологии в целом, мы легко можем прийти к выводу о том, что на уровне словоформы разрешающим является правый контекст. Это объясняется тем, что агглютинативная словоформа развертывается слева направо, в этом же направлении идет и конкретизация грамматических значений. Следовательно, слева от омонимичной морфемы, для которой требуется разрешение, будут находиться морфемы, выражающие более общие, универсальные категории. Справа же располагаются морфемы, которые выражают более частные значения и потенциально могут служить разрешающим контекстом.

Таким образом, возможности применения разрешающего контекста на уровне словоформы ограничиваются следующими факторами:

- омонимичная морфема не должна быть последней в аффиксальной цепочке;
- морфемы в правом контексте не должны относиться к «общим» категориям;
- морфемы в альтернативных разрешающих контекстах должны относиться к непересекающимся парадигмам, либо непересекающимся ветвям парадигм.

К «общим» категориям в данном случае мы относим такие морфемы, как вопросительные мЫ и мЫни, а также ДЫр. За исключением случаев, когда данные морфемы прямо или косвенно сигнализируют о синтаксической функции, на основе которой многозначность может быть разрешена. Например, вопросительные морфемы мЫ и мЫни, как правило, присоединяются к сказуемому вопросительного предложения общего типа. Так, тип V,PCP_PS(ГАн)|V,PST_INDF(ГАн) представляет омонимию причастия (сыйфат фигыль) и изъявительного наклонения (хикәя фигыль), при этом причастие обычно не используется в функции сказуемого.

3. Выделение разрешающих контекстов на уровне словоформы

Примеры применения разрешающих контекстов на уровне словоформы приведены в таблице 1. В столбце «Количество подтипов» указано общее количество уникальных подтипов из двух омоформ, выделяемых морфоанализатором и имеющих в своем составе омонимичную морфему. В столбце «Количество разрешающих контекстов» приведено количество подтипов из общего числа, разрешаемых на уровне словоформы.

Таблица 1.

ПРИМЕРЫ ПРИМЕНЕНИЯ РАЗРЕШАЮЩИХ КОНТЕКСТОВ НА УРОВНЕ СЛОВОФОРМЫ

Тип омонимии	Кол-во подтипов	Разрешающих контекстов
V,PCP_PS(ГАн)|V,PST_INDF(ГАн)	224	186
N,Nom,PL(ЛАр)|V, FUT_INDF(ЫР)	11	8
V,2SG(сЫң),PRES(Й)| V,Nom,OBL(ЙсЫ),POSS_2SG(Ың)	27	15
V,FUT_INDF_NEG(мАс)|V,PCP_FUT(мАс)	30	22

Приведенные в таблице типы омонимии представляют омонимию двух аффиксальных цепочек, одна из которых входит в именную, а другая в глагольную парадигму, например, тип N,Nom,PL(ЛАр)|V, FUT_INDF(ЫР): *атлар* ‘лошади’ (лошадь+Nom+PL) и ‘шагнет’ (шагнуть+FUT_INDF). Остальные три типа из таблицы демонстрируют омонимию аффиксальных цепочек, относящихся к разным ветвям глагольной парадигмы. Из них типы V,PCP_PS(ГАн)|V,PST_INDF(ГАн) (в прошедшем времени) и V,FUT_INDF_NEG(мАс)|V,PCP_FUT(мАс) (в будущем времени) представляют известную неоднозначность тюркских глагольных форм причастия и изъявительного наклонения 3 лица: *барган кеше* ‘ходивший человек’ и *кеше барган* ‘человек ходил’; *үлмәс җаны* ‘(его) душа, которая не умрет’ и *җаны үлмәс* ‘(его) душа не умрет’.

4. Разрешающие контексты для морфемы ГАн

Рассмотрим разрешающие контексты на примере типа V,PCP_PS(ГАн)|V,PST_INDF(ГАн).

Тип V,PCP_PS(ГАн)|V,PST_INDF(ГАн) представлен значительным количеством подтипов (в проанализированной корпусной выборке зафиксированы 224 омонимичные аффиксальные цепочки с морфемой ГАн в составе). Морфема ГАн является омонимичной (многозначной) и представлена двумя функциями: прошедшее время изъявительного наклонения (PST_INDF) и причастие (PCP_PS). Был проанализирован состав аффиксальных цепочек, с точки зрения позиции данной морфемы (см. таблицу 2).

Таблица 2.

АФФИКСАЛЬНЫЕ ЦЕПОЧКИ С МОРФЕМОЙ ГАн

Позиция в цепочке	Кол-во подтипов	Наличие разрешающего контекста на уровне словоформы

Единственный элемент аффиксальной цепочки	1	-
Последний элемент цепочки, есть левый контекст	36	-
Есть правый контекст, именная парадигма	91	PCP_PS
Есть правый контекст, глагольная парадигма	96	PST_INDF

Определены дифференцирующие признаки, релевантные для снятия неоднозначности данного типа:

- Занимает ли омонимичная морфема последнюю позицию в аффиксальной цепочке?
- К какой парадигме (именной или глагольной) относится часть аффиксальной цепочки в правом для омонимичной морфемы контексте? (см. рис. 1)

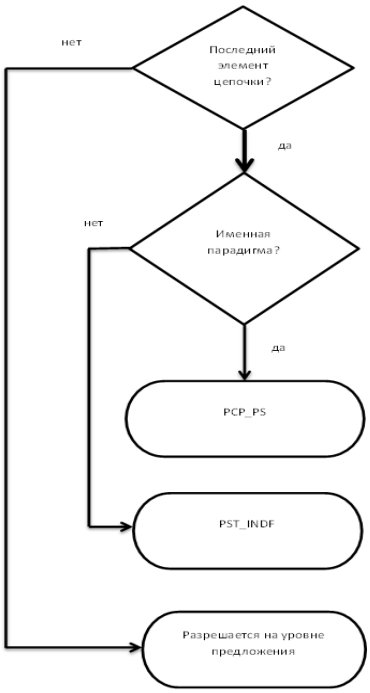


Рис. 1. Обобщенная структура правила разрешения морфемы ГАн на уровне словоформы

Использование аффиксов именной и глагольной парадигмы в качестве дифференцирующих признаков обусловлено особенностями глагольных форм причастия и изъявительного наклонения. Согласно татарской грамматике, причастие совмещает признаки глагола и прилагательного. Прилагательные в татарском языке являются неизменяемой частью речи, однако при субстантивации принимают аффиксы именной парадигмы.

Подобное явление не наблюдается для собственно «глагольной» формы изъявительного наклонения. Последняя, в свою очередь, принимает личные глагольные аффиксы, что невозможно для причастия. Кроме того, к именной парадигме мы в данном случае условно относим вопросительный аффикс мЫ, в том случае, когда он является единственным в правом контексте и отсутствуют другие аффиксы, однозначно имеющие принадлежность к именной либо глагольной парадигме. Несмотря на то, что аффикс мЫ является универсальным и не привязан к какой-либо из парадигм, в данном случае он может служить разрешающим контекстом, так как присоединяется к словоформе, выполняющей в предложении синтаксическую функцию сказуемого. С другой стороны, причастие PCP_PS в несубстантивированной форме, т.е. без иных аффиксов в правой части выполняет функцию определения и практически не может быть сказуемым.

5. Частотность разрешающих контекстов для морфемы ГАн

На выборке из корпуса «Туган тел» объемом 22 млн. словоупотреблений нами был произведен подсчет частотности аффиксальных цепочек в зависимости от разрешающего контекста (см. Таблицу 3).

Таблица 3.

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ АФФИКСАЛЬНЫХ ЦЕПОЧЕК С МОРФЕМОЙ ГАН

Позиция в цепочке	Словоупотреблений в корпусной выборке	Доля в %
Единственный элемент аффиксальной цепочки	449206	75,2
Последний элемент цепочки, есть левый контекст	68128	11,4
PCP_PS (правый контекст)	67090	11,2
PST_INDF (правый контекст)	12994	2,2
Всего	597418	100

Как видно из таблицы, в корпусе доля примеров с разрешающими контекстами на уровне словоформы составляет 13,4 %. В то же время, в общей сложности 86,6 % примеров необходимо разрешать на уровне предложения.

Нами также был произведен анализ разрешающих контекстов для исследуемого типа многозначности на уровне предложения. В частности, для вышеуказанного типа V, PCP_PS(ГАн)|V,PST_INDF(ГАн) была проанализирована выборка из подкорпуса со снятой вручную омонимией, были определены наболее частотные разрешающие контексты для каждого

из двух альтернативных вариантов омонимичной морфемы. Так, в 66 % примеров, в которых морфема ГАн используется в значении причастия, в правом контексте непосредственно соседствует существительное. В то же время, в 88,5 % примеров, в которых морфема ГАн используется в значении изъявительного наклонения, в ближайшем правом контексте присутствуют знаки препинания (точка, запятая, двоеточие, вопросительный знак, восклицательный знак и др.).

6. Заключение

Подводя итог, следует отметить, что рассмотренная в статье проблема требует дальнейших исследований с привлечением корпусных данных. На данный момент получены следующие результаты:

- произведен анализ разрешающих контекстов для некоторых татарских омоформ и омонимичных аффиксов;
- рассмотрены типы разрешающих контекстов в зависимости от омонимичной языковой единицы и состава аффиксальной цепочки, проанализированы особенности некоторых разрешающих контекстов на основе корпусных данных;
- определены факторы, ограничивающие возможности применения разрешающего контекста на уровне словоформы.

ЛИТЕРАТУРА

1. Гатауллин Р.Р., Гильмуллин Р.А., Хакимов Б.Э. (2015) Методы разрешения морфологической многозначности в татарском языке. *Сохранение и развитие языков и культур в поликультурном и поликонфессиональном обществе: мировой опыт и современные технологии: материалы Международной научно-практической конференции (Казань, 14-16 октября 2015 г.)*. — С. 56-57.
2. Хакимов Б.Э. Гильмуллин Р.А., Гатауллин Р.Р. (2014) Разрешение грамматической многозначности в корпусе татарского языка. Ученые записки Казанского университета, кн. 5. — С. 236-244.
3. Невзорова О.А., Зинькина Н.В., Пяткин Н.В. (2006) Метод контекстного разрешения функциональной омонимии: анализ применимости. *Труды международной конференции Диалог'2006.*- С.399–402.
4. Татарская грамматика: Морфология. Т. 2. (1993). Казань: Татарское книжное издательство.
5. Татар грамматикасы: Морфология. Т. 2. (2002). Казань: Фикер, Москва: Инсан.
6. Татарский национальный корпус «Туган тел» http://tugantel.tatar.
7. Gataullin, R., Gilmullin R. (2015) Grammatical Disambiguation in Tatar Language Corpus. In Proceedings of the International Conference on Turkic Language Processing TURKLANG-2015 (pp. 290-296). Kazan: TAS Press.

8. Gataullin, R, Khakimov, B, Suleymanov, D, Gilmullin, R. (2017) Context-Based Rules for Grammatical Disambiguation in the Tatar Language In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol.10449 LNAI (pp..529-537).
9. Khakimov, B., Gilmullin, R., Gataullin, R. (2016) Grammatical Disambiguation in the Tatar National Corpus. In *EPiC Series in Language and Linguistic*s. *Volume 1, 2016, CILC2016. 8th International Conference on Corpus Linguistics*. (pp. 229 – 235).

SPELL CHECKING ANALYSIS OF UZBEK TEXTS USING DJARO WINKLER ALGORITHM

U.Tuliyev[1]., N.Abdurakhmonova[2],[1]National university of Uzbekistan named after Mirzo Ulugbek, mirzobek.tuliyev@gmail.com
[2]Tashkent state university of Uzbek language and literature named after Alisher Navoi, abdurahmonova.1987@mail.ru

The article is devoted to tackle the problem of spell checker program for Uzbek text using Djaro Winkler algorithm.

Key words: *Spell checking analysis, Djaro Winkler algorithm.*

АНАЛИЗ ОРФОГРАФИЧЕСКИЙ ПРОГРАММ УЗБЕКСКОГО ТЕКСТА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА ДЖАРО ВИНКЛЕРА

Тулиев Улугбек[1]., Абдурахмонова Нилуфар[2], [1] Национальный университет Узбекистана им. Мирзо Улугбека, mirzobek.tuliyev@gmail.com
[2] Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, abdurahmonova.1987@mail.ru

В статье описывается программа проверки орфографии узбекского текста с использованием алгоритма Джаро Винклера.

Ключевые слова*: проверка орфографии, алгоритм Джаро Винклера.*

The tasks of text processing immediately arose almost after the advent of computer technology. However, despite half century of research in the field of artificial intelligence, a huge leap in the development of IT and its related disciplines, a satisfactory solution to most practical problems of text processing is not available yet. It is clear that there are some incredible results for some languages such as English, German, French, Russian and Turkish.

There are many issues related on text mining such as Authorship attribution, identifying domain of given text, translating, summarization and so on. That is why analyzing text is divided different types.

Distance of Djaro

Distance of Djaro between two given string is calculated as:

$$d_j = \begin{cases} 0 & when\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & in\ other\ cases \end{cases}$$

where: $|s_i|$ — length of string s_i;

m – number of similar characters;

t – half number of transpositions.

Two characters taken from strings s_1 and s_2 are said similar to each other if only if they are the same and no farther than $\left[\frac{\max(|s_1|,|s_1|)}{2}\right] - 1$.

Each character of the string c1 is compared with all the corresponding symbols in c2. The number of similar (but in different positions) characters, which is divided to 2, is the number of transpositions. For example, in comparing a word ODAMLAR with a word OMADLAR, al characters from two given words a similar and m=7, but M and D are different positions. That is why t=1. Nevertheless, if we compare ODAM with OMAD, only characters O and A are similar. Because positions of D and M in two words farther than 1. That's why they can not be said similar characters.

Distance of Djaro-Winkler

Distance of Djaro-Winkler uses a scaling coefficient p, it gives more clear rate for strings, which are similar to each other from the beginning till length l, which is called prefix. Distance of Djaro-Winkler for given two strings is:

$$d_w = d_j + \left(lp\left(1 - d_j\right)\right)$$

where: d_j – is Djaro distance for strings s_1 and s_2

l – is the length of common prefix from the beginning of strings till maximum 4-th character

p – is scaling coefficient. It shouldn't be bigger than 0,25. In the opposite cases distance can be bigger than 1. But it should be between 0 and 1. If distance is equal to 0, it means that two strings are totally dissimilar to each other and if it is equal to 1, it means that two strings are exactly the same. Standard value of scaling coefficient is 0,1 in Winkler's works. We also accept this value.

Let compare two words TAROQ and TAYOQ. A table is created for this:

	T	A	R	O	Q
T	1	0	0	0	0
A	0	1	0	0	0
Y	0	0	0	0	0
O	0	0	0	1	0
Q	0	0	0	0	1

For similar characters we put 1 and in other case we put 0. Every character of the first word is compared with all characters of the second word to complete the table. The number of 1's gives value of m=4. There is no transpositions letters so t=0.

$$d_j = \frac{1}{3}\left(\frac{4}{5} + \frac{4}{5} + \frac{4}{4}\right) = 0.8(6)$$

For these two words l=2 and distance of Djaro-Winkler:

$$d_w = 0.8(6) + \left(2 \cdot 0.1 \cdot \left(1 - 0.8(6)\right)\right) = 0.89(3)$$

It is clear that if we compare two words which have not the same prefix distance of Djaro and distance of Djaro Winkler will be the same.

	T	A	R	A	M	O	Q
Q	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0
R	0	0	1	0	0	0	0
A	0	0	0	1	0	0	0
M	0	0	0	0	1	0	0
O	0	0	0	0	0	1	0
Q	0	0	0	0	0	0	1

For the words above TARAMOQ and QARAMOQ. m=6, t=0, $|s_1|=|s_2|=7$:

$$d_j = \frac{1}{3}\left(\frac{6}{7} + \frac{6}{7} + \frac{6}{6}\right) = 0.9047$$

l=0, p=0.1:

$$d_w = 0.9047 + \left(2 \cdot 0.1 \cdot (1 - 0.9047)\right) = 0.92376$$

There is some result of different test of program that is written in python programming language.

	DISTANCE OF DJARO	DISTANCE OF DJARO-WINKLER
BAHOR	0.8666666666666667	0.9066666666666667
NAHOR	0.7333333333333334	0.7333333333333334
XABAR	0.7	0.7
KITOB	0.8666666666666667	0.9066666666666667
KILOB	0.7333333333333334	0.7866666666666667

REFERENCE:

1. Choudhury, R., K. Kashyap, and N. Deb, 2016. A survey on the different approaches of context sensitive spell-checking. *International Journal of Engineering Science and Computing*, 6(6):6872–6873.
2. Fierman, W., 1992. *Language planning and national development: the Uzbek experience*. Berlin, Germany; New York, USA: Mouton de Gruyter.
3. Gupta, N. and P. Mathur, 2012. Spell checking techniques in NLP: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(12):217–221.

4. Li, X., J. Tracey, S. Grimes, and S. Strassel, 2016. Uzbek-English and Turkish-English Morpheme Alignment Corpora. In *Proceedings of LREC 2016*. pages 2925–2930.
5. Li, X., J. Tracey, S. Grimes, and S. Strassel, 2016. Uzbek-English and Turkish-English Morpheme Alignment Corpora. In Proceedings of LREC 2016. Pages 2925–2930.
6. Pirinen, T. and K. Linden, 2014. State-of-the-art in weighted finite-state spell-checking. In CICLing 2014: Computational Linguistics and Intelligent Text Processing, volume 8404 of LNCS. Berlin, Heidelberg: Springer.
7. Sayfullaev, A., 2016. Problems of rendering prepositions in translation (on the material of English and Uzbek languages). In *Scientific enquiry in the contemporary world: theoretical basics and innovative approach*. pages 91–96.
8. Singh, S.P., A. Kumar, L. Singh, M. Bhargava, K. Goyal, and B. Sharma, 2016. Frequency based spell checking and rule based grammar checking. In Proc. Of International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE.

МАТНЛАРГА АВТО-ЛИНГВИСТИК ИШЛОВ БЕРИШ ТИЗИМЛАРИ

М. Абжалова, Навоий давлат кончилик институти, Навоий, Ўзбекистон, manzura_ok@mail.ru

Мазкур мақолада матнларни автоматик таҳрир ва таҳлил қилувчи тизимлар борасида сўз юритилди. Шунингдек, мазкур масалада ўзбек адабий тили доирасида олиб борилаётган изланишлар баён этилди.

Таянч сўзлар: *дастурий таъминот, спелл-чекер, стемминг, лемматизация, токенизация, парсер.*

АВТОЛИНГВИСТИЧЕСКИЕ СИСТЕМЫ ОБРАБОТКИ ТЕКСТОВ

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

В статье обсуждается вопрос об автоматическом редактировании и автоматическом анализе систем текстов. А также говорилось о проводимых исследованиях по развитию узбекского литературного языка.

Ключевые слова: *программное обеспечение, спелл-чекер, стемминг, лемматизация, токенизация, парсер.*

AUTOLINGUISTIC TEXT PROCESSING SYSTEMS

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

In this article discussed about automatic editing and automatic analysis systems of texts. And also, it was spoken about conducted researches on development of the Uzbek literary language.

Key words: *software, spell-checker, stemming, lemmatization, tokenization, parser.*

Матнларни лингвистик таҳрир қилиш жараёнига азал-азалдан асосий филологик масала сифатида қаралган бўлиб, ушбу вазифа йиллар давомида инсон томонидан амалга ошириб келинмоқда.

Матнларга ишлов беришнинг автоматик таҳрир йўналиши XX аср ўрталарига келиб ривожланди. У матн муҳаррир дастурлари билан биргаликда янгича имкониятлар билан ривожланмоқда. Оддий

муҳаррирлардан фарқи шундаки, унда таҳрир автоматик тарзда қисқа вақт ичида катта ҳажмли матнлар тез текширилиб, хатоларни самарали тўғрилаш каби имкониятлари бўлади.

Лингвистик таҳрирлаш (корректура) – турли кўринишдаги (илмий, бадиий, публицистик ва расмий услублардаги) матнларнинг орфографик, грамматик, стилистик ҳамда мантиқий қурилишини текшириш демакдир.

Бугунги кунда матнларни таҳрир қилувчи ёхуд унинг таҳлилини ҳам амалга оширувчи автоматик лингво-тизимлар яратилган бўлиб, уларнинг ишлаш принциплари янада такомиллаштирилмоқда. Қуйида шундай тизимлар борасида сўз юритилди.

Имлони текшириш тизими (спелл-чекер ингл. spell checker) – компьютер дастури бўлиб, киритилган матннинг орфографик текширувини — таҳрирни амалга оширади. Аниқланган имло хатолари махсус тарзда белгиланади, яъни хато ёзилган лексеманинг тагига чизилади. Кўп ҳолларда матн терувчига имловий хатоларга ишора қилишдан ташқари дастур махсус эслатмаси сифатида сўзнинг тўғри ёзилиш вариантларини ҳам таклиф қилади. Шунингдек, матнга қандай тузатиш киритиш мумкинлигига изоҳлар ҳам берилади.

Биринчи матн териш текшируви тизимлари 1970-йилларнинг охирларида фойдаланила бошланган. Жоржтаун университетининг олти нафар тилшуносидан иборат гуруҳи IBM компанияси учун биринчи тизимни ишлаб чиқди. CP/M ва TRS-80 шахсий компьютерларига 1980 йилда ушбу тизим киритилган, 1981 йилда IBM PCда тизимнинг биринчи пакетлари пайдо бўлди.

Орфографик текширув (Speller) – матнни орфографик жиҳатдан тўлиқ текширувчи модул: унинг қулайлиги шундаки, келтирилган кўрсатмалар орқали матнни киритувчи дастурнинг лингвистик таъминотига янги сўзлар, сўзшаклларни киритиб автоматик луғатни шу заҳоти бойитиши мумкин.

Лемматизация (lemmatization) – бу сўзшаклларини унинг луғатдаги оддий шакли – леммага келтириш жараёни[7].

Лемматизация сўзларнинг лексиконидан фойдаланиб, уларнинг морфологик таҳririни амалга оширувчи аниқ жараён бўлиб, лемматизация жараёнида фақат флекцияга учраган аффикслар учирилади ва лемма деб аталмиш таъминотдаги сўзнинг асосий ёки луғат шаклига қайтарилади.

Стемминг (stemming) – бу киритилган сўзнинг асоси (ўзак)ини топиш жараёни. Бунда топилган сўз асоси морфологияда қабул қилинган сўз ўзагига мос келиши талаб қилинмайди. Боиси дастур таъминотига «тайёр қолипли сўзлар» асос сифатида алоҳида категория қилиб киритилиш эҳтимоли юқори бўлади. Масалан, ҳуқуқ, тадқиқ каби лексемаларга эгалик қўшимчаси қўшилганда лексема сўнгидаги қ ундоши ўзгаришга учрамагани

[7] Атамаларга айнан изоҳ беришда https://ru.wikipedia.org/wiki/ сайтидаги маълумотлардан фойдаланилди.

боис, бундай истисноли лексемалар «тайёр қолипли сўзлар» категориясига ҳуқуқи, тадқиқи каби киритилади ва улар стеммингда асос деб қабул қилинади.

Стемминг, одатда, тахминий жараён дейилади. Сабаби, стеммингда сўзшакллари аффикслари охиридан бошлаб лингвистик таъминотга киритилган асос шаклга қадар кесиб келинади. Кўп ҳолларда бу ҳолат ўзини оқлаган.

Стемминг бир асосдан юзага келган сўзшаклларидаги белгилар билан ишлайди, лемматизация эса бир лемманинг флектив (аффикс қўшилиши натижасида ўзгаришга учраган шакл) шаклини эътиборга олади.

«Токенизация» инг. tokenizing сўзидан олинган бўлиб, информатиклар томонидан киритилган атама ҳисобланади ва тилшуносликда лексик таҳлил бирикмаси билан тушунтирилади. Токенизация – бу киритилаётган белгилар кетма-кетлигини муайян гуруҳларга, яъни лексемаларга ажратиб, таҳлил қилиш жараёни. Ушбу жараён чиқишда «токен» («сўзлар билан гуруҳланган ҳарфлар») деб аталмиш идентификация қилинган кетма-кетликни олиш мақсадида амалга оширилади.

Парсер (инг. parser; parse – таҳлил қилиш) ёхуд синтактик анализатор — дастур қисми ёки фақат синтактик таҳлилни амалга ошириш учун яратилган махсус дастур. Кириш маълумотларини (матн) муайян форматга келтириш орқали таҳлил қилади.

Ўзбек тилидаги матнларни автоматик таҳрир қилиш дастурини яратишда, албатта, тилшунослар томонидан яратиладиган лингвистик маълумотлар базаси муҳим аҳамиятга эга. Бу борада кўп босқичли ўзбек тилининг таҳрир ва таҳлил дастурининг лингвистик таъминоти устида иш олиб борилмоқда. Матнларни автоматик равишда таҳрир қилувчи дастурларнинг лингвистик модулларининг яратилиши мукаммал дастурларнинг ишлаб чиқилишига замин яратади, бу эса ўзбек адабий тилидаги матларнинг саводли ёзилиши ва дунё тиллари қаторидан ўрин олишига хизмат қилади.

АДАБИЁТЛАР:

1. Русский орфографический словарь: ок. 200 000 слов / РАН. Ин-т рус. яз. им. В. В. Виноградова / Под ред. В. В. Лопатина, О. Е. Ивановой. — Изд. 4-е, испр. и доп. — М.: АСТ-Пресс Книга, 2012. — С. 709. — (Фундаментал. словари рус. яз.). — ISBN 978-5-462-01272-3.

INTELLECTUALLY ANALYZING DOCUMENTS IN UZBEK LANGUAGE

U. Tuliyev, National University of Uzbekistan named after Mirzo Ulugbek Tashkent, Uzbekistan, u.tuliyev@mail.ru

This paper gives a brief description about intellectually analyzing documents in Uzbek language and some special properties of Uzbek language.

Key words: *corpus, structure, preprocessing.*

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТА ДОКУМЕНТОВ В УЗБЕКСКОМ ЯЗЫКЕ

У. Тулиев, Национальный университет Узбекистана имени Мирзо Улугбека Ташкент, Узбекистан, u.tuliyev@mail.ru

В статье приводится краткое описание интеллектуального анализа текста документов в узбекском языке, а также некоторых особенностей узбекского языка.

Ключевые слова: *корпус, структура, предобработка.*

Nowadays, intellectually analyzing data is one of the most actual research area of computer science and there are a lot of methods of this subject. But these methods are useful for only well structured data. If these data is given in a comfortable form to analyze. As for data given in text form is a weak structured data. Because it has not got any clear structure. Furthermore analyzing documents in uzbek language is more difficult as other tukish languages. Because uzbek language is very weak structured language. Suffixes and affixes make difficult this language to work. As one word can have more models adding suffixes to it.

However we can get more new knowledge analyzing documents. Because most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases.

There are several techniques to use text documents such as information extraction, categorization, clustering, visualization and summarization. So there are several tasks being done by human that we should teach to computer. These problems depends on each others solution. Computional Linguistics works on this kind of problems. We discuss some of them below.

Morphological analyzing. This task can be seen as a beginning problem of text mining domain. It requires to analyze individual words in text. It contains spellchecker, stem defining, tokenizing. But it does not care about semantic

meaning of We should collect uzbek language corpus to do these tasks. So we can find several form of given word from different documents in uzbek language. Because in uzbek language one word has a lot of tokens. If we have uzbek language corpus we can compare words in different documents and find the nearest words to the given word. There are several metrics to find a distance between two words. Damerau-Levenshtein metrics can be suggested for uzbek language.

Phonological analyzing. It learns phonetics and its rules. There are a lot of phonetical changings in words when suffix is added to it. It contains problems of translation text to speech and speech to text correctly. We can see there is more successful solution for English, Russian, Germany, Turkish and other languages. But there is no satisfactory solution for Uzbek language.

Syntactic analyzing. Syntax learns the structure of sentences, rules of relations between words and order of words in sentences. Furthermore it learns general properties of natural language. Especially this task can be more difficult in Uzbek language. Because words in Uzbek language can connect to each other in a lot of different forms.

Semantical analyzing. Semantical analyzing is connected with the meaning of words, collocations and sentences. There are some words that give different meaning in different subjects. For example a word «daraxt» is understood as a structure in computer science but it is a plant in biological subjects. If this word is translated to another foreign language this aspect should be considered.

Lexical analyzing. It describes the lexicon of a specific natural language. It can be its individual words, their grammatical and semantic properties and so methods of writing a dictionary.

Applications. We can teach computer to do our special tasks connected with language using this kind of analysis. But to do it we should use artificial intelligence methods, especially machine learning methods.

But because of text consists of different sentences and sentence consists of different words, furthermore probability of one word has several meanings and several words have the same meaning can be difficult problem for machine to analyze them automatically. This kind of data can be seen as a weak structured data. We should transfer them to well structured form such as digital matrix or graph before apply a method or its algorithm on it. In this process choosing a structure also is another problem and it can be determined by experiment. Different information can be extracted analyzing documents transferred to well structured form. For example, we can clarify which domain document belongs to. We need to collect thematical dictionary data base to do it. It is clear the that number of words in this dictionary can be large and it can increase number of variants should be compared. In this purpose, at first, up to hundred the most active words must be separated from every subject area documents. This kind of documents set up uzbek language corpus. Using web sites (we chose

www.ziyonet.uz) is usefull to do it. After creating a database we can use intellectual analyzing methods comparing documents with this database. It requires frequency dictionary. Frequency dictionary is a set of words the most used in the doc, it contains the number of repetition of word. As a result, which domain the most frequent words of document belongs to, we can conclude that the document also belongs to this subject area. Defining subject area of text can help to translate it into another language. Because computer can use a meaning of some words in the given text match to the subject area, if it has a knowledge about subject area. It also can be used to create automatic responser to letters in different organizations according to the meaning of letters. It can analyze the content of letter and response to the author itself. All mentioned tasks above can be done by using computional linguistics and text mining methods.

REFERENCES:

1. Соколов Артем Михайлович, «Методы нейросетевого распределенного представления и поиска сходных символьных последовательностей в задачах классификации на основе рассуждений по примерам» — Диссертация на соискание ученой степени кандидата технических наук. Киев — 2008. 159 б.
2. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. «Автоматическая обработка текстов на естественном языке и анализ данных» Изд-во НИУ ВШЭ, 2017. — 269с. ISBN 978–5–9909752-1-7.
3. Абдураҳмонова Н. З. «Машина таржимасининг лингвистик таъминоти» монография, Тошкент: Муҳаррир нашриёти, 2018. – 188б. ISBN 978-9943-5269-4-5.
4. Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil «Text Mining Methods and Techniques» International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014.
5. К. В. Воронцов, «Комбинаторный подход к оценке качества обучаемых алгоритмов» УДК 519.7:004.855.5.
6. Селезнев К. «Обработка текстов на естественном языке» 18.12.2003.

СЕКЦИЯ 6. ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ ДЛЯ ОБУЧЕНИЯ ТЮРКСКИМ И ИНОСТРАННЫМ ЯЗЫКАМ

RECOMMENDATIONS ON IMPLEMENTING THE CEFR FOR THE ASSESSMENT OF THE AZERBAIJANI LANGUAGE

Pirdas H. Muradova, Institute of Information Technology of ANAS, Azerbaijan, Baku, pirdas.davudova@gmail.com

This paper gives short information about The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) for describing language ability which was accepted in symposium by European Council. Special attention was paid here to understanding the CEFR language levels. In Azerbaijan, this standard is used for teaching and assessment of foreign languages (L2), but it didn't apply for the assessment of Azerbaijani language skills. For this reason, there were analyzed the existing problems of teaching level and assessment of Azerbaijani language. In the article, it was proposed to create the "Azerbaijani Language Examination System" based on four skills (writing, reading, listening, speaking), as in exams, such as IELTS, TOEFL, etc.

Key words: *CEFR, language standard, Azerbaijani language, assessment, teaching, learning, language ability.*

РЕКОМЕНДАЦИИ ПО РЕАЛИЗАЦИИ CEFR ДЛЯ ОЦЕНКИ СОХРАНЕНИЯ АЗЕРБАЙДЖАНСКОГО ЯЗЫКА

Пирдас Мурадова, Институт информационных технологий Национальной академии наук Азербайджана, Баку, Азербайджан, pirdas.davudova@gmail.com

В статье дана краткая информация об «Общеевропейской компетенции владения иностранным языком: изучение, преподавание, оценка» (CEFR) для описания языковых компетенций, которые были

приняты на симпозиуме Европейским Советом. Особое внимание было уделено пониманию уровней языка CEFR. В Азербайджане этот стандарт используется для обучения и оценки иностранных языков (L2), но он не применяется для оценки навыков азербайджанского языка. Были проанализированы существующие проблемы уровня обучения и оценки азербайджанского языка. В статье было предложено создание система экзаменов по азербайджанскому языку, основанную на четырех навыках (письмо, чтение, аудирование, разговор) на экзаменах, таких как IELTS, TOEFL и т.д.

Ключевые слова: *языковой стандарт, азербайджанский язык, оценка, преподавание, изучение, языковые компетенции.*

Introduction

One of the most important issues of our time in the globalization is to acquire a beautiful ethical speech culture in accordance with all the rules of the language. In general, on the basis of the formulation of difficulties and problems during use of whether the written or oral form of language stays the lack of proper language teaching and the lack of correct assessment language. During language teaching, gradually the language learner should be able to speak fluently, communicate in written and verbal form of language.

European experience in the development of language skills

In the second half of the 20th century, the concepts of plurilingualism gave rise to a new approach to language learning in Europe. In November 1991, the Council of Europe held a great symposium in Switzerland and in the symposium was offered the idea of «Common European Framework of References for Languages» (the CEFR) that is a universal framework used to evaluate any language skills. In 1992, was established a scientific working group with the support of the ALTE (The Association of Language Testers in Europe) association of Council of Europe, and the Swiss National Science Foundation, and the purpose of this research group was to create and develop a standardized system for measuring language proficiency. As a result, the research group creates a systematic scale the CEFR for measuring the level of language knowledge and skills. The original version of this document was distributed to experts of the Council of Europe countries in Strasburg at the end of 1996 and was again reviewed, and the last version of the CEFR was presented at a conference in April 1997. Finally, in January 2001, the latest official version was published with the support of the Council of Europe and Cambridge University. (Council of Europe, www.coe.int)

Today, the CEFR has been playing a central role in the teaching of many languages, not only in Europe, but also in the world (Asia and America) from the very first day of its publication Europe. At present, the CEFR has been translated into almost 40 languages. (Council of Europe, www.coe.int) As a result, the CEFR has become an internationally accepted standard for language

knowledge testing and assessment. CEFR describes in a comprehensive way of what language learners have to learn to use language for communication and what knowledge and skills they have to be able to act effectively in the language. It also covers the culture context in which language is set. The framework also give an opportunity to learners identify their level of proficiency at each stage of the learning. The Common European Framework is designed to overcome the barriers between professionals working in the field of modern languages arising from the different educational systems in Europe. The CEFR's specificity lies in the fact that it is not associated with a specific language in creating a systematic framework. It is not set of rules for a particular language. It is almost a new approach in teaching foreign languages. In other words, the CEFR does not tell practitioners what to do or how to do it, this is not a new method, but here offered ideas of different methodological options based on long years of experiences and each linguist works on their own methods and suggestions, adapting the information he received to his own experience and language features. Based on this, it is possible to obtain curriculum that responding to the characteristics and needs of a definite group of languages. (Using the CEFR 2011:4)

The CEFR adopts an action-oriented approach to the description of communicative proficiency and describes what skills the learner achieves in each level, taking into account the real needs of the learners. Activities in the CEFR are presented in four modes of communication: reception (listening and reading), production (speaking and writing), interaction (speaking and writing) and mediation (make communication possible between persons who are unable to communicate with each other directly – translation, interpretation, paraphrase etc.). (Council of Europe, CEFR 2018: 32-33)

The CEFR organizes language proficiency in six levels (A1-A2, B1-B2, and C1-C2), which can be grouped into three main levels: Basic User, Independent User and Proficient User, and that can be further subdivided according to the needs of the local context. (Council of Europe, CEFR 2001: 23) The levels are defined through ‘can-do‘ descriptors. For example, if any person who learns a foreign language has intermediate level in writing and reading, he or she may have an elementary level in listening and speaking.

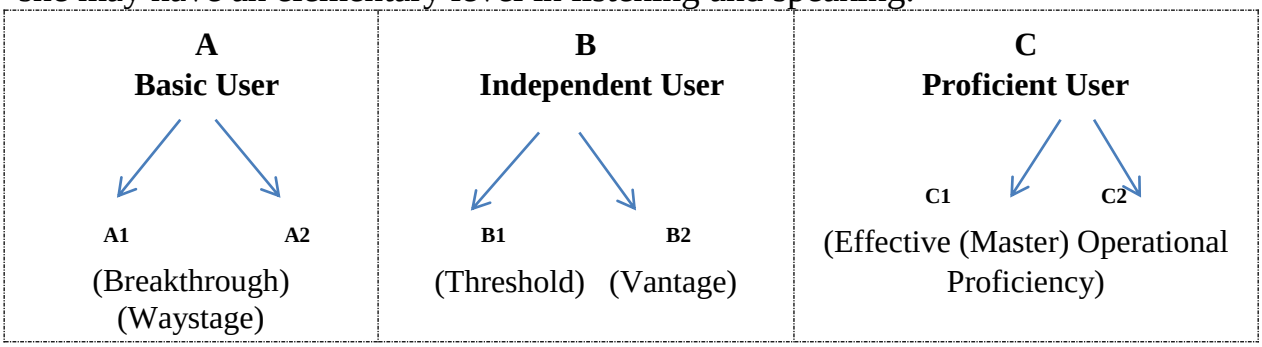


Fig. 1. Three broad categories with six reference level ((Council of Europe, CEFR 2001: 23)

Another reason the CEFR carries out this kind of division is that which audience belong the language learners. In the CEFR each act of language use is grouped in spheres of action in which social-life organized. These particular situations in spheres of action are called domains and divides into four domains: personal, public, occupational and educational.

- the **personal** domain, in which the person concerned lives as a private individual, centred on home life with family and friends, and engages in individual practices such as reading for pleasure, keeping a personal diary, pursuing a special interest or hobby, etc.;
- the **public** domain, in which the person concerned acts as a member of the general public, or of some organization, and is engaged in transactions of various kinds for a variety of purposes;
- the **occupational** domain, in which the person concerned is engaged in his or her job or profession;
- the **educational** domain, in which the person concerned is engaged in organised learning, especially (but not necessarily) within an educational institution. (Council of Europe, CEFR 2001: 45)

If say more specifically it comes from the real-life needs of the language learners, for example, to get a visa for higher education or to gain professional work experience. The description of all levels is based on skills and is indicated by the "can do" descriptors as shown in the table below (Table 1) (Council of Europe, CEFR 2001: 24). As mentioned in the «Common European Framework of Reference for Languages: Learning, Teaching, Assessment»:

«Can do descriptors are provided for reception, interaction and production… Can do descriptors are provided for some of strategies employed in performing communicative activities. Strategies are seen as a hinge between the learner's resources (competences) and what he/she can do with them.» (Council of Europe, CEFR 2001: 24-25)

Table 1.

COMMON REFERENCE LEVEL: GLOBAL SCALE (COUNCIL OF EUROPE 2001: 24)

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/her fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed

		text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/her and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

The descriptors of communicative activities are very suitable for teacher or self-assessment with regard to real-life tasks. The CEFR also offers a draft for self-assessment direction tool based on the six levels. Its aim to help language learner to profile their language skills and give more detailed descriptors in order to self-assess their level of proficiency (Council of Europe (CEFR) 2001: 26-27). In general, the results have shown that the CEFR has a major impact on language education in the Europe. In some countries, the CEFR has influenced both strategies of the language policy the development of language teaching methodology and the assessment of language proficiency. Some countries try to use all the capabilities of the CEFR in language teaching, as well as in the development of new teaching aids and assessment of language proficiency.

The suggestions on teaching Azerbaijani language as a foreign language.

Today, not only in the world but also in Azerbaijan, during the development of science and technology, the other task (issue) is to introduce Azerbaijani language to the world. In other words, one of the issues about the language industry is related to the migration of foreigners. Every year, hundreds of foreigners migrating to our country with the aim of education, establishing trade relations and so on. It is necessary to organize the teaching of Azerbaijani language in order to eliminate language problems of foreign students during education and to help them feel easy in society, at the same time to introduce Azerbaijan and its culture to them. In many countries, one of the main conditions of these countries is in a short time to get a certificate by giving certain examinations (IELTS, TOEFL, etc.) either for education or to get visa to live or to work. If you are applying for a visa to move to or stay in the UK, IELTS(International English Language Testing System) result is one of the needed document. IELTS is the world's most popular English language test and a high-level English language test for study, migration, and work. There are three types of IELTS: IELTS General, IELTS Academic and IELTS Life-skills. These differ in content and address different target groups:

- **IELTS Academic** is intended for people who want to attend study programmes at universities and other institutions of higher education, which are taught in English. You can also take this version of IELTS to register within a professional body in an English-speaking country.
- **IELTS General Training** is a good choice if you want to migrate to an English-speaking country and work there, or if you plan to attend a secondary school.
- **IELTS Life Skills**, the test for UK Visas and Immigration, is appropriate if you wish to immigrate to or obtain citizenship in the United Kingdom. Unlike the other two versions, the Life Skills test assesses only your speaking and listening English skills, at levels A1 or B1 of the Common European Framework of Reference for Languages (CEFR). (IELTS, www. ielts.org)

In the globalization, the influence of information and communication technologies, the development of the economy, culture and science has created opportunities for wide use and development of the Azerbaijani language. The aim of the article is to apply results of years of research in Europe in Azerbaijan, not only in the teaching methods of language but at the same time, in the compilation of textbooks with additional course materials for language teaching according to the CEFR's requirements and new assessment criteria for language proficiency. After many world language level exams stands the CEFR, formed as a result of the efforts of the Council of Europe.

Today, for recognition of Azerbaijani language between other languages, it is necessary to carry out certain studies and then must be realized achieved results. The promotion of the Azerbaijani language in international arena, the organization of the teaching of Azerbaijani language to foreign citizens, in general, teaching Azerbaijani language as foreign language the following suggestions were made based on international experience and by adapting this practice to the Azerbaijani language:

1. Preparation of the curriculum for the teaching of Azerbaijani as a foreign language
2. Preparation of curriculum textbooks and additional course materials (audio, video, software, etc.)
3. Development of new assessment criteria based on international standards (on levels)
4. Creation of a single exam in the Azerbaijani language that is a "Test of Azerbaijani language as a Foreign Language"
5. Creation of an electronic database with the support of ICT for the realization of the above-mentioned proposals.

It is possible to use suggested system for the Azerbaijani language to check Azerbaijani language proficiency of foreign citizens (students) and during the admission of students form other language section (both Russian and English) to the university and as well as during the employment of Azerbaijani citizens (if Azerbaijani language is required at a high level, like jobs in radio, television, etc.).

Conclusion

The investigations show that there enough things to do in the field of linguistics and in solving these issues should be done in collective work. The research should also be conducted to study and introduce new teaching methods based on European experience. At the same time, it is necessary to develop new criteria for language skills assessment. In the article, it was proposed to create the "Azerbaijani Language Examination System" based on four skills (writing, reading, listening, speaking), as in exams, such as IELTS, TOEFL, etc. In recent years, through the increase of migration of foreigners to Azerbaijan, have been created conditions for the promotion of the language in the world. In the article were proposed suggestions about teaching Azerbaijani language to foreign citizens based on the European system of teaching and assessment.

REFERENCES:

1. Council of Europe. www.coe.int
2. Council of Europe (2001) Common European Framework of References for Languages: Learning, Teaching, Assessment, 2001, 260p. www.coe.int/lang-CEFR

3. Council of Europe (2018) Common European Framework of References for Languages: Learning, Teaching, Assessment, Companion voume with new descriptors, 2018, 238p. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989
4. https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages
5. Using the CEFR: Principles of Good Practice, october (2011), University of Cambridge, 41p.
6. Cambridge Assessment English http://www.cambridgeenglish.org/exams-and-tests/cefr/
7. IELTS, www. ielts.org

HEMISPHERIC ASYMMETRY AND LANGUAGE LEARNING

Alymjan Zakirov[1], Yulia Seredina[1], Tashbolot Sadykov[2], [1]International Alatoo University, [2]Bishkek Humanity University, Bishkek, Kyrgyzstan, tash_sadykov@mail.ru, ali_zakir_50@bk.ru

The article highlights the questions of lateralization and hemispheric asymmetry in learning of second languages in the context of unrelated languages: the English and Turkic stock languages. Hemispheric asymmetry is believed to carry an important function in cognitive processes. Each hemisphere plays a specific but separate role in comprehension and realization of language competence. The peculiarities of the work of hemispheres of learner influences greatly the process of learning and teaching languages.

Key words: *lateralization, hemispheric asymmetry, language learning and teaching, comprehension, language competence, function of language.*

ПОЛУСФЕРНАЯ АСИММЕТРИЯ И ОБУЧЕНИЕ ЯЗЫКАМ

Алимжан Закиров[1], Юлия Середина[1], Ташполот Садыков[2]
[1]Международный университет Ала-Тоо,
[2]Бишкекский гуманитарный университет, Бишкек, Кыргызстан
tash_sadykov@mail.ru, ali_zakir_50@bk.ru

В статье рассматриваются вопросы асимметрии и латерализации полушарий головного мозга при изучении вторых языков в контексте неродственных языков: английского и тюркских языков. Однозначно известно, что асимметрия полушарий головного мозга несет особую функцию в познавательном процессе. Каждое полушарие играет специфическую, но определенную роль при понимании и реализации языковой компетенции. Особенности работы полушарий индивидуума, изучающего языки, оказывают большой эффект при обучении и преподавании языков.

Ключевые слова: *латерализация, асимметрия полушарий, обучение и преподавание языков, понимание, языковая компетенция, функция языка.*

The lateralization of brain function is referred to how some cognitive processes or neural functions tend to be more predominant in one hemisphere than in the other one. The brain of humans is divided into two hemispheres — right and left. Scientists and researchers go on to investigate how some cognitive functions have a tendency to be dominated by one side or the other, or how they are lateralized. Lateralized regions of brain sub serve functions such as

visuospatial processing and language. It was conjectured that individuals can be right-brain dominant or left-brain dominant. This dominance is based on personality and cognitive style.

Language and speech understanding are usually regarded by linguists and neuroscientists as heavily lateralized functions. Many aspects of language are found to be localized in the left hemisphere, while less are localized in the right hemisphere as the left one is usually dominant. Until the late twentieth century, scientists believed that almost everything you can do, including the ability to learn foreign languages, was determined by which hemisphere of brain you are using. However, the left or right-brain construct helps to define a useful learning style continuum, with implications for second language learning and teaching, therefore the theme of the current research is relevant. The practical value of this work consists in the fact that this theoretical information can be taken into account by foreign language teachers and used in creating lesson plans. The aim of the current article is to examine the hemispheric asymmetry and its influence on language learning and teaching. To achieve the aim, the following objectives have been set: To give the definition of hemispheric asymmetry.

2. To describe the influence of hemispheric asymmetry on language learning and teaching.

Hemispheric lateralization is referred to the division of functions of the left and right hemispheres of the brain. The medial longitudinal fissure separates the human brain into two cerebral hemispheres which are connected by the corpus callosum. Even though the macrostructure of the two hemispheres appears to be almost the same, dissimilar composition of neuronal networks allows for particular function that is different in each hemisphere. If one hemisphere is more deeply involved in a particular function, it is often thought as being dominant.

The two brain hemispheres are not identical. Some functional differences between the right and the left sides of the brain, so-called functional hemispheric asymmetries, are observed for some cognitive functions. For example, most people show a right-hemispheric dominance for visuospatial processing and a lefthemispheric dominance for processing and production of a language . Besides these functional hemispheric asymmetries, anatomical differences between the two sides of the brain,such as volume or size of a definite area, can be found in many brain regions, they are called structural hemispheric asymmetries (Деглин, 1975)

It is now believed that the left hemisphere in right-handed people plays a more predominant role in the expressive and impressive speech, reading, writing, verbal memory and verbal thinking. The right hemisphere is more important for non-verbal thinking, for example, an ear for music, visual-spatial orientation, non-verbal memory and critical thinking.

It has been shown that the left hemisphere is more focused on the prediction of future situations, and the right — to interact with the experience and with the relevant events occurring.

In the process of individual development the vividness of asymmetry is changing — lateralization of brain functions appears. Recent studies suggest that the hemispheric asymmetry contributes to the manifestation of a person's high intelligence. But there still exists an interchangeability of the cerebral hemispheres.

It is important to note that a particular type of hemispheric response is not formed in the early childhood. In the early stages of ontogenesis most children reveal shaped, right-brain type of response, and only at a certain age (generally from 10 to 14 years) the right or the left-brain dominance is fixed. This is also proved by the fact that functional asymmetry of the brain of illiterate people is less than that of literate people.

1. Asymmetry becomes stronger in the learning process: the left hemisphere becomes specialized in gestural operations, and the right hemisphere — in the figurative ones (Спрингер, Дейч, 1983)

Asymmetry is totally important for completing the complex tasks we take for granted. In the conversation with someone, the left brain is processing the verbal language, while the right brain interprets inflection and tone. If the person is asked to imagine a scene, the left brain will create the details, while the right brain handles the overall sizes, shapes and their spatial locations. Without the both sides processing all of these details at the same moment, we would never be able to function at the advanced level we do.

Each hemisphere plays a specific but separate role in the understanding, use and production of language. This becomes obvious when looking at patients that have one-sided hemisphere damage and their language deficits can be observed. For example: when the left hemisphere is damaged, the right hemisphere is used to take over some functions through brain plasticity, and this damage of one hemisphere and compensation by the another hemisphere creates language understanding and deficits that can be studied and help to determine the interaction of brain areas in the language processes.

The language production and language comprehension involve the coordination of different subprocesses at the same time. Even though there are the debates on how these subprocesses work in cooperation and how thinking and comprehending can be changed, the anatomical basis and role of a loop including the Wernicke's and the Broca’s area are usually agreed upon.

Neuroscientists usually agree that around the lateral sulcus in the left hemisphere, there is a neural loop that is involved in producing and understanding spoken language. At the front end of this loop lies the Broca’s area, which is generally associated with the language production, or language outputs. At the other end, or in the superior posterior temporal lobe, lies the

Wernicke's area, that is frequently associated with the processing of words that we hear, or language inputs. Broca’s and Wernicke's areas are connected by a large bundle of nerve fibres that are called the arcuate fasciculus (Amunts, 2010).

Scientists have long held the theory that the left and right hemisphere of' your brain control different functions when it comes to learning. It is thought that the left hemisphere controls language, math and logics, while the right hemisphere is thought to be responsible for visual imagery, spatial abilities, the ability to recognize faces and music. The left hemisphere also controls the movement on the right side of human's body.

Human's language control center is made up of the parts of temporal lobe, parietal lobe and the occipital lobe, that are contained in the left hemisphere of the brain. ln these lobes the two areas known as the Wernicke’ s area and the Broca’s area help you recognize and understand, read and speak language patterns — as well as the ability to learn foreign languages. Language is a code consisting of symbols that can be connected to form letters, words and phrases. Human' s brain allows people to crack that code and relate the letters, words and phrases to a specific meaning. People learn the sounds that form the words and put these words in a sequence so that the listener can understand. The brain tells the tongue, mouth and voice box to work on a millisecond-by- millisecond basis, which allows you to speak.

While each hemisphere of the brain is responsible for completing certain tasks and learning definite functions, the integrati01NåGo hemispheres is responsible for most of the characteristics shared by humans. While the left hemisphere is mainly responsible for human's ability to learn a new language, scientists found that the opposite hemisphere can really take over those responsibilities in some cases. Specifically, the most important parts of language like grammar and vocabulary are typically "lateralized" to the left hemisphere. It means that thinking about grammar and vocabulary mostly happens in the left half of the human's brain. This is especially true in the right-handed people. But not all aspects of language are controlled by the left hemisphere. Other important aspects of language, such as intonation and accentuation are often lateralized to the right hemisphere.

Productive language skills such as writing and speaking is an example of how hemispheric asymmetry can be different for right-handed and left-handed people. The production of language is lateralized to the left hemisphere in about 90% of right-handed people. But almost in 50% of left-handed people language production is controlled by both hemispheres, or more by the right hemisphere.

2. If you force left-handed children to write with their right hand they may have the following abnormalities: mental retardation, psychosis, speech defects (Brown, 2007).

3. However, the left or right-brain construct helps to define a useful learning style continuum, with implications for second language learning and teaching. It

is considered that left-brain-dominant second language learners prefer a deductive style of teaching, while learners with right-brain-dominance appear to be more successful in an inductive classroom environment. It was concluded that left-brain dominant second language learners are more successful at gathering the specifics of language, producing separate words, carrying out sequences of operations, and dealing with classification, abstraction, reorganization and labeling. But right brain-dominant second language learners are usually better at dealing with metaphors,with whole images, with generalizations, and with emotional reactions and artistic expressions. This may suggest a great need to perceive whole meanings in those early stages, and to analyze and monitor oneself more in the later stages (Blake, 2015).

The work was devoted to hemispheric asymmetry and language learning. The aim of the current research paper is to examine the hemispheric asymmetry and its influence on language learning and teaching. To achieve the aim, two objectives were set. All the objectives were completed. The notion of hemispheric asymmetry was given. It was described as an important factor in learning a foreign language. Neuroscientists have not yet provided clear data whether the phenotypic differences in the strength of right-dominant or left-dominant networks exist. Popular psychological articles often say that each side of the brain does things that the other does not. For example, a widespread mistake is the idea that the left side of the brain does all of a person's logical thinking,while the right brain does all the creative thinking. It is often said that a creative person is right-brained, and a person who is more logical is left-brained.

The following conclusion can be drawn: the both hemispheres work together on both logical and creative thinking. Important parts of language like grammar and vocabulary are usually "lateralized" to the left hemisphere. This means that most thinking about grammar and vocabulary happens in the left half of the brain. But not all parts of language are controlled mostly by the left hemisphere. Other important parts of language, like intonation and accentuation are often lateralized to the right hemisphere.

ЛИТЕРАТУРА:

1. Деглин В. Функциональная асимметрия — уникальная особенность мозга человека // Наука и жизнь. 1975. #1.С.104-115
2. Закиров А. К вопросу психолингвистических особенностей обучения иностранным языкам //Международная научно-практическая конференция « Международное научное обозрение проблем и перспектив современной науки и образования», Лондон, Великобритания, 7-8 октября 2015 . « Проблемы науки».
3. Закиров А. Интеллект- карты Энтони Бьюзена в изучении народного эпоса «Манас». Общие принципы когнитивного подхода»

//Международная научная конференция «Fundamental and Applied science», Шеффилд, Великобритания, 30 октября-07 ноября 2016.
4. Спрингер С., Дейч Г. Левый мозг, правый мозг. Москва: Мир, 1983. С. 230-234.
5. Amunts K. Structural indices of asymmetry in the two halves of the brain. Cambridge, MA: The MIT Press, 2010. P. 145-176.
6. Brown D. Principles of Language Learning and Teaching. Englewood Cliffs: Prentice Hall Regents, 2007. P. 125- 126.
4. Blake C. Left or Right Hemisphere of the Brain: Learning a Foreign Language. URL: http://education.seattlepi.com/left-right-hemisphere-brain-leaming-foreignlanguage- 1343 .html (Accessed: 20.12.2015).

CREATING CALL IN UZBEK LANGUAGE

Z. Xoliqova, U. Hamroyev, TSUULL, zakhrokholikova1995@gmail.com

This article is about creating CALL in Uzbek language. Teaching program embrace writing, listening, speaking, reading. General conception of teaching of language by program is explained in this paper.

Key words: *lingvodidactic computer programme, language's base, vocabulary, lexicography, morpho- syntactic errors, interactive means.*

СОЗДАНИЕ CALL ДЛЯ УЗБЕКСКОГО ЯЗЫКА

З. Холикова, У. Хамроев, Ташкентский государственный университет узбекского языка и литературы им. Алишера Навои, Ташкент, Узбекистан zakhrokholikova1995@gmail.com

В статье описывается создание программы CALL на узбекском языке. Программа обучения включает в себя письмо, аудирование, говорение, чтение. Здесь рассматривается общая концепция преподавания языка по программам.

Ключевые слова: *лингводидактическая компьютерная программа, языковая база, словарный запас, лексикография, морфосинтаксические ошибки, интерактивные средства.*

In the 20th century the interest to teach language by computer became stronger. It is named CALL (Computer — Assisted Language Learning) and the first conference hold in 1985 in Budapest. In this conference teaching foreign language by computer was discussed, scientists suggested new methodics and lingvodidactic computer programme. Using audio and video programmes was observed.

Then the whole world researches connected to CALL was named World CALL. From its successful beginning in Melbourne (1998), and equally ambitious second conference in Banff (2003), World CALL has consistently attracted a broad spectrum of international CALL researchers, representing both highly developed and less well-served nations and highly experienced and neophyte researchers alike.Those conferences represented collective aims of the founding organizations(EUROCALL, CALICO, ATELL, CCALL) and tended to reflect th early strength of CALL in Europe and the English speaking world. LET (The Japan Association for Language Education and Technology), the hosts of World CALL III in Fukuoka in August 2008, were anxious to continue the same wide- ranging, global tradion whilist at the same time providing an

attractive, accessible venue for the rapidly growing numbers CALL colleagues in Asia-Pacific regions.[1]

The programmes which taught languages were created. They are SOAS (taught and online distance learning), UCL (Study programme), DAAD (German program) etc.

How to create CALL in Uzbek language? What should we do? There are lot of means to learn Uzbek language. They:

- New texnologies, New pedagogies
- Developing Languge Skills through Texnology
- Materials Design and Development
- Learner training
- Teacher education

To do all counted works must create uzbek language's base. This base might include whole language features. It requires special models vocabulary. For example, in English sentence sequence: **subject+verb+adjective+object+adverb – I gave my book to Lucy**. In Uzbek language: **subject+adjective+object+adverb+verb – Dildora bu mavzuni yaxshilab o'rgandi.** Models can lead to learn language easily.

New technologies, New pedagogies. It covers a range of contemporary technologies in use language learning together with their attendant pedagogies. They use various methods to learn: Game methods, audio and videos, conversation methods.

Developing Languge Skills through Texnology. It considers the learning and teaching of pronunciation, writing, and grammer with specially designed programs and applications to address these goals. These part includes some of the more technical contributions to the volume, illustrating some of the programming and linguistic challenges being adressed in the CALL arena. **Materials Design and Development.** It represents a well-established area of activity in CALL. This part describes CALL structures, pictures and it does programmes sight. Materials should be better if they use Uzbek national features.

Learner training. It overvalues or overlooks, helps ensure optimal value in any CALL activity, and therefore must form an important component in teacher preparation.

Teacher education. It provides for perspectives on teacher education. It also includes psychologic skills and other means are used to motivate learning Uzbek language.

There is not the programme which taught Uzbek language. We can model other foreign languages, but Uzbek language features should notice. Firstly, Uzbek phonetic system and alphabet are understood: A (ei) in English – A (a) in Uzbek.

Then we should understand themes with sections (lexicography, phonetics, morphology…). One of merits in CALL is to design computer algorithms that are

capable of providing effective diagnostic feedback to language learners. The program can discuss language learner modeling and ways in which the computer may relate morpho- syntactic errors to the student's underlying knowledge of the grammatical system.

The program should ensure writing, listening, speaking, understanding, reading. Uzbek audios (by voice recorder), videos, interactive means, Uzbek texts, Peer review, Uzbek songs will help to do these functions.

Creating CALL in Uzbek language can help to learn Uzbek language easily and foreign learners also know about Uzbek languages features and Uzbek traditions.

REFERENCES:

1. Levy, M., Blin, F., &Takeuchi (2011). WorldCALL International Perspectives on Computer -Assisted Language Learning.
2. Rakhimov, A.,(2011). Computer linguistics.

ONA TILI DARSLARIDA AKTDAN FOYDALANISH USULLARI

K. Mavlonova[1], D. Hasanova[2], M. Xudayarova[3], D. Kabulova[3],
[1]Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti,
[2]Qo‘qon pedagogika instituti,[3] Nukus davlat pedagogika instituti

Tezisda o‘zbek tilini o‘qitishda AKT lardan unumli foydalanish va uni darsga tadbiq qilish borasida fikrlar o‘rtaga tashlangan. Unga ko‘ra auditoriyaning yoshi va mavzuning semantikasiga qarab pedagogik texnologiyalar va metodlar tanlanishi lozimligi qayd qilingan.

Tayanch so‘zlar: *metod, ta’lim, AKT, alifbo.*

WAYS OF USAGE ICT IN THE LESSON OF NATIVE LANGUAGE

K. Mavlonova[1], D. Hasanova[2], M. Xudayarova[3], D. Kabulova[3],
[1] Tashkent State University of Uzbek Language and Literature named after Alisher Navoi,
[2] Kokand pedagogical institute, [3] Nukus State Pedagogical Institute

This article deals with the use of ICT in the process of teaching native language and its methodological tools. According to this, each method is selected according to topics and is determined by the age of the audience.

Key words: *Method, Education, ICT, Alphabet.*

СПОСОБЫ ИСПОЛЬЗОВАНИЯ ИКТ НА УРОКАХ РОДНОГО ЯЗЫКА

К. Мавлонова[1], Д. Хасанова[2], М. Худаярова[3], Д. Кабулова[3]
[1]Ташкентский государственный университет узбекского языка и литературы
имени Алишера Навои, [2]Кокандский педагогический институт
[3]Нукуский государственный педагогический институт

В статье рассматривается использование ИКТ в процессе обучения родному языку и его методологические инструменты. Каждый метод обучения подбирается по темам и определяется возрастом аудитории.

Ключевые слова*: метод, образование, ИКТ, алфавит.*

O‘zbekistonda umumta’lim fanlarini o‘qitishning uzluksizligi va izchilligini ta’minlash, zamonaviy metodologiyasini yaratish, umumiy o‘rta va o‘rta maxsus, kasb-hunar ta’limi davlat ta’lim standartlarini kompetensiyaviy yondashuv asosida takomillashtirish masalasiga alohida e’tibor berib kelinmoqda [1; 117-b.].

Umumiy o‘rta va o‘rta maxsus, kasb-hunar ta’limi muassasalarida ona tili fanini o‘qitishning asosiy vazifalaridan biri «ona tilining keng imkoniyatlaridan unumli foydalangan holda to‘g‘ri va ravon bayon eta olishni rivojlantirishga qaratilgan lingvistik kompetensiyalarni shakllantirishdan iborat»ligi belgilab qo‘yilgan.

O‘quvchilarning murakkab til hodisalarini oson egallashlari, o‘zlaridagi topqirlik xislatlarini tarbiyalashlari, murakkab masalalarni yechish orqali aqllarini peshlashlarida AKTdan foydalanish muhim ahamiyatga ega. Quyida ana shunday usullar haqida so‘z boradi.

«Jadvalli labirint». Bu usulda o‘quvchilarga proyektor orqali quyidagi jadval namoyish etiladi. O‘quvchilar birinchi ustunda turgan iboralarga mos bo‘lgan fe’lni ikkinchi ustundan qidirib topishlari va ularni birlashtirishlari lozim.

T /r	Iboralar	Fe’llar
1	Oyog‘iga bolta urmoq	Urushmoq
2	Og‘ziga tolqon solmoq	vafot qildi
3	Holdan toymoq	Hayron qolmoq
4	Boshi ko‘kka yetmoq	Halaqit bermoq
5	«Ali» desa «bali» demoq	Aytishmoq
6	Yoqasini ushlamoq	Charchamoq
7	«San-man»ga bormoq	Jim turmoq
8	Yer tishlamoq	Xursand bo‘ldi
9	Tutuni osmonga chiqmoq	Jahli chiqmoq
10	Bir yostiqqa bosh qo‘ymoq	Urmoq
11	Almisoqdan qolgan	Eskirmoq
12	Ikki oyog‘ini bir etikka tiqmoq	Turmush qurmoq
13	Qo‘l ko‘tarmoq	Shoshilmoq

O‘qituvchi o‘quvchilar bilan birga ushbu iboralar ishtirokida gap tuzish, so‘z birikmalari tuzish va ular orasidagi farqni aytish topshirig‘ini bajaradi. Bu usul orqali o‘quvchilarda predmetga oid nutqiy va lingvistik kompetensiyalarni shakllantirish imkoniyati yuzaga keladi.

«Alifbo va so‘z». O‘quvchilarning so‘z boyligini oshirish, ularda tezkorlik va topqirlik malakalarini rivojlantirishda «Alifbo va so‘z» usulini qo‘llash katta

samara beradi. Monitorda alifbo namoyish qilinadi va o‘quvchilarga alifbo tartibida fe’lga oid bo‘lgan so‘zlarni aytish so‘raladi. Bir o‘quvchi «A» harfi bilan boshlanadigan fe’lga misol aytsa, keyingi o‘quvchi «B» harfi bilan boshlanadigan fe’lga misol aytadi. Bir so‘zni birdan ortiq qaytarish mumkin emas. O‘yin shu tariqa davom etadi. Ushbu o‘yinning ikki o‘quvchining orasida ham, bir necha o‘quvchilar guruhi bilan ham o‘tkazish mumkin. Masalan: aytmoq-bosmoq-demoq-ekmoq-farqlamoq-gaplashmoq-ho‘plamoq-irg‘itmoq-janjallashmoq-kesmoq-lapshaymoq-maydalamoq va hokazo. Ushbu usul o‘quvchini faollashtiradi, ijodkorlikka va rivojlanishga undaydi. O‘rganilayotgan mavzuga ajratilgan vaqtni o‘g‘irlamagan holda qo‘llaniladigan mazkur usul dars samaradorligini oshirishga xizmat qiladi. **«Bu bizniki» usuli.** O‘quvchilar bordi, yugurdi, yozdi kabi fe’llarning harakat bildirishini yaxshi tushungani holda, uxladi, tingladi, mudradi so‘zlarining holat bildirishini izohlashda qiynaladilar. Shuning uchun ularga harakat va holat tushunchasini aniqlab olish yo‘llarini batafsil tushuntirish darkor.

O‘qituvchi sinfdagi o‘quvchilarni uch guruhga bo‘lgan holda, har bir guruh uchun quyidagi tushunchalar yozilgan tarqatmalarni oldindan berib qo‘yadi.

1. Jismoniy faoliyat: yugurmoq, turmoq, kuylamoq, o‘smoq;

2. Ichki kechinmalar: o‘ylamoq, xursand bo‘lmoq, esga olmoq, sevinmoq;

3. Bir holatdan ikkinchisiga o‘tish jarayoni: qizarmoq, yaltiramoq, jimirlamoq.

So‘ng slayd orqali gaplar uchligini taqdim etib boradi. Gaplar tarkibidagi fe’llar qora kursiv bilan ajratilgan bolib, o‘qituvchi o‘quvchilar e’tiborini ana shu fe’llarga qaratadi va ushbu fe’llarning o‘quvchilardagi tushunchalar bilan aloqasi bor yoki yo‘qligini so‘raydi. Gapda berilgan fe’l qaysi guruhdagi tushuncha bilan bog‘liq bo‘lsa, o‘sha guruh «Bu bizniki» deya javob beradi va o‘z javobini izohlaydi.

O‘z javobini izohlagan o‘quvchidan ana shunday fe’lga yana bir misol aytish so‘raladi. Keyingi guruh ana shu so‘z ishtirokida yangi gapni hosil qiladi. Namuna:

1. *Bolaga ins-jins ziyon yetkazmasligi maqsadida go‘daklar yostig‘i ostiga non* ***qo‘yiladi*** – jismoniy faoliyat;
2. *Ra’no javob berish o‘rniga* ***ikkilanib qoldi –*** ichki kechinma;
3. *Dov-daraxt, o‘t-o‘lan xuddi yer kabi* ***oltin tus oladi*** – birinchi holatdan ikkinchisiga o‘tish.

Ushbu topshiriq bir necha gaplar uchligini taqdim etgan holda olib boriladi. Topshiriq bajarib bo‘lingach, o‘qituvchi slayd orqali quyidagi qoidani ko‘rsatadi. Unga o‘z izohini berib o‘tadi. Shaxs va narsalarning **jismoniy faoliyati** natijasida ro‘y bergan harakatni bildiruvchi fe’llar harakat fe’llari sanaladi. Shaxsning **ichki kechinmalari** va narsalarning **bir holatdan ikkinchi holatga o‘tish jarayoni**ni ifodalovchi fe’llar esa holat fe’llari sanaladi.

Mazkur usullarni amalga oshirishda o‘qituvchi tashabbuskor, tashkilotchi bo‘lishi kerak. U o‘zi qo‘shilgan jamoani uyushtiradi, ishtirokchilarda mas’uliyat hissini o‘stiradi, ularni jadal ishlashga o‘rgatadi. Tavsiya etilayotgan interaktiv usullar o‘rganilayotgan tilning qurilishini o‘rganishga va amaliy ko‘nikmalarni egallashga yo‘naltirilgan o‘yinlar hamda nutqning ravon bo‘lishi va erkin muloqotga xizmat qiluvchi o‘yinlar [2] shaklida amalga oshiriladi.

AKTdan foydalangan holda qo‘llaniladigan bunday usullar orqali yaxshi va bo‘sh o‘zlashtiradigan o‘quvchilarning imkoniyatlari, kuchlari bir oz bo‘lsa-da, tenglashadi. Ayniqsa, guruhlarga bo‘linib o‘ynaladigan o‘yinlarda shunday holat ta’minlanadi. Shunga ko‘ra ham darsdagi so‘z o‘yinlari – o‘quvchilarning qobiliyatlari, ijodkorligini rivojlantirishning muhim vositalaridan sanaladi.

ADABIYOTLAR:

1. Ўзбекистон Республикаси Вазирлар Маҳкамасининг «Умумий ўрта ва ўрта махсус касб-ҳунар таълимининг давлат таълим стандартларини тасдиқлаш тўғрисида»ги Қарори. Ўзбекистон Республикаси қонун ҳужжатлари тўплами. 2017 йил, 14-сон.
2. Bowster J. The Primary English Teacher’s Guide, London, 1992.

THE DEVELOPMENT OF COMPUTATIONAL LINGUISTICS IN UZBEKISTAN

T. Nasrullaeva, Samarkand state institute of foreign languages, Samarkand, Uzbekistan, toz81@mail.ru

The article aims to study the practice of teaching computational linguistics at universities in Central Asia. This article also addresses the problem of developing and expanding blended learning for graduate students in computer linguistics with linguistics and computer science.

Key words: *computational linguistics, general linguistics, morphology, inflectional morphology, semantics, lexeme, morpheme.*

РАЗВИТИЕ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В УЗБЕКИСТАНЕ

Т. Насруллаева, Самаркандский государственный институт иностранных языков, Самарканд, Узбекистан, toz81@mail.ru

В статье описывается исследование практики преподавания компьютерной лингвистики в университетах Центральной Азии. Также в этой статье рассмотрена проблема развития и расширения смешанного обучения для студентов магистратуры по компьютерной лингвистике с языкознанием и компьютерными науками.

Ключевые слова: *компьютерная лингвистика, общая лингвистика, морфология, флективная морфология, семантика, лексема, морфема.*

The Internet nowadays is a global information source, where enormous information of all kind is stored, requires intelligent information processing tools (i.e., computer applications) to help the information seeker to retrieve the stored information. To build these intelligent information processing tools, we need to build computer applications that understand human language. That is why where computational linguistics becomes important in Uzbekistan. We need to develop a systematic understanding of Uzbek language for developing Master Program on Computational Linguistics for students.

In the following article, I will describe MA programs presented by Uzbekistan in Computational Linguistics. MA Program in Computational Linguistics is a specialized degree offered by CLASS project which is sponsored by Erasmus+ Action 2 CBHE. We offer an accessible, intensive two-year curriculum for students with a linguistics, language, computer science, mathematics, or science background, including students without prior study of computer science and linguistics.

The curriculum is designed to progressively prepare students with the tools they need for a successful long-term career in CL. The two-year time frame allows all students to establish a strong foundation in both computer science and linguistics—including filling in any gaps in a computer science, linguistics, or mathematics background—as well to complete advanced coursework in statistical natural language processing, machine learning, and a range of topics within applied and theoretical CL.

The main topics including in the curriculum of Uzbek Universities are Computational Morphology, Computational Semantics, Machine Learning for NLP, Programming for NLP.

By the end of the M.A. program, students are supposed to have obtained the following qualifications:

- Extensive knowledge of modern formal language and grammar theories and methodology and techniques of language description, as well as the ability to implement these in Computational Linguistics
- The ability to work both practically and scientifically using the techniques and methods of computational linguistics for the analysis of natural languages
- Insight into the motivation and rationale of programming methods of computational linguistics, relevant data types, grammars and abstract machines
- Competence with the academic fundamentals and with academic problem-solving either in Applied or Theoretical Computational Linguistics
- Extended knowledge in a discipline neighboring Computational Linguistics, e.g., General Linguistics, Uzbek Linguistics or Kazakh Linguistics.

For development and implementation of this program teachers and trainers of Uzbekistan and Kazakhstan work and design in collaboration with partner universities. Moreover in Spain University of A Coruna was organized the special training courses for Uzbek and Kazakh teachers which was held from 17.06.2018 to 04.07.2018 . The program present the full information on the Morphology, Semantics, Syntax,NLP and many other subjects on CL. In the lesson of Morphology we studied and implement in our uzbek curriculum. As we know Morphology has two branches:

Derivational morphology is study of how new words are formed from an existing word, often by adding an affix,for example: happiness and unhappy derive from happy .

Inflectional morphology (only for inflective languages, i.e.: English, Spanish and other IE languages)

INFLECTTIONAL MORPHOLOGY

- Study of the processes that distinguish the forms of words in certain grammatical categories (**complementary distribution**)
- English has a fairly limited inflectional system
- Other key concepts in morphology which were presented:
- Lexeme
- Grammatical category
- Morpheme
- Morph
- Allomorph
- There are the difference of lexeme and morpheme:
- Boys
- boy- (lexeme = young mele human being)
- -s (morpheme = more than one, plural)
- He stopped
- stop- (lexeme = cease movement)
- -(p)ed (morpheme = past tense)

Moreover we can connect morphology with computational morphology. Computational morphology is a processing of word forms graphemic (written) phonemic (spoken)

Computational Linguistics is around us as it has many different applications:

Spell-checkers

Automatic hyphenation

These tasks = hard problems for a computer programme

Above mentioned materials is one of the samples we intend to conduct in our lessons for MA students in Uzbekistan. Since this project has been approved on the understanding that it will benefit the CA languages, the focus of at least the first version of the Master's should be on preparing students to develop these resources, while also learning how they can be applied to future developments.

BIBLIOGRAPHY:

1. Igor A. Bolshakov and Alexander Gelbukh Computational Linguistics Models, Resources, Applications, Mexico 2004

2. N.Abdurakhmanova N. Problems and perspectives of computational linguistics in Uzbekistan.

3. Interdisiplinary Master Program on CL at Central Asian Universities, English Morphology Prof. Isabel Moskovich. (PPT)

СЕКЦИЯ 7. НАЦИОНАЛЬНАЯ ЛОКАЛИЗАЦИЯ КОМПЬЮТЕРНЫХ СИСТЕМ И ТЕРМИНОЛОГИЯ

ПРЕИМУЩЕСТВА АВТОМАТИЗАЦИИ ТЕРМИНОГРАФИЧЕСКОЙ РАБОТЫ

А. Гурбанова, Институт Информационных Технологий НАНА Баку, Азербайджан, afruz1961@gmail.com

В статье рассматриваются понятия «лексикография» и «терминография», их история, теоретические и практические особенности. Отмечена необходимость разработки единых принципов и методов составления словарей, описаны необходимые для этого условия. Определены перспективные направления лексикографии и терминографии, отмечены преимущества их компьютеризации. Проанализированы процессы автоматизации работ по лексикографии и терминографии и отмечена необходимость создания терминологической базы данных, позволяющей накопление и быструю обработку больших объемов терминологической информации. Показаны возможности терминологической базы данных, созданной в рамках Национальной терминологической информационной системы Азербайджана.

Ключевые слова: *лексикография, терминография, терминологический словарь, компьютерная лексикография, компьютерная терминография, терминологическая база данных.*

ADVANTAGES OF TERMINOGRAPHIC AUTOMATION

A.Gurbanova, Institute of Information Technology of ANAS, Baku, Azerbaijan, afruz1961@gmail.com

The article reviews the concepts of "lexicography" and "terminography", their history, theoretical and practical features. The need to develop common principles and methods for compiling dictionaries is noted, and the necessary conditions for this are described. Prospective directions of lexicography and

terminography are determined, and the advantages of their computerization are noted. The processes of automation of works on lexicography and terminography are analyzed. The necessity of building terminological database enabling the accumulation and rapid processing of large volumes of terminological information is noted. The opportunities of terminological database established within the framework of the National Terminological Information System of Azerbaijan are shown.

Key words: *lexicography, terminography, terminological dictionary, computer lexicography, computer terminography, terminological database.*

Введение

На сегодняшний день информационные технологии являются неотъемлемой частью лингвистики, как и любой другой сферы профессиональной деятельности. Использование компьютеров и прикладных программ в лингвистических исследованиях, переводах и изучении языка сегодня имеет большое значение. Первым шагом к использованию информационных технологий в лингвистике был переход из печатной и устной формы в цифровую.

При определении соответствующих разделов лингвистики, связанных с использованием информационных технологий, необходимо обратить внимание на различия между теоретической и прикладной лингвистикой.

Теоретическая лингвистика – это раздел лингвистики, изучающий становление государственного языка, его историю, соответствие законам. Прикладная лингвистика, развивающаяся с 20-х годов прошлого столетия, является разделом лингвистической науки, специализирующейся в области разработки методов решения практических проблем, связанных с использованием языка.

Будучи разделом прикладной лингвистики, вычислительная лексикография имеет целью создание компьютерных словарей, разработку программного обеспечения для лингвистической базы данных и поддержки лексикографических работ. Основными направлениями традиционной и вычислительной лексикографии являются определение структуры словарей и места лексических статей, а также разработка принципов проектирования различных типов словарей (Щипицина Л.Ю., 2013).

Одним из перспективных направлений вычислительной лексикографии и прикладной лингвистики является работа над электронными терминологическими словарями и терминологическими базами.

Терминография, входящая в состав лексикографии, занимается разработкой специальных терминологических словарей. Терминография тесно связана с терминоведением, наукой об особой лексической единице языка. Соответственно, вычислительная терминография – это наука о разработке электронных терминологических словарей.

В данной работе рассматриваются перспективы развития лексикографии, связанные с автоматизацией работ терминографии, являющейся одной из её ветвей.

1. Лексикография и терминография

Искусство разработки словарей имеет давнюю историю, но как научная дисциплина с её концептуальным аппаратом и методами исследования лексикография была сформирована в двадцатом веке. Так, начиная с 70-х годов прошлого столетия лексикография, выйдя из сферы разработки словарей, перешла в отдельный научно-практический предмет, появилась теория работы над словарями. В научной литературе лексикография определяется как один из разделов лингвистики, специализируется в теоретической и практической разработке словарей, выполняет функции общественной важности: изучение языка, представление и нормализация языка, межязыковое и внутриязыковое общение, изучение языка с научной точки зрения (Бобунова М. А., 2009).

В истории мировой лексикографии известны такие выдающиеся лексикографы как Владимир Даль (Россия), Е. Гримм, К. Гримм и Г. Пауль (Германия), Э. Литтре и П. Ларусс (Франция), И. Малорни (Италия), В. Вартбург (Щвейцария), Х. Вебстер (США) (Комарова З. И., 2016).

Как было отмечено выше, современная лексикография традиционно делится на 2 части (Дамбуев И.А., 2011). К ним относятся:

- теоретическая лексикография, которая изучает методы разработки словарей и составные элементы словарей;
- практическая лексикография, которая исследует типологию словарей и осуществляет классификацию словарей по отдельным формам (исторические, фразеологические, этимологические и др.).

Начиная с 80-х годов прошлого столетия начал формироваться один из разделов лексикографии, а именно терминография. Следует отметить, что в работах учёных в научной литературе, связанной с проблемами лексикографии и терминографии, даётся неоднозначная интерпретация данных терминов.

Известный испанский лексикограф Хулио Касарес в своей книге «Введение в современную лексикографию» отмечает, что лексикография – это искусство разработки словарей. (Касарес Х**.,** 1958).

Терминография – это интегративный научно-практический предмет по истории, теории и практике терминологических словарей. В него включены методология, методы, методика и технология для оптимального проектирования, разработки и использования терминологических словарей, созданных для решения различных научных и практических вопросов (Комарова З. И., 2016).

Терминологические словари имеют немаловажное значение для всестороннего налаживания связей на международном и

внутригосударственном уровнях, а также для создания и обеспечения информационных служб и систем (Герд А. С., 2005).

По мнению российского лингвиста А. С. Герда, «терминография» и «лексикография» имеют одно и то же значение. Объект научно-технической (или терминологической) лексикографии состоит из практики и теории создания специальных (терминологических) словарей (Герд А. С., 1996).

Термин «терминография» выведен из словосочетания «терминологическая лексикография». Терминография занимает одно из приоритетных направлений терминоведения (Лейчик В. М., 2006).

Терминологическая лексикография или терминография – **это** наука о проектировании, разработки и использовании словарей специализированной лексики (Гринёв-Гриневич С. В., 1993).

По мнению российского филолога К.Я.Авербуха, функция создания терминологических словарей традиционной лексикографии по-прежнему **использует** свои основные правила, но материал, используемый разработчиками терминологических словарей, поставленные задачи и методы их решения привели к появлению нового научного направления – терминографии (Авербух К. Я., 2004).

В статье азербайджанского языковеда А.М.Гурбанова представлена главная цель и назначение лексикографии, заключающиеся в сборе, включении в систему слов из словарного состава какого-либо языка, а также фразеологических единиц, в объяснении их происхождения, значения и принадлежности (Гурбанов А.М., 2003).

Многие из проблем, изучаемых терминологами, возникают на практике в ходе разработки терминологических словарей, и решение этих проблем влияет на методы составления словарей. В то же время изучение любой сферы специальной лексики связано с терминографией, поэтому, результаты работ по обнаружению, исследованию и регулированию терминологии обычно составляются в форме словаря. Терминография тесно связана с терминоведением, поскольку в рамках терминоведения решаются такие проблемы, как многозначность терминов, принятие синонимов и омонимов, выбор их эквивалентов в других языках. В связи с этим многие терминоведы считают, что терминоведение является теоретической базой терминографии или же терминография является одним из разделов терминоведения (Гринев С.В., 1995).

Как известно, словари играют важную роль в жизни людей. Они**,** являясь инструментом для изучения родного и иностранного языка, способствуют расширению запаса знаний и укреплению существующих знаний. Сегодня они играют большую роль и в ускоренном развитии науки и техники. Следует отметить, что в последнее время число терминологических словарей стремительно растёт, что приводит и к увеличению числа других словарных форм.

В 1920-1952 гг. в Азербайджане количество терминологических словарей в 10 раз превышало количество других словарных форм. С 50-х годов прошлого века по сегодняшний день издано до 300 двуязычных, многоязычных и толковых словарей (Садыгова С., 2011).

Поскольку область знаний объективна, а термины и терминологические системы привязаны к конкретному языку, важным вопросом терминографии является стандартизация и унификация терминов, а также их однозначная интерпретация в различных языках.

Унификация терминологических систем основывается на терминологических стандартах. В настоящее время в мире существует более 20 000 стандартов, связанных с организацией терминологических систем. Помимо этого, существуют терминологические стандарты разных уровней – международного, государственного и даже уровня отдельной компании или фирмы. В связи с этим, вопрос об унификации терминологии и терминологических систем также должен быть неотъемлемой частью государственной и локальной языковой политики (Баранов А Н., 2007).

В соответствии со всем вышеизложенным, перспективными направлениями современной лексикографии и терминографии являются:

- проектирование, внедрение и эффективная методология использования различных типов словарей для разных целей, реализация методики и технологии;
- выявление новых проблем теории лексикографии и терминографии;
- утверждение новых научных парадигм, интегрированных с лексикографией и терминографией;
- совершенствование существующих видов и жанров словарей;
- создание новых типов словарей, основываясь на новые принципы и проблемы;
- информатизация и компьютеризация работы над словарями для более ускоренного создания словарей, а также упрощения работы лексикографа / терминографа.

2. Автоматизация лексикографии и терминографии

Известны преимущества компьютеризации лексикографии и одного из её разделов – терминографии: компьютер может быстро предоставить информацию о синонимах и антонимах слова, цитатном материале и.т.д.

Задачи, стоящие перед современной лингвистикой, ставят использование компьютерных технологий на первое место.

Здесь можно выделить нижеследующие направления:

- лингвистическое обеспечение различных типов информационных систем;
- машинный перевод;
- разработка систем, понимающих естественный язык;

- разработка системы использования информации, состоящей из звукового речевого сигнала и т.д.

Ещё в 70-х годах прошлого века были отмечены возможности применения компьютера в работе лексикографии (Марчук Ю.Н., 2007). Таким образом, компьютер может выстроить в алфавитном порядке лексические единицы, а также выполнять более тяжёлую работу, которая потребовала бы у лексикографов затраты большего времени и труда.

Современные вычислительные системы позволяют автоматизировать лексикографическую работу практически на каждом этапе – от выбора цитат до редактирования и печати словаря. Автоматизация такой работы, широкое использование компьютерного программного обеспечения увеличивает производительность труда лексикографов (терминографов). В результате формируется новое направление – лексикографическая информатика. Лексикографическая информатика включает в себя создание автоматических словарей, а также разработку программ поддержки лексикографической работы.

Основные аспекты лексикографической информатики нижеследующие:

- автоматическое извлечение различных словарей из текстов с помощью компьютерных инструментов;
- теоретические и практические аспекты составления компьютерных словарей для системы обработки естественного языка;
- создание и реализация машинной версии традиционных словарей.

Одной из актуальных проблем является включение популярных словарей и запросов в машинные носители и создание на их основе новых словарей.

В статье (Rundell M., Hanks P., Schryver G.M., 2015) описаны три разных метода создания и сбора лингвистических данных, применяемых для создания словаря: контент пользователя; модель вики; модель краудсорсинга. При этом отмечено, что все три метода — важный потенциал для лексикографии и терминографии.

Одним из ключевых вопросов терминографии является создание сложных автоматизированных терминологических словарей.

Таким образом, терминологическая информатика будет развиваться по двум основным направлениям:

- создание печатных словарей на основе компьютерной технологии;
- создание электронных словарей, действующих только в памяти компьютера или на магнитных носителях.

Известно, что в настоящее время осуществляется переход от печатных словарей к электронным. Именно за счёт электронных словарей произошло ускоренное развитие теории и практики лексикографии и терминографии, и в первое десятилетие XXI века уже существовало 600 типов электронных

словарей, доступных для 40 языков (Дубичинский В. В., 2009). Основными типами электронных словарей являются (Гринёв-Гриневич С. В., 2008):

1. информационно-поисковый тезаурус, который служит информационно-поисковой системе, используя при этом языки поиска информации;
2. база данных;
3. база терминологических данных;
4. база знаний;
5. база терминологических знаний;
6. машинные корпусы и др.

Повышение эффективности поиска связано с повышением качества информационно-поисковых тезаурусов. По этой причине были разработаны методы оптимизации информационно-поисковых тезаурусов, которые позволяют выбрать более оптимальный лексикон для тезауруса [Мамедова М.Г., Скороходько Э.Ф., 1985).

Некоторые исследователи полагают, что «…компьютеризация работы над словарями определяет будущий характер лексикографии. Благодаря компьютерной обработке лексикографических данных разработка словарей будет продолжаться стремительным шагом, возможно, XXI век назовут Золотым веком лексикографии» (Дубичинский В. В., 2009).

В статье (Kosem I., Gantar P., Logar N., Krek S., 2014) были представлены два различных лексикографических и терминографических исследовательских проекта. В обоих проектах использовался один и тот же метод автоматического выведения данных корпуса и были представлены аналогичные и отличительные особенности полученных данных. Даны перспективы автоматизации в лексикографии и терминографии.

3. Терминологическая база данных

Как было упомянуто выше, в цели уменьшения трудоемкости создания словарей в последнее время были разработаны автоматизированные методы. Автоматизация лексикографической работы привела к необходимости создания терминологической базой данных (ТБД). Современные компьютерные технологии позволяют обрабатывать и хранить в ТБД большое количество терминов из различных областей знаний.

Создание ТБД требует больших затрат, и замена его состава является с крупномасштабной работой. Поэтому очень важно предварительно изучить содержание терминологической информации ТБД. Необходимость организации структуры и автоматической обработки терминологической информации в больших количествах сделала эту область актуальной в терминографии.

ТБД включает и себе следующие функции:

- предоставление информационных и справочных услуг специалистам в различных областях знаний;
- обеспечение традиционного перевода научно-технической литературы;
- обеспечение системы машинного перевода;
- лингвистическое обеспечение автоматизированных информационных систем;
- обеспечение работы по терминологическому регулированию;
- подготовку и публикацию терминологических словарей;
- унификацию определённых терминов;
- подготовку научных отчётов по языку.

В настоящее время создаются ТБД по различным областям знаний и практической деятельности.

В Оттавском университете Канады разработали концептуальную основу Терминологической базы знаний (ТБЗ). Реализация концепции ТБЗ нацелена на облегчение процесса приобретения знаний (Meyer I., Skuce D., Bowker L., Eck K., 1992).

ТБД России есть во Всероссийском научно-исследовательском институте классификации, терминологии и информации по стандартизации и качеству (ВНИИКИ Госстандарта России) (www.vniiki.ru). Крупнейшая в мире ТБД, принадлежащая немецкой фирме "Siemens", содержит 2,5 миллиона терминологических записей на 8 языках. В Германии ТБД есть в Институте стандартизации и других руководящих органах. Большая международная ТБД была создана в Люксембурге в Бюро терминологий Европейской комиссии (www.iate.europa.eu). Созданная под названием «Eurodicautom» ТБД функционирует на нескольких языках. ТБД были созданы и используются во Франции, Канаде, Швеции, Бельгии и других странах мира. У Международной организации по стандартизации есть гигантская ТБД (www.iso.org/obp/ui).

Следует отметить, что в Азербайджане разработана концепция Национальной терминологической информационной системы (НТИС) (Rasim M. Alguliyev, Afruz M. Gurbanova, 2018). В рамках НТИС создан Национальный терминологический веб-портал с целью сбора терминологических словарей, разработанных в различных областях в единой информационной системе (www.terminologiya.az).

Таким образом, ТБД имеют большой потенциал развития, они стали инструментом для проведения серьезных научных исследований на основе автоматизированного хранилища, а также имеют большое практическое значение.

Вывод

В данной работе рассмотрен вопрос определения перспектив развития, касающихся автоматизации терминографических работ, и были получены следующие результаты:

- установлено, что основными задачами современной терминографии являются создание надёжной классификации специализированных словарей, анализ спроса на отдельные формы словарей, анализ способов повышения их качества, изучение методов отбора и организации терминологической информации в словарях и методов автоматизации терминологической деятельности.

- автоматизация терминографических работ даст возможность решить следующие проблемы:

- обеспечение отдельного подхода к разработке терминологических словарей:
- предоставление выбора оптимальной последовательности для определения ключевых характеристик при разработке словарей;
- предоставление возможности многочисленным специалистам, занимающимся разработкой словарей, для организации своей работы и объективного оценивания результатов.

- создание в Азербайджане ТБД способствует повышению эффективности терминотворчества и поможет решить нижеследующие задачи:

- моделирование терминологической системы азербайджанского языка;
- создание общетеоретических и общенаучных тезаурусов;
- исследование азербайджанской терминологии.

ЛИТЕРАТУРА:

1. Авербух К. Я. Общая теория термина, Москва: Издательство МГОУ, 2004, 252 с.
2. Alguliyev R.M., Gurbanova A.M. The Conceptual Foundations of National Terminological Information System, International Journal of Education and Management Engineering(IJEME), 2018, pp.31-49
3. Баранов А Н. Введение в прикладную лингвистику, Учебник; МГУ, Филол. фак., Изд. 3-е., Москва: Изд-во ЛКИ, 2007, 358 с.
4. Бобунова М. А. Русская лексикография XXI века, Учеб. Пособие, М., Флинта : Наука, 2009, 200 с.
5. Всероссийский научно-исследовательский институт классификации, терминологии и информации по стандартизации и качеству, http://www.vniiki.ru
6. Герд А. С. Прикладная лингвистика. СПб., СПбГУ, 2005, 268 с.
7. Герд А. С. Научно-техническая лексикография, Прикладное языкознание. СПб., Изд-во СПб, 1996, 287 с.
8. Гринёв-Гриневич С. В. Введение в терминоведение, М., Москва, Лицей, 1993, 230 с.
9. Гринев-Гриневич С.В. Введение в терминографию, М., Изд-во МПУ, 1995, 158 с.

10. Гринёв-Гриневич С. В. Терминография, Терминоведение, Учеб. Пособие, М., Академия, 2008, 304 с.
11. Гурбанов А.М. Современный азербайджанский литературный язык. I том. Баку, «Нурлан», 2003, 450 с.
12. Дамбуев И.А. Современная лексикография: статус и направления развития, Бурятия, Россия // Вестник Бурятского Государственного Университета, 2011, стр. 16-21.
13. Дубичинский В. В. Терминография и стандарты, Лексикография русского языка : учеб. Пособие, М., Флинта: Наука, 2009. 432 с.
14. IATE — European Terminology Database, http://www.iate.europa.eu
15. International Organization for Standardization, Online Browsing Platform, http://www.iso.org/obp/ui
16. Касарес Х**.** Введение в современную лексикографию, Изд-во иностранной лиературы, 1958, 354 с.
17. Kosem I., Gantar P., Logar N., Krek S. Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies, Proceedings of the XVI EURALEX International Congress: The User in Focus, Lexicography and Language Technologies, 2014, pp. 355-364.
18. Комарова З. И. Когда и почему возникла терминография как научно-прикладная дисциплина? // Актуальные проблемы германистики, романистики и русистики: сб. науч. Работ, Екатеринбург, Изд-во УрГПУ, 2016, ч. 3, с. 90—107.
19. Лейчик В. М. Терминоведение (Предмет, методы, структура), М., КомКнига, 2006, 256 с.
20. Мамедова М.Г., Скороходько Э.Ф. Метод оптимизации информационно-поискового тезауруса // НТИ, сер. 2 – Информационные процессы и системы, 1985, стр. 12–17 (Москва).
21. Марчук Ю.Н. Компьютерная лингвистика: учеб. Пособие, М., АСТ: Восток-Запад, 2007, 317 с.
22. Meyer I., Skuce D., Bowker L., Eck K. Towards a New Generation of Terminological Resources: An Experiment in Building a Terminological Knowledge Base, 1992, COLING '92 Proceedings of the 14th conference on Computational linguistics, Vol. 3, p. 956-960, http://acl.ldc.upenn.edu/C/C92/C92-3146.pdf
23. Rundell M., Hanks P., Schryver G.M. Crowdsourcing, wikis, and user-generated content, and their potential value for dictionaries // International Handbook of Modern Lexis and Lexicography, 2015, pp. 1-16.
24. Садыгова С. Терминология азербайджанского языка. Баку, «Элм», 2011, 380 с.
25. Терминологическая Комиссия при Кабинете Министров Азербайджанской Республики, http://www.terminologiya.az
26. Щипицина Л.Ю. Информационные технологии в лингвистике : учеб. пособие, М., ФЛИНТА: Наука, 2013, 128 с.

ТРЕХЪЯЗЫЧНЫЙ ТЕРМИНОЛОГИЧЕСКИЙ СЛОВАРЬ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ: ЦЕЛИ, МЕТОДОЛОГИЯ, ПЕРСПЕКТИВЫ

Д. А. Киселев, О. Я. Юсупов, Самаркандский государственный институт иностранных языков, Самарканд, Узбекистан, dkiselyov@umail.uz, otabekuz10@mail.ru

В статье излагаются цели и методические основы создания трехъязычного (англо-узбекско-русского) терминологического словаря по компьютерной лингвистике, а также перспективы его использования. Работа над составлением словаря ведется группой преподавателей СамГИИЯ в рамках проекта ERASMUS+ CLASS, направленного на создание междисциплинарной магистерской программы по компьютерной лингвистике для вузов Узбекистана.

Ключевые слова: *компьютерная лингвистика, тюркские языки, терминологический словарь, дефиниция, толкование.*

TRILINGUAL DICTIONARY OF COMPUTATIONAL LINGUISTICS' TERMINOLOGY: AIMS, METHODOLOGY, AND PERSPECTIVES

D. A. Kiselev, O.Ya.Yusupov, Samarkand state institute of foreign languages, Samarkand, Uzbekistan, dkiselyov@umail.uz, otabekuz10@mail.ru

The article presents aims and methodological basis of compiling of trilingual (English-Uzbek-Russian) dictionary of computational linguistics' terminology as well as perspectives of its use. The work on the dictionary is being carried out by the workgroup of Samarkand State Institute of Foreign Languages in the framework of ERASMUS+ CLASS project which aims the design of an interdisciplinary MA program on Computational linguistics for Uzbek universities.

Key words: *computational linguistics, Turkic languages, dictionary of terminology, definition, gloss.*

Компьютерная лингвистика представляет собой одну из наиболее динамично развивающихся и наиболее востребованных отраслей языкознания на современном этапе. Современная вычислительная техника предоставляет возможность автоматической обработки в сжатые сроки значительных объемов информации, например текстов, что

выводит на качественно новый уровень достоверность статистических данных, лексикографическую работу и многое другое. Однако существующий интерфейс – технические средства и способы взаимодействия человека и компьютера, например, клавиатура, сканер и т.д. – потенциально уступают такому гибкому инструменту коммуникации как естественный язык. Как следствие, организация непосредственного общения человека и компьютера требует глубокого понимания закономерностей и особенностей использования естественного языка в процессе общений людей между собой [Боярский 2013: 3-4].

Зародившись в середине XX в., к настоящему времени компьютерная лингвистика эволюционировала с самостоятельную отрасль научного знания, имеющую высокий прикладной потенциал. Изучение структуры естественных языков и принципов его функционирования, с одной стороны, и динамичное развитие вычислительной техники, с другой стороны, сделало возможным не только решение широкого ряда специфических лингвистических задач, в частности создание национальных корпусов языка, но и выход на новый уровень взаимодействия человека и компьютера, что находит свое отражение в создании и постоянном совершенствовании различных форм искусственного интеллекта, в том числе созданных на основе нейронных сетей голосовых помощников (Google Assistant, Amazon Alexa, Яндекс Алиса) и многое другое, делающее возможным непосредственное двустороннее общение человека с компьютером.

Степень развития компьютерной лингвистики в различных регионах мира варьируется, как и степень вовлеченности различных языков. Так, большинство программных продуктов в настоящее время ориентировано на английский язык. Таким образом, развитие компьютерной лингвистики на базе национальных языков предоставляет широкий исследовательский простор и формирует устойчивый тренд развития научной дисциплины.

Потребность в развитии компьютерной лингвистики в Узбекистане на материале узбекского языка крайне высока по ряду взаимообусловленных причин. Так, в настоящее время вузы Узбекистана не обеспечивают подготовку достаточного количества квалифицированных специалистов по компьютерной лингвистике. В результате, теоретические основы автоматической обработки естественных языков (АОЕЯ) на материале узбекского языка оказываются недостаточно разработанными, как следствие, это негативно влияет на решение прикладных задач, в частности создание национального корпуса языка, совершенствование алгоритмов машинного перевода и т.д. Однако в целом, ситуацию нельзя

охарактеризовать как критическую: преподавание компьютерной лингвистики в качестве учебной дисциплины ведется в ряде вузов Узбекистана, в частности в Узбекском государственном университете узбекского языка и литературы им.А.Навои, Ургенчском государственном университете; осуществляется реализация некоторых локальных задач прикладного характера и т.д.

Существенную помощь в решении вопроса по подготовке в вузах Узбекистана квалифицированных кадров в сфере компьютерной лингвистики может оказать проект ERASMUS+ CLASS (Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities), имеющий целью разработку и внедрение в вузах центральноазиатских стран, в частности Узбекистана и Казахстана, междисциплинарной магистерской программы по компьютерной лингвистике. Залогом достижения указанной цели является участие в консорциуме проекта целого ряда европейских партнеров, имеющих богатый опыт в преподавании данной дисциплины, в частности университеты Испании, Португалии, Греции и Польши, а также некоторых центральноазиатских вузов, имеющих значительный практический опыт в данной сфере, в частности Евразийского национального университета им. Л.Н.Гумилева (Казахстан).

Критический анализ текущей практики преподавания компьютерной лингвистики в вузах различных стран мира, проведенный экспертами из вузов-участников проекта и последовавший обмен мнениями участников проекта позволили определить оптимальную структуру и содержание будущей образовательной программы для магистратуры. Вместе с тем был проанализирован уровень преподавания компьютерной лингвистики в вузах Казахстана и Узбекистана с целью выявления положительного опыта и максимальной адаптации создаваемой учебной программы к локальным и региональным потребностям. Так, в частности, было определено, что создаваемая магистерская программа будет носить междисциплинарный характер в плане содержания учебных курсов и образовательного опыта студентов, которые будут обучаться по данной программе. Это позволит привлечь к обучению студентов, прошедших обучение в бакалавриате, как по филологическому направлению образования, так и по направлению точных наук. Последнее обстоятельство имеет ряд преимуществ, но также содержит и определенные риски. В частности, необходимым становится прохождение всеми обучающимися по программе общего курса, который позволит, с одной стороны, повысить уровень филологических знаний студентов, получивших образование в сфере точных наук, и, с другой стороны, расширить знания о вычислительной технике, программировании и т.п. Таким образом, образовательный опыт

студентов должен быть приведен к некоему общему знаменателю. Условно говоря: филология, ее методология и терминология должны быть понятны программистам, а принципы программирования и соответствующая терминология должны быть понятны филологам.

Решение этой задачи имеет несколько аспектов, одним из которых может стать составление терминологического словаря, включающего термины, имеющие хождение в сфере компьютерной лингвистики, программирования и языкознания. Принимая во внимание потребность будущих студентов магистратуры по этой программе, а также, вероятно, и преподавателей в единстве тематической терминологии, рабочая группа Самаркандского государственного института иностранных языков определила своей задачей создание трехъязычного (англо-узбекско-русского) терминологического словаря по компьютерной лингвистике. Выбор трех языков обусловлен рядом объективных факторов.

Так, английский язык можно считать базовым средством профессионального и научного общения в сфере вычислительной техники, программирования и компьютерной лингвистики. На английском языке написано большинство научной и учебной литературы по компьютерной лингвистике. Хотя зачастую сами термины этимологически восходят к древнегреческому и латинскому языкам, они возникли и получили устойчивое толкование изначально в англоязычной профессиональной среде. Кроме того, предполагается, что на начальном этапе обучения по программе компьютерной лингвистики, английский будет не только полноправным, наряду с узбекским и, возможно, русским, языком обучения, но также послужит материалом для исследования и эксперимента, т.е. на примере английского будут изучаться структура и принципы функционирования языка, создаваться учебные алгоритмы автоматической обработки языка и многое другое.

Учитывая эти обстоятельства, при составлении англоязычные термины можно принять в качестве базовых, корневых – именно они определяют порядок расположения терминов в словаре. Выбор узбекского языка более чем закономерен – он является целевым языком, т.е. программные продукты будут создаваться для узбекского языка. При этом, и студенты, обучающие по программе должны владеть единой терминологической базой. Вместе с тем, вследствие того, что компьютерная лингвистика в Узбекистане находится на этапе становления, создание терминологической базы на узбекском языке становится приоритетной задачей. Включение терминологии и дефиниций на русском языке определяется тем фактом, что русский язык широко распространен в академической и научной среде Узбекистана и, как следствие, наличие толкования терминов на русском языке является

дополнительной гарантией их широкого и адекватного использования, как в образовательном процессе, так и научной среде.

Исходя из того, что составление терминологического словаря ведется рабочей группой проекта, крайне важной является соблюдение единых методологических принципов [Ахманова 1969: 11-12; Dubois 2015: 7].

Это стандартизирует способ представления терминов и их толкования, а также в перспективе облегчит пользование словарем [Ducrot 2016: 4-5].

Редактирование толкований требует значительных усилий, т.к. термины, включенные в состав словаря, тематически разнообразны – они относятся и к языкознанию, и к компьютерной лингвистике, и к кибернетике в целом.

Можно с уверенностью утверждать, что разрабатываемый трехъязычный терминологический словарь будет востребован как студентами, которые будут получать образование по магистерской программе по компьютерной лингвистике, так и преподавателями, как единая многоязычная справочная и учебная литература.

ЛИТЕРАТУРА:

1. Ахманова О.С. Словарь лингвистических терминов. Изд. 2ое. -М.: Сов. Энциклопедия, 1968. – 608 с.
2. Боярский К.К. Введение в компьютерную лингвистику. Учебное пособие. – СПб: НИУ ИТМО, 2013. – 72 с.
3. Dubois J., Giacomo M. et alii. Le dictionnaire de linguistique et des sciences du language. Edition Larousse, 2015. — 1012 p.
4. Ducrot O., Schaeffer J.-M. Nouveau dictionnaire encyclopédique des sciences du language. Paris: Edition Points, 2016. – 872 p.

КОМПЬЮТЕРИЗИРОВАННЫЕ РЕШЕНИЯ В СИСТЕМЕ ЯЗЫКОВЫХ ПРОЦЕССОВ МИРОВОЙ ЦИВИЛИИЗАЦИ

Б. Р. Каримов [1], Ш.Ш Муталов [2], [1]Отдел научных исследований Комитета по межнациональным отношениям и дружественным связям с зарубежными странами при Кабинете Министров Республики Узбекистан, Ташкент, Узбекистан, karimov.bahtiyor@yahoo.com
[2]Научный исследователь, переводчик, Ташкент, Узбекистан, shahahmad.mutal@gmail.com

Посредством использования методов математической и компьютерной лингвистики предлагается создать усредненные языки на основе усреднения лексических фондов, языковых правил и норм генеалогически родственных языков. Предлагается идея создания среднемирового языка посредством усреднения усредненных и изолированных языков. Рассмотрены социальные, культурные, лингвистические, информационные, коммуникативные проблемы усредненных языков и пути использования мультимедийных технологий в этой сфере.

***Ключевые слова**: усредненные языки; языковая идентичность, этнолингвопанизм, математическая и компьютерная лингвистика, мультимедийные и информационные технологии.*

COMPUTERIZED SOLUTIONS IN THE SYSTEM OF LANGUAGE PROCESSES OF WORLD CIVILIZATION

B.R.Karimov[1], Sh.Sh.Xmutalov[2], [1]Research Department of the Committee on Interethnic Relations and Friendly Relations with [2] Foreign Countries under the Cabinet of Ministers of the Republic of Uzbekistan Tashkent, Uzbekistan, karimov.bahtiyor@yahoo.com;
[2]Научный исследователь, переводчик, Tashkent, Uzbekistan, shahahmad.mutal@gmail.com

It is proposed to create averaged languages by using the methods of mathematical and computer linguistics based on averaging vocabularies and grammar rules and norms of genealogically cognate languages has been proposed. The idea of creating averaged world language is proposed by means of averaging the averaged and isolated languages. Social, cultural, linguistic,

information, communication problems of averaged languages and ways of using multimedia technologies in the area have been investigated.

Key words: *averaged languages; language identity; ethnolingvopanism; mathematical and cumputer linguistics, multimedia and information technologies.*

Введение

При обмене информацией обменивающиеся ею стороны должны понимать язык, на котором закодирована информация. Язык социальной коммуникации, то есть устный язык (звуковая кодировка информации), письменный язык (текстовая кодировка информации), а также частично и знаковый язык, выработанный для сообщества немых, лежат и в современную эпоху в основе информационного обмена между людьми. Без знания и понимания языка социальной коммуникации в социуме, в котором задействованы мультимедийные технологии и информационные технологии коммуникации эти технологии теряют свою эффективность, так как без языка теряется семиотическая и семантическая целостность системы кодов социальной коммуникации. Поэтому в большинстве исследований мультимедийных технологий и информационных технологий исходят из презумпции знания языка обмена информацией. Однако реальная ситуация в системе мировой цивилизации в корне отличается от этой идеализированной модели, которая таким образом оказывается применимой лишь к социальной группе знающей определенный подразумеваемый в презумпции общий язык. Для приближения к адекватному познанию реальности необходимо выйти за рамки этой идеализированной модели и создать более широкую модель, охватывающую её как частный случай.

Языковая ситуация в мире изучается во многих исследованиях, которые показывают сложность, противоречивость, конфликтность, драматизм, трагичность отношений языковых групп в мировой цивилизации. Хотя мировая цивилизация признаёт, что каждый язык является общечеловеческой ценностью, исследования показывают, что каждые две недели погибает один из живых языков, превращаясь в мертвый язык. Языковая идентичность тесно связана с национальной и этнической идентичностью, поэтому языковые процессы взаимосвязаны с национально-этническими процессами.

Современный ход языкового развития мировой цивилизации разворачивается на основе процесса создания национальных государств, в которых язык так называемой государствообразующей нации признается государственным языком, а языки других наций на этой территории, в большинстве случаев, приобретают второстепенное положение. Так как большинству наций не удается создать такое свое национальное государство, то язык этих наций оказывается во второстепенном статусе.

Такие нации составляют большинство наций мира, а их языки – большинство языков мира.

1. Постановка проблемы

В статье предложен иной путь решения этой фундаментальной языковой проблемы социальной коммуникации в современной мировой информационной цивилизации, основанный на рациональной научной регламентации информационного обмена и языковых процессов основанный на методах математической и компьютерной лингвистики.

Несмотря на некоторые особенности процесса формирования наций, большинство существующих наций развились из родственных племен и общенациональный язык, литературный язык нации развивался на основе племенных диалектов в результате их взаимодействия. С точки зрения генезиса кажется естественным, что зародыши наций находятся среди родственных племен, населявших соседствующие территории и имевших более тесные коммуникации. Кроме того, чем ближе были их языки, тем легче была коммуникация.

Согласно ойкуменической теории нации [1] такие этнические общности как род, племя, народность и нация следует рассматривать как исторические стадии развития общего родового феномена этнация. Этнация формируется и развивается вместе со своей ойкуменой [1], то есть с системой этнаций с которыми она взаимодействует, поддерживает коммуникацию.

Человечество в будущем будет единым целым. Стадии развития, когда родственные нации образуют общности более высокого уровня, представляются неизбежными. Но следует иметь в виду, что могут существовать неверные пути консолидации наций, такие как имевшие место в прошлом попытки сформировать единые народы. Проводя политику консолидации наций, следует принимать во внимание, что этническое самосознание является общечеловеческой ценностью.

Этнический лингвистический панизм, сокращенно этнолингвопанизм [1] является применением философской концепции потенциального единства к определенной этнической группе. Близость определенной группы народов и сознание того, что люди этой группы могут понимать друг друга в определенной мере, даже если каждый из них знает только свой язык, является основой любого специфического вида этнолингвопанизма. Подобие процесса коммуникации между людьми, говорящими на диалектах одного языка, и представителями вышеупомянутых народов является гносеологическим основанием возникновения соответствующего панизма. Если есть возможность, то этнолингвопанизм использует и такие характеристики, как близость культуры, религии, литературы, происхождения, исторических судеб и так

далее этносов, близость чьих языков служила основой при обосновании соответствующего этнолингвопанизма.

В настоящее время единство человечества необходимо для его выживания, целесообразно сознательно искать пути для формирования и сохранения такого единства. Объединение близких наций в федерации, конфедерации или унитарные государства согласно их волеизъявлению является одним из таких путей.

В многонациональных государствах возникает языковая проблема. В таких государствах должен быть язык, служащий в качестве общего языка общения, называемый языком-посредником. В мире существует много языков-посредников. Модно рассматривать английской язык как мировой язык-посредник, и в мире существуют много регионов с их собственными языками-посредниками. Например, суахили является таким языком в Восточной Африке, хинди – в Индостане, испанский и португальский в – Латинской Америке и т. д. Но национальный язык в качестве языка-посредника имеет, по меньшей мере, два неприемлемых свойства.

Во-первых, с точки зрения справедливости в духовной жизни, т.к. носители других языков будут чувствовать себя ущемленными в своих правах применения их родных языков. Во-вторых, с экономической точки зрения, все народы, кроме носителей языка-посредника, будут отставать на годы из-за необходимости изучать язык-посредник наряду со своими родными языками для получения мировой информации. Таким образом, можно прийти к заключению, что ни один национальный язык не пригоден для роли языка-посредника.

Региональным языком-посредником должен быть язык, в равной мере близкий к языкам народов региона, где предполагается, что он будет языком общения. Язык всемирного общения должен отвечать таким же требованиям. По нашему мнению, региональные языки межнационального общения должны быть созданы на базе групп родственных языков согласно специальному методу, обеспечивающему его нейтральность, а также и оптимальную степень понятности как можно большему числу людей среди носителей этих языков.

Ускорение темпов социальных процессов, формирование системы глобальных проблем человечества приводит к необходимости перехода к ноосфере, в частности, к моделированию и управлению информационными и языковыми процессами. Метод создания языка-посредника должен соответствовать объективному механизму процесса формирования койне в ходе общения носителей родственных языков, имитационно моделировать его. Чем у большего числа: 1) родственных языков и 2) индивидов носителей этих языков встречается определенная форма при реализации некоторого языкового феномена, тем больше вероятность включения этой нормы в систему норм койне.

Механизм, или метод, создания языка-посредника важен, так как в осознании своей этнической принадлежности личностями, считающими своим родным языком родственные языки, важную (хотя и не решающую) роль играет степень близости этих языков, осознание личностью своих языков как самостоятельных языков или как диалектов одного языка. В системах диалектов различных языков мера близости диалектов, приводящая к осознанию носителями диалектов принадлежности к одному языку, существенно различается.

Для лексики необходимо выбрать определенной длины верхнюю часть частотного словаря одного из языков. Этот язык служит базовым. Значения полисемантических слов, которые помечены в словаре основного языка цифрами, считаются отдельными лексемами. Выбранные слова в списке нумеруются. Для каждой лексемы упорядоченного списка должны быть найдены эквиваленты в языках или диалектах, на базе которых строится усредненный язык.

2. Концепция решения проблемы

Метод создания языка-посредника должен соответствовать объективному механизму процесса формирования койне в ходе общения носителей родственных языков, имитационно моделировать его. Чем у большего числа: 1) родственных языков и 2) индивидов носителей этих языков встречается определенная форма при реализации некоторого языкового феномена, тем больше вероятность включения этой нормы в систему норм койне.

3. Реализация концепции

Сутью предложенного метода [2, 3, 4, 5] является выявление общего фонда для группы диалектов или языков посредством специальной математической процедуры и кодификация данного общего фонда в качестве стандарта, т. е. как единицы соответствующего уровня создаваемого языка. Процедура следующая: единицам уровней языка сопоставляются векторы. Как правило, единицы уровней имеют варианты: различные значения слов, позиционные варианты фонем и морфем, различные способы выражения синтаксических отношений. Векторное пространство, сопоставленное единицам уровня, построено таким образом, что главное значение единицы сопоставлено со значением 1, другие варианты соответствуют числам, меньше 1, согласно формуле

$$X^{i}_{\mathrm{n},\alpha} = 1 - \frac{i-1}{2\mathrm{s}}, \qquad (1)$$

где "n" – номер единицы в списке единиц; «α» – определенный язык; «i» – номер варианта единицы уровня с номером "n"; "s" – число вариантов уровня с номером "n". Эти числа образуют набор компонентов вектора. Мы назвали этот метод методом усреднения, а языки, построенные таким способом, – усредненными.

Следующим шагом является упорядочение корней в соответствии со значениями данной функции.

Затем таблицу 1 с синонимическими рядами нужно преобразовать в таблицу 2, где однокоренные слова-эквиваленты расположены на одной строке.

Для однокоренных слов в таблице 2 вводится функция приоритетности синонима $F_n^j(\vec{x})$ по формуле

$$F_n^j(\vec{x}) = \sum_{\alpha=1}^{A}\left(1+\frac{K_\alpha}{\overline{K}}\right)\mathrm{X}_{\mathrm{n},\alpha}^{j}, \qquad (2)$$

где A – полное число усредняемых языков, K_α – численность носителей языка α, $\overline{K}$ – среднее арифметическое носителей языка одной группы родственных языков, которое вычисляется как частное от деления общего числа носителей всех языков группы на число языков в группе, j означает порядковый номер однокоренных слов в таблице 2.

Роль коэффициента $1+\frac{K_\alpha}{\overline{K}}$ – это учет численности носителей языка в группе, для которой строится усредненный язык.

Корни, помещенные на начальных первых местах в Таблице 3, в которой слова-синонимы расположены в порядке убывания функции приоритетности синонима рекомендуются для включения в словарь усредненного языка в качестве основных лексических единиц. Другие единицы могут формировать запас слов-синонимов для обогащения словаря усредненного языка при необходимости.

Вся работа может быть выполнена при помощи компьютеров. Преимущества предлагаемого решения мировой языковой проблемы заключаются в следующем:

1) так как усреднение проводится на базе этнических языков (диалектов), носители данных языков (или диалектов), на основе которых строится усредненный язык, будут понимать его в определенной мере без предварительного изучения;

2) не принадлежа ни одной этнической группе, усредненный язык не дает преимущества ни одной из них, так что он не будет способствовать национальной розни на основе языковой политики;

3) усредненный язык устраняет некоторый произвол в выборе одного из местных языков в качестве официального государственного языка, так же как межэтнические конфликты, связанные с этим произволом;

4) усредненный язык позволяет не вводить язык бывших колонизаторов в качестве единственного официального государственного языка, ослабляет зависимость от бывшей метрополии в сфере культуры и образования;

5) многие народы, говорящие на одном языке или родственных языках разделены государственными границами. Таким образом, усредненный язык, построенный предложенным методом, мог бы играть для них роль макропосредника.

Проблемы этнической консолидации различных диалектных групп в основном связаны со степенью их близости. Степень близости, когда носители диалекта считают себя говорящими на одном языке, сильно различается, т. е. в некоторых случаях носители не чувствуют себя относящимися к одному и тому же языку, хотя существует объективная близость диалектов, не осознаваемая носителями. Для объективной, научно-обоснованной оценки этой близости языков и диалектов целесообразно использовать метод количественной оценки их близости [6].

Здесь можно обратить внимание на еще одно преимущество усредненных языков – они могут оказаться спасением для исчезающих языков, многие их которых являются родственными языками малочисленных народов и диалектами.

В связи с созданием усредненных нейтральных языков для родственных языков возникает вопрос о перспективах развития коммуникации между носителями неродственных языков в мировом масштабе, глобальной коммуникации всех народов.

Целесообразно создание среднемирового языка посредством усреднения в многообразии усредненных языков и изолированных языков на основе ностратической (борейской) концепции, концепции языковых универсалий и статистических методов усреднения языковых феноменов [1, 2, 3, 4, 5]. Создаваемый таким путем всемирный вспомогательный язык межкультурного, межнационального общения, накопления мировой информации и глобального обучения способствовал бы формированию единого мирового информационного пространства, решению многих глобальных проблем мировой цивилизации и духовному взаимообогащению всех локальных цивилизаций и народов.

В процессе формирования мировой информационной цивилизации для каждого языка целесообразно создание компьютерных программ, которые преобразуют тексты на одном алфавите в тексты на другом алфавите. Целесообразно создание системы компьютерных программ для перевода с одного языка на другой. При этом перевод текста на усредненный язык определенной группы языков мог бы служить основным этапом для последующего перевода на другие генеалогически родственные ему языки [4, 5]. Необходимо увеличить информационные и мультимедиа ресурсы в Интернет на национальных языках и на усредненных языках. В частности, целесообразно развитие Википедии на национальных и усредненных языках. Осуществление этих предложений способствовало бы развитию как

каждой из локальных цивилизаций, так и системы мировой информационной цивилизации в целом.

Предлагаемые преобразования соответствуют тенденциям развития и расширяют горизонты развития межкультурной коммуникации в процессе формирования мировой информационной цивилизации в XXI веке.

Заключение

Мы полагаем, что всемирный вспомогательный язык может быть создан тем же путем, что и усредненные языки для групп родственных языков, и он будет следующим шагом в нейтрализации языков. Создание всемирного вспомогательного языка важно для решения проблемы информационных барьеров, так же как и необходимости преодоления информационного колониализма, и необходимости достижения равенства народов в условиях, когда существует клуб так называемых мировых языков, и стремление установить новый информационный порядок. Эта актуальная глобальная проблема может быть решена группой специалистов под эгидой ООН. При создании искусственного всемирного языка рекомендуется принять во внимание факты всех языков, и в связи с этим использование среднемирового языка создаваемого методом усреднения имеет то преимущество, что обеспечивается нейтральность языка и равноправие языков народов мира [2, 3, 4, 5].

Развитие мультимедийных технологий обеспечивает их применение во все более широких областях человеческой деятельности. В частности, в обмене информацией и образовании (дистанционное образование, международное, непрерывное образование). Эти сферы тесно связаны с употреблением языка, язык, в свою очередь, связан с самосознанием и национальной самоидентификацией. Для обеспечения равноправия наций и локальных цивилизаций в языковом аспекте целесообразно развивать усредненные языки, создать методом усреднения среднемировой язык и внедрять их в сферы обмена информацией и образования.

ЛИТЕРАТУРА:

1. Karimov B.R. The oikumenic concept of the nation and development of languages. Ойкуменическая концепция нации и развитие языков. Qarshi, 2003.

2. Каримов Б.Р., Муталов Ш.Ш. К вопросу о языковой политике в многоязычных развивающихся странах // Тезисы докладов республиканской научно-теоретической конференции молодых ученых и специалистов по общественным наукам. Ташкент, 2-3 марта 1982 г., секция 1. Ташкент, 1982. С.94-95. (Karimov B.R., Mutalov Sh.Sh. K voprosu o yazykovoy politike v mnogoyazychnykh razvivayushchikhsya stranakh. In: Tezisy dokladov respublikanskoy nauchno-teoreticheskoy konferentsii molodykh

uchyonykh i spetsialistov po obshchestvennym naukam. Tashkent, 2-3 marta 1982 g., sektsiya 1. Tashkent, 1982, s. 94-95).
3. Каримов Б.Р., Муталов Ш.Ш. Уртатурк тили. Тошкент, 1992. (Karimov B.R., Mutalov Sh.Sh. Ortaturk tili. Toshkent, 1992).
4. Karimov B.R., Mutalov Sh.Sh. Averaged languages: an attempt to solve the world language problem. Tashkent: Fan, 1993. (второе издание в 2008 г.). (vtoroye izdaniye v 2008 g.)
5. Каримов Б.Р., Муталов Ш.Ш. Усредненные языки: попытка решения мировой языковой проблемы. Т.: Фан, 2008. (Karimov B.R., Mutalov Sh.Sh. Usrednyonnye yazyki: popytka resheniya mirovoy yazykovoy problemy. T.: Fan, 2008)
6. Каримов Б.Р., Муталов Ш.Ш. О количественной оценке синхронической близости родственных диалектов и языков // Тюркское языкознание. Материалы международной конференции. Ташкент, 1985. (Тезисы этой конференции были опубликованы в 1980 году) (Karimov B.R., Mutalov Sh.Sh. O kolichestvennoy otsenke sinkhronicheskoy blizosti rodstvennykh dialektov i yazykov. In: Tyurkskoye yazykoznaniye. Materialy mezhdunarodnoy konferentsii. Tashkent, 1985) (Tezisy etoy konferentsii byli opublikovany v 1980 godu).

PROTECTION OF AZERBAIJANI LANGUAGE IN E-GOVERNMENT AND DEVELOPMENT PROSPECTS

R.Alguliyev, F. Yusifov, A. Gurbanova,
Institute of Information Technology of ANAS, Baku, Azerbaijan,
r.alguliev@gmail.com; farhadyusifov@gmail.com; afruz1961@gmail.com

The protection of language diversity has become one of the topical issues in the rapidly globalizing modern world influenced by information technology. The article studies the protection of Azerbaijani language in e-government. The approaches to the impacts of globalization on linguacultural space and linguistic processes are explored. The protection of linguistic diversity and the application opportunities of linguistic technologies in e-government are analyzed. The approaches to the protection of Azerbaijani language in e-Azerbaijan segment of the global information space are provided. A conceptual model that provides effective mechanisms for the application of linguistic technologies is proposed. In general, there are various ideas and approaches related to language protection on the Internet. There are many factors endangering language in the globalization process, and it is not possible to maintain the language at the expense of controlling each of these factors. In this regard, the Internet can also be used to raise awareness about the language loss and language protection. The application of cloud-based linguistic technologies enables the provision of different services over the Internet (translation, cataloging, data storage, availability etc.). They can also be used to preserve the integrity of spoken language. Furthermore, many of these technologies used for sound and speech recognition can be applied to protect spoken languages. The use of linguistic technologies in e-Azerbaijan segment of the unique information space will contribute to the protection and development of Azerbaijani language by providing variety of e-services.

Key words: *E-government; Globalization; Linguistic diversity; Linguistic technologies; Language loss; Language protection.*

СОХРАНЕНИЕ АЗЕРБАЙДЖАНСКОГО ЯЗЫКА В РАМКАХ ЭЛЕКТРОННОГО ПРАВИТЕЛЬСТВА И ПЕРСПЕКТИВЫ ЕГО РАЗВИТИЯ

Р.Алгулиев, Ф. Юсифов А. Гурбанова, Институт информационных технологий
Национальной Академии наук Азербайджана, Баку, Азербайджан
r.alguliev@gmail.com; farhadyusifov@gmail.com afruz1961@gmail.com
farhadyusifov@gmail.com, ph.d@iit.science.az

Сохранение языкового разнообразия стало одной из актуальных проблем в процессе быстрой глобализации современного мира, подверженного влиянию информационных технологий. В статье рассматриваются вопросы поддержки и сохранения азербайджанского языка в электронном правительстве. Исследуются подходы к описанию воздействия глобализации на лингвокультурное пространство и языковые процессы. Анализируется лингвистическое разнообразие и возможности применения лингвистических технологий в электронном правительстве. Предложены подходы к защите азербайджанского языка в азербайджанском электронном сегменте глобального информационного пространства. Предложена концептуальная модель, обеспечивающая эффективные механизмы применения лингвистических технологий. В целом, в интернете существуют различные идеи и подходы, связанные с сохранением языков. Есть много факторов, угрожающих языку в процессе глобализации, и невозможно поддерживать язык за счет контроля каждого из этих факторов. В этом отношении интернет также может быть использован для повышения осведомленности об исчезновении языка и о его защите. Применение облачных лингвистических технологий позволяет предоставлять различные услуги через интернет (перевод, каталогизация, хранение данных, доступность и т. д.). Они также могут быть использованы для сохранения цельности разговорного языка. Кроме того, многие из этих технологий, используемых для распознавания звука и речи, могут применяться для защиты разговорных языков. Использование лингвистических технологий в электронном азербайджанском сегменте уникального информационного пространства будет способствовать защите и развитию азербайджанского языка путем предоставления разнообразных электронных услуг.

Ключевые слова: *электронное правительство; глобализация; языковое разнообразие; лингвистические технологии; утрата языка; сохранение языков.*

1. Introduction

Nowadays, technological revolution that has emerged around ICT. Protection of language diversity has become one of the topical issues in the modern world, which is rapidly globalizing with the impact of information technology (Maurais, & Morris, 2003; Smetanina-Baldvin & Maslova, 2009). Achievements in the development of linguistic technologies will gradually lead to greater access of people to information and services in their own languages. The globalization process has a significant impact not only on technological and economic fields, but also on the social, political and cultural spheres. This trend indicates that the globalization grounds on the English-American model of society and its economy, politics, culture. Obviously, this model is closely

linked to English language, which plays the role of world language. The globalization process, as well as the changes in the language sphere require serious measures in the field of protection and development of Azerbaijani language in the electronic environment (Smetanina-Baldvin & Maslova, 2009; Gurbanova, 2010).

In the light of increasing efforts towards the formation of Information Society, most countries seek to develop e-government solutions through ICT application. Implementation of e-government initiative acts as a platform for the maintenance of tolerance, cultural, religious and linguistic diversity along with achieving effectiveness in public administration, and calls for taking serious steps at national, regional, and local levels.

Forming a new information environment and enhancing public sector activity, e-government has emerged as a means of increasing the effectiveness of services provided (Alshehri, & Drew, 2011; Alguliyev, & Yusifov, 2014; Vasilyeva, & Kononenko, 2016). At the same time, developing countries do not have full access to the advantages of e-government yet, and therefore, they are facing many challenges in implementing e-government projects. These problems include political-administrative, infrastructural, demographic, linguistic, and other social factors, which play an important role in the implementation of e-government projects. Demographic features of the citizens, their age, gender, education, language diversity, etc. refer to the factors significantly affecting behavior of users and their access to online services. In this regard, protection of language diversity is one of the topical issues in generating e-government services. Rapid development of ICT and gradual dominance of English language hinder the access of population and ethnic minorities to e-services (Mittal, & Kaur, 2013; Torgby, & Asabere, 2014). Particularly, in the countries, where multilingualism exists and other languages are used along with the official state language, such as in Latin America and Africa, India, and the People's Republic of China, the use of dominating language on e-government websites may restrict access of majority of population to services (Lata, & Chandra, 2010; Mittal, & Kaur, 2013; Torgby, & Asabere, 2014; Pérez-Salazar, Aguilar-Edwards, & Mata-Martínez, 2016). Emergence of dominant languages in e-government upsurges the relevance of language diversity and language protection issues.

2. The effect of globalization on linguistic space

The concept of a unique information space is regarded as one of the key approaches to understanding the dynamics of linguistic changes. From this point of view, a unique information space enables the multifaceted activities of international and local media to be presented as a single and complete system.

Any information is transferred via the symbols related to this or that language and culture. In this aspect, linguistic space becomes the most important component of a unique information space. The range of English-

American linguistic space is extending far beyond the borders of the states where it is used and forming a vast English-language information space (Dobrosklonskaya, 2012; Laletina, 2011; Crystal, 2001). Along with the socio-political changes taking place in post-Soviet countries, serious changes in the linguistic-cultural information environment are also noteworthy. From this point of view, the consequences of the globalization's effects on the linguistic sphere can be attributed to the fact that linguistic processes are subjected to quality changes more due to the application of new information technologies.

To evaluate the impact of online media on linguistic processes, three leveled (geo-linguistic, inter-linguistic and intra-linguistic) assessment methods are proposed in scientific literature (Dobrosklonskaya, 2012). Geo-linguistic level involves the analysis of media's impact on the status and development of linguistic processes in the world or in the region. Inter-linguistic level analyzes the relations among languages, interaction of languages, the effects of languages on one another, the functionality of languages and so forth. Intra-linguistic level incudes the analysis of language processes, including norms and boundaries of spoken language, linguistic changes, and deficiencies in the spoken language within the linguacultural areal.

In the globalization process, the real threat to this or that language is undoubtedly provided by dominant languages, namely English, which is now believed to be a world language (Crystal, 2003; Gritsenko, 2011; Alakbarova, 2012). English-language resources, which are rapidly disseminated due to economic, political and socio-cultural reasons in the information space, significantly prevail the resources available in all other languages. Unquestionably, one of the main reasons for such a rapid spread of English is the availability of authoritative media tools.

As a result of the political, economic and technological changes taken place in Europe over the last decade, the role of English has significantly increased, becoming a language for communication and cooperation for Europe as well as for the West and East. Dutch linguist Olga Fischer believes that the role of language in the international communication has its own negative features, such as low level language proficiency, cultural diversity, differences in structural and meaning diversity, and the risk of losing its national identity and falling under the control of alien culture presented in English (Fischer, 2006).

The global impact of English in the modern world has been thoroughly investigated in the book, "English as a Global Language," by the famous British linguist David Crystal (2003). From the 90s, the concepts such as "linguistic imperialism", "cultural imperialism", "media imperialism" and "information imperialism" have emerged and resulted in extensive discussions (Dobrosklonskaya, 2012; Crystal, 2001, 2003). Given the fact that linguistic influences in the information society are mainly realized through media

channels, the influence of the dominant English language on a unique information space, including its Azerbaijani segment, is evident. In this regard, protection of Azerbaijani language in a unique information space becomes essential, and its geo-linguistic, inter-linguistic and intra-linguistic assessment and the development of effective mechanisms are the demands of the day.

3. Linguistic diversity and linguistic technologies in e-government

Multilingualism and freedom of expression are considered to be a basic value and one of the basic principles of e-democracy formation by many countries, including the European Union. For example, although European multilingualism has fundamental cultural and social values, language diversity can be a barrier significantly affecting the communication (Linguistic Diversity Roadmap, 2010; Language Cloud, 2015). From this point of view, preserving "unity and diversity" becomes very complicated issue. At present, this issue is also relevant for many languages in relation to the dominance of English and requires taking serious measures to protect the languages of ethnic minorities.

The effects of linguistic fragmentation can be shown clearly in social media such as Twitter. If we review the scope of communication language, we can clearly see that conversations and discussions are mainly limited to national languages and geographical boundaries (Alguliyev, & Mahmudov, 2018). At present, the languages of the peoples settled in Azerbaijan are mainly related to sevДилбаperal language groups, though Azerbaijani language, which belongs to Oghuz group of Turkic languages, is dominant at the modern ethnolinguistic background of Azerbaijan. Representatives of all nationalities, ethnic minorities and groups in Azerbaijan are forming an information environment within a family and mutual understanding. From this point of view, e-Azerbaijan segment of a unique information space is mainly limited to Azerbaijani language.

In the online environment, several technological solutions for preventing linguistic fragmentation or protecting language are available. Advances in language technology include machine translation, text analytics, semantic analysis, and speech recognition technologies that help to eliminate language barriers and preserve language diversity and promote multilingualism in the digital world (Gurbanova, 2010; Language Cloud, 2015; Alguliyev, & Mahmudov, 2018). As a result of the development of language technologies, people can read, write and speak their language on the Internet, while others have access to information in the languages understandable to them. The use of linguistic technologies in online environment will abolish the language barrier and create a unique information space among countries by developing language diversity in the field of language protection and e-government. For example, the availability of the content in multiple languages, which is presented in of the e-commerce environment capable of influencing the development of the country's economy, will enable the users to communicate and interact with each other

without any interference (Alguliyev, & Mahmudov, 2018). This, in turn, serves to the development of country's economy and the efficiency of e-government services.

4. Protection of Azerbaijani language in e-government

Language protection is to prevent the loss and endangering of languages, especially those of ethnic minorities. If a language is not taught to younger generations, it is exposed to endangering, and since language carriers are mainly spoken by older generations, their death leads to the death of language. On the other hand, the emergence of a dominant language in the context of globalization restricts the use of other languages. Language is an important element of any society and state, and people can communicate and express their thoughts through language. The death of language means that future generations lose a vital part of life that is essential for a comprehensive perception of culture. From this point of view, protection of language, as an important component of culture, is essential.

According to the UNESCO publication Atlas of the World's Languages in Danger (2010), about 6,000 languages are spoken all over the world, and half of the world's population speak in 8 wide-spread languages. Approximately, 425 languages are expected to be endangered (Alguliyev, & Mahmudov, 2018; UNESCO Atlas of the World's Languages, 2016; Krasikova, 2010; Gurbanova, 2015).

There are various factors that endanger languages. The main reasons for the language loss include emergence of dominant languages, globalization and migration. Moreover, language loss often occurs due to imbalanced distribution of languages among population. Massive migration of people results in forgetting to speak their own language after a certain period of time. In addition, political and military conflicts can also threaten the language loss. In the rapidly globalizing world, the development of the Internet, formation of information society and emergence of a unique information space, including the dominant status of English are the key factors accelerating the process of language loss.

Implementation of relevant programs and projects on the realization of initiatives for language protection in e-government should be carried out on phases.

From this point of view, the following issues are scheduled to be solved for the protection of Azerbaijani language in the e-government:

Integrating e-services into a unique information space, and forming e-Azerbaijan environment;

Standardizing e-services based on international standards, and supporting multilingualism, providing access to services to all citizens regardless of their national identity, communication language, and location;

Developing mechanisms to ensure e-services accessible for all categories, taking into account language diversity.

- Creating a virtual language teaching platform;
- Developing a national terminological information system;
- Developing linguistic technologies (machine translation, text analytics, semantic analysis, language sanitization, speech recognition);
- Creating transliteration systems;
- Taking measures to develop language industry;
- Establishing dialectical base, and forming language culture in online environment, etc.

In Fig. 1 shown the conceptual model of Azerbaijani language protection in e-government. The application of cloud-based linguistic technologies to protect the Azerbaijani language in e-Azerbaijani segment of the global information space will eventually create an effective mechanism for geo-linguistic, inter-linguistic and intra-linguistic assessments.

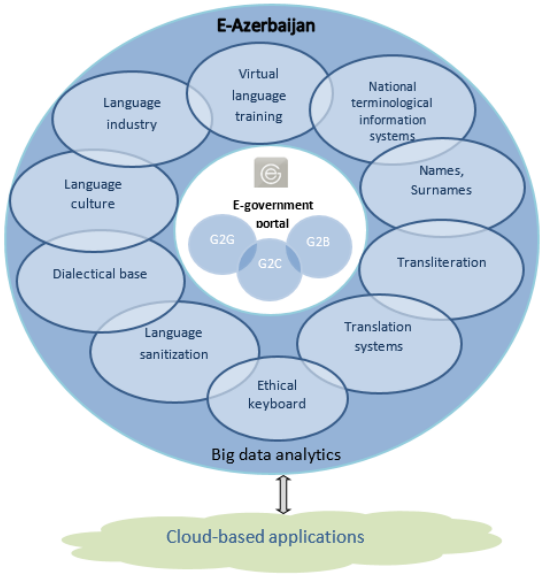


Fig. 1. Conceptual model of protection of Azerbaijani language in e-government

Conclusion

In general, there are various ideas and approaches related to language protection on the Internet. There are many factors endangering language in the globalization process, and it is not possible to maintain the language at the expense of controlling each of these factors. In this regard, the Internet can also be used to raise awareness about the language loss and language protection. The application of cloud-based linguistic technologies enables the provision of different services over the Internet (translation, cataloging, data storage, availability etc.). They can also be used to preserve the integrity of spoken language. Furthermore, many of these technologies used for sound and speech recognition can be applied to protect spoken languages. The use of linguistic technologies in e-Azerbaijan segment of the unique information space will

contribute to the protection and development of Azerbaijani language by providing variety of e-services.

REFERENCES:

1. Alakbarova, I.E. (2012). English in a globalized world, Lingua mobilis, No. 4 (37), pp.104-110.
2. Alguliyev, R.M., & Mahmudov, R.Sh. (2018). Language Industry: opportunities, prospects and problems, Journal of Information Society Problems, No. 1, pp. 3-26.
3. Alguliyev, R.M., & Yusifov, F.F. (2014). Some actual scientific-theoretical problems and solution prospects of the formation of electronic government, Journal of Information Society Problems, pp. 3-13.
4. Alshehri, M., & Drew, S. (2011). E-government principles: implementation, advantages and challenges. International Journal Electronic Business, 9(3), pp. 255-270.
5. Crystal, D. (2001). Language and the Internet. Cambridge, UK : Cambridge University Press, 272 p.
6. Crystal, D. (2003). English as a Global Language. Cambridge UK : Cambridge University Press, 2003, 160 p.
7. Dobrosklonskaya, T.G. (2012). Linguistic consequences of information globalization, languages in the modern world: Proceedings of the X International Conference, Moscow, 2012. pp. 59-71.
8. Fischer, O. (2006). Morphosyntactic Change, Functional and Formal Perspectives, 398 p.
9. Gritsenko, E.S. (2011). Language and security in the context of globalization. Scientific journal Vlast, Moscow, No. 11, pp. 9-11.
10. Gurbanova, A.M. (2010). Azerbaijani Language in Virtual Space: Some Problems and Solutions, Information Society Problems, No 1, pp. 63-70.
11. Gurbanova, A.M. (2015). Terminological threats to Azerbaijan in globalization, Information Society Problems, No.2, pp.87-95
12. Krasikova, E.N. (2010). Globalization of English language in a multicultural environment, www.ncfu.ru/uploads/doc/krasikova_konfmt.pdf
13. Laletina, A.O. (2011). Globalization as an object of linguistic research, Political Linguistics, No 3, pp.39-45
14. Language Cloud (2015). The European Language Cloud, or how to enable a multilingual Europe, www.euractiv.com
15. Lata, S., & Chandra, S. (2010). Development of Linguistic Resources and Tools for providing multilingual Solutions in Indian Languages – A Report on National Initiative, Proceedings of the International Conference on Language Resources and Evaluation, LREC'2010, 2010, Valletta, Malta, pp. 2851-2854.

16. Linguistic Diversity Roadmap (2010). The European Roadmap for Linguistic Diversity, Towards a new approach on languages as part of the European Agenda 2020, www.npld.eu/uploads/publications/313.pdf
17. Maurais, J., & Morris, M.A. (2003). Languages in a Globalising World, Cambridge, UK: Cambridge University Press, 345 p.
18. Mittal, P., & Kaur, A. (2013). E-Governance — A challenge for India, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 2, issue 3, pp. 1196-1199.
19. Pérez-Salazar, G., Aguilar-Edwards, A., $ Mata-Martínez, C.N. (2016). Linguistic diversity and accessibility in Mexican government web sites: executive branch, Comparative Cultural Studies: European and Latin America Perspectives, No 2, pp.1-13.
20. Smetanina-Baldvin, Yu.V., & Maslova, E.V. (2009). Internet as an informational, linguistic and social phenomenon and its role in globalization and localization of world languages, News of VSU. Series: Philology. Journalism, No 1, pp. 176-183.
21. Torgby, W.K., & Asabere, N.Y. (2014). Challenges of Implementing and Developing E-Government: A Case Study of the Local Government System in Ghana, International Journal of Computer Science and Telecommunications, vol. 5, issue 8, pp. 39-48.
22. UNESCO Atlas of the World's Languages in Danger, 2010, www.unesco.org
23. Vasilyeva, E.G., & Kononenko, D.V. (2016). Modern interpretations of the concept of electronic state (electronic government), Bulletin of Volgograd State University. Series 5, Jurisprud. No. 1 (30), pp. 9-16.

TERMINOLOGY IN MACHINE TRANSLATION

N. Temirova, M. Komilova, M. Kholmuminova
Tashkent state university of Uzbek language and literature named after Alisher Navoi, komp_ling@mail.ru

In this article we study what the terms «term» and «terminology» mean and translation categories of medical terms It is also discussed that the importance and problems of terminology in machine translation and given some reasonable solutions to those problems.

Key words: *terminology, term, word by word translation, transliteration, transcription, calque, polysemantic, medical terms.*

ТЕРМИНОЛОГИЯ В МАШИННОМ ПЕРЕВОДЕ

Н. Темирова, М. Комилова, М. Холмоминова,
Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, komp_ling@mail.ru

В этой статье мы изучаем, что означают термины «термин» и «терминология», и категории перевода медицинских терминов. Также обсуждается важность и проблемы терминологии в машинном переводе и даются некоторые разумные решения этих проблем.

Ключевые слова: *терминология, термин, дословный перевод, транслитерация, транскрипция, калька, полисемантика, медицинские термины.*

Introduction

In the era of globalization it can be observed that new words and phrases with their new concepts in every sphere. Thanks to modern discoveries nearly every day a lot of different terms to name them are coming into usage. Firstly, it should be given what special words «term» and «terminology» mean before noting the importance of it in machine translation(MT). The origin of «term» goes back to the word «terminus» in the Latin language. According to Merriam-Webster Dictionary a term is «a word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession or subject». In ABBYY Lingvo it is given that «a word or phrase used to describe a thing or to express a concept, especially in a particular kind of language or branch of study». Whereas «terminology» means «the technical or special terms used in a business, art, science, or special subject; a nomenclature as a field of study» in Merriam-Webster Dictionary and «the body of terms used with a particular technical application in a subject of study, theory, profession» in ABBYY Lingvo.

Depending on the above stated definitions, we can generalize that a term is a word or phrase that has a specific meaning in a particular field, whereas the terminology is a branch of linguistics that studies terms of any sphere. So, terminology is a wider concept that all terms are included in.

Terminology is very significant to compile dictionaries of precise terms differing from the existed ones in scientific translation. And adjoining such kind of word lists to the base of MT improves speed and quality of translation process.

Body part

By now transferring particular terms into another language has been carried out by human translators. But it requires translators to have general knowledge in both linguistics and that specific sphere. To own such knowledge translators need a great deal of energy and time. The reason of this, although terms do not have polysemantic meanings, some following difficulties can be seen in translation process:

1. Absence of certain term in another language exactly;
2. Expressing terms as a word combination;
3. Defining the style of the text and selecting the meanings of scientific terms which are appropriate to the sense of text in collaboration with its style;
4. Utilizing a lexeme in different meanings in a particular sphere.[35,2]

Creating specialized dictionaries for machine translation requires a plenty of vital researches. In developed countries this kind of work has been carried out relatively enough, whereas in Uzbekistan it has not been investigated yet. Because trained personnels are taught only in one sphere. Linguist can know source language and target language perfectly, but does not have enough knowledge about other field’s terminology. So, linguist and specialist should work together in order to create dictionary without any drawbacks.

Standardizing terms is very difficult. Bringing a new term into a particular language can be done by the help of various methods. For instance, while we were translating the medical terms from English into Uzbek, we used following ways:

1. Word by word translation: medullary cavity (a hollow centre of a long bone, containing bone marrow) — ilik kovagi (suyakning ilik joylashgan kovak qismi); muscle tissue (the specialised type of tissue which forms the muscles and which can contract and expand) — mushak to‘qima (qisqarish tuzilmalariga ega bo‘lgan hujayralar guruhidan tashkil topgan to‘qima).

2. Transliteration : multivitamin (a preparation containing several vitamins and sometimes minerals, used as a dietary supplement) – multivitamin (parhez davrida iste'mol qilinadigan bir qancha vitamin va minerallarni o‘z ichiga olgan preparat); Micrococcus (a genus of bacterium, some species of which cause arthritis, endocarditis and meningitis) – mikrokokk (meningit, endokardit, artrit va boshqalarni keltirib chiqaruvchi bakteriyalarning bir guruhi, turkumi);

3. Transcription: microcephaly (a condition in which a person has an unusually small head, sometimes caused by the mother having had a rubella infection during pregnancy) – mikrosefaliya (boshning juda kichikligi, tug‘ma rivojlanish nuqsoni; ba’zan ona xomiladorligida rubella infeksiyasi yuqtirganligi sabab ham yuzaga kelishi mumkin); microcheilia (the condition of having unusually small lips) – mikrochiliya (lablarning haddan ortiq kichikligi, tug‘ma rivojlanish nuqsoni);

4. Calque: microwave therapy (treatment using high-frequency radiation)– mikroto‘lqinli terapiya (yuqori chastotali radiatsiya yordamida davolash); muscular rheumatism (a disease giving pains in the back or neck, usually caused by fibrositis or inflammation of the muscles) — mushak revmatizmi (bo‘yin yoki belga og‘riq beruvchi, mushaklar yallig‘lanishi keltirib chiqaradigan kasallik).

We can see the importance of terminology in MT in these cases, majority of terms are in forms of word combination and phrases. However, machine considers that they are two or more different words come together and translates them automatically as a solitary word since there is no special vocabulary including correspondent translations of the word combinations and phrases. For instance, the combination «large intestine» may be translated by machine as «katta ichak» while there is not any term «katta ichak» among uzbek medical terms. In almost every dictionary we can not find this combination’s translation, so we search them separately: «large» – «katta, yirik», «intestine» – «ichak». And this causes serious mistakes during translation. If a specific dictionary exists in the base of machine, the machine can translate the term as «yo‘g‘on ichak». Adding such kind of dictionary to the base, actually, improves the quality and correctness of translation.

ЛИНГВИСТИЧЕСКИЙ СИНТЕЗ В ИССКУСТВЕННЫЙ ИНТЕЛЕКТ

Б. Х. Хайитқулов,
Национальный университет Узбекистана, b.hayitqulov@mail.ru

В статье дается определение термина «компьютерная лингвистика». Рассматриваются основные понятия предметной области, приводится классификация лингвистического программного обеспечения. Особое внимание уделяется соотношению терминологии компьютерной лингвистики и теоретической лингвистики.

Ключевые слова: *компьютерная лингвистика, машинный перевод, искусственный интеллект, алгоритмы и программы, обработка единиц естественного языка.*

LINGUISTICAL SYNTHESIS IN ARTIFICIAL INTELLECT

B. X. Khayitkulov, National university of Uzbekistan,
b.hayitqulov@mail.ru

Definition of the term «computational linguistics» is given. Main notions of its subject field are described, and classification of linguistic software is suggested. The emphasis is made on correlation between terminology in computational linguistics and in theoretical linguistics.

Key words: *computational linguistics, machine translation, artificial intelligence, algorithms and programs, natural language processing.*

СУНЪИЙ ИНТЕЛЛЕКТДА ЛИНГВИСТИК СИНТЕЗ

Б. Ҳ. Ҳайитқулов, O‘zbekiston Milliy universiteti
b.hayitqulov@mail.ru

Мақолада компьютер лингвистикаси термини моҳияти, соҳа доирасида ўрганилувчи масалалар, лингвистик таъминот тушунчаси ўрганилади. Шунингдек, лингвистик амалларнинг сунъий интеллект томонидан бажарилиши таҳлили амалга оширилган.

Калит сўзлар: *компьютер лингвистикаси, машина таржимаси, сунъий интеллект, алгоритмлар и дастурлар, табиий тилни қайта ишлаш.*

Компьютер лингвистикаси — математика, компьютер ёки ҳисоблаш лингвистикаси деб аталадиган янги соҳадир. Дастлаб ушбу тушунча табиий тиллар билан ишлаш учун компьютердан фойдаланишнигина ифодалаган

бўлса, ҳозирда сунъий интеллектдан фойдаланишни англатади [2]. У компьютер фанлари, математика, сунъий интеллект ва тилшунослик асосларида пайдо бўлди. Хусусан, ушбу йўналишнинг ривожланиши умумий тилшуносликка асосланган.

Компьютер лингвистикасининг анъанавий табиий тилларга ишлов бериш усулларидан фарқи биринчидан, лингвистикани яхлит ўрганишни моделлаштиришга қаратилган иккинчидан, тилни тавсифлайдиган математик моделини қуришга қаратилганлигидир. Компьютер лингвистикасининг асосий вазифасини автоматик матнни қайта ишлаш учун мослаштирилган моделлар ва тегишли алгоритмлар ва дастурлар мажмуи сифатида белгилаш мумкин [3]. Матнни қайта ишлаш ўз ичига уни таҳрирлаш, транслитерациялаш ва таржима қилиш амалиётини олади.

Машина таржимасидан одатда техник матнлар учун фойдаланилади. Ихтисослаштирилган луғатларда тўғри созлашлар амалга оширилганда сифатли таржима олинади. Машина таржимаси яхши натижага эришиши учун матн техник ёки соҳанинг расмий услубини қўллаган ҳолда амалга оширилади. Автоматлаштирилган таржиманинг инсон таржимасидан фарқи, таржима қилишга кетадиган вақтни камайтиради ва таржима матн сифатини яхшилишга ёрдам беради. Статистик компьютер таржимаси ҳам мавжуд. Ушбу ҳолатда, икки хил дастурлаш тилларида ёзилган дастурларни таққослаганда, бир хил маънони беради. Бундай дастурлар ўз-ўзини ўқитиш имкониятларига эга. Бундай таржимадан фойдаланишга Google таржимони мисол бўлади [7].

Машиналар фақат матнни таржима қилиш учун эмас, балки уни қайта ишлаш учун ҳам қўлланилиши мумкин. Нутқни таний олиш ва синхронлаш, матнларни таҳлил қилиш ва яратиш — бу йўналишнинг асосий вазифаларидир. Сўзнинг ўзини «тушуниш» сунъий интеллектнинг асосий вазифаларидан биридир. Тилнинг ноаниқликлари — бу табиий тилни қайта ишлашнинг асосий муаммосидир ва уларни ҳал қилиш табиий тилларнинг ташқи кўринишини ички тузилишга айлантириш орқали содир бўлади.

Тасодифий танланган сўзлардан матн ҳосил қилишда биз сифатсиз натижага эришамиз, шунинг учун матнлар олдиндан киритилган қолип иборалар ёрдамида яратилади. Аммо ҳозиргача етарли даражада матнларни яратишга қодир қолип ибораларни ҳосил қилувчи дастурий таъминотлар жуда озчиликни ташкил қилади.

Бугунги кунда компьютер технологияларининг қўлланилиш соҳалари кенгайиб бораётган бир вақтда турли соҳада машина ва инсоннинг оғзаки мулоқотига бўлган эҳтиёж тобора ортиб бормоқда. Шу билан бирга турли соҳадаги мулоқотни автоматик аниқлаш, таҳлил қилиш ва уни синтез қилишдаги кўпгина муаммоларни ҳал қилиш лозим бўлади. Инсон ва машина ўртасида оғзаки мулоқотни амалда қўллаш истиқболлари жуда кенгайиб бормоқда.

Инсон ва машина ўртасидаги оғзаки мулоқот тизимини яратиш учун фонетика, лингвистика, компьютер технологиялари, ҳисоблаш тизимлари назарияси, сигнални қайта ишлаш, тизимни моделлаштириш ва ахборот назарияси маълум даражадаги билимларни талаб қилади.

Нутқни аниқлаш тизими инсон эшитиш ва нутқ органларининг иш моделига асосланган. Шу билан бирга нутқ синтезатори нутқни шакллантириш жараёни асосида моделлаштирилган ва оғзаки нутқни матндан тўғридан-тўғри синтез қила олади [3].

Кўп сонли одамлар, жумладан, тилшунослар томонидан қўлланиладиган компьютер лингвистикасининг яна бир ютуғи электрон луғатлар ва онлайн луғатлардир. Агар сизнинг мобил ёки бошқа электрон қурилмангизда электрон луғат бўлса, унда луғат китобни кўтариб юришга эҳтиёж қолмайди. Бундай луғатларда сиз керакли сўзни тезда қидириб топишингиз мумкин [4].

Компьютер лингвистикасининг ривожланиши ҳатто робот тизимларига ҳам етиб борди. Матнни роботли тизимлар — объектларни ва уларнинг ҳолатларини таниган, турли анализаторлардан фойдаланадиган ва инсон билан мулоқотга асосланган кейинги ҳаракатларни белгилайдиган тизимлар. Ушбу тизимларда муаммоли муҳит моделларини тавсифловчи тил услуби ишлаб чиқилган [5]. Бундай модел, семантик тизим сифатида қаралади, унда синтактик муносабатлардан ташқари семантик йўналишларга, яъни семантик алоқаларга, мақсадга эришиш йўлида улар билан биргаликда ишлашга имкон бериши керак. Семантик алоқаларни жорий этиш натижасида моделнинг тавсифини қисқача кўрсатиш имкони мавжуд ва маълумотлар муайян семантик конвертация қилиниши мумкин.

Сунъий интеллект бўйича тадқиқотларда компьютерни табиий тил билан ишлашида ўзаро алоқасини таъминлаш жуда муҳимдир [6]. Дастурларни мослашувчан интерфейс билан жиҳозлаш талаб қилинади, чунки фойдаланувчиларнинг катта аудиторияси компьютер билан сунъий тилда мулоқот қилишни истамайди. Амалий тизимлар табиий тилларни тушуниш имконини берадиган интерфейс билан жиҳозланган, аммо маълум чекловлар мавжуд. Шу сабабли табиий тилларни қайта ишлашда кўплаб муаммолар ҳал қилинмаган.

Компьютер лингвистикаси замонавий фаннинг муҳим йўналиши бўлиб, сунъий интеллект тадқиқотларини чуқурлаштириш учун эмас, балки ушбу соҳадаги ҳар қандай ривожланиш бизга тилни ривожлантириш ва инсон фикрлашининг эволюция жараёнини тушунишга ёрдам беради [1].

Бугунги кунда сунъий интеллектга оид кўплаб назарий тадқиқотлар амалиётда қўлланилган. Яратилган роботлар гап туза олиши, расмларни таниши, мураккаб шароитлардан чиқиш йўлларини топа олиши, аниқ механик операцияларни бажариши мумкин. Шунга қарамай, сунъий

интеллектнинг асосий муаммоларидан бири табиий тилни компьютерда ўрганиш бўлиб қолмоқда.

АДАБИЁТЛАР:

1. Анисимов А. «Компьютерная лингвистика для всех — Мифы, Алгоритмы, Язык »// М. 1991 г.
2. Касевич В.Б. «Элементы общей лингвистики»//М. 1977 г.
3. Шемакин Ю.И. «Начала компьютерной лингвистики».// М. 1992 г.
4. Кузнецова Т.И., Кузнецов И.А. «Особенности развития иноязычной коммуникации в техническом ВУЗе»// М. 2012
5. Катранов С.Н., Кузнецов И.А. «Принципы подготовки переводчиков в сфере профессиональной коммуникации в системе дополнительного образования высшей школы»// Филологические науки. Вопросы теории и практики. 2016 г.
6. Кузнецова Т.И., Марченко А.Н., Кузнецов И.А. «Теория и практика обучения английскому языку в техническом ВУЗе»// М. 2014 г.
7. Rakhimova D.I., Kolpakova G.V., Kuznetsova T.I., Litvinov A.V., Samokhvalova A.G. Management of civil position’s formation of the student youth // International review of management and marketing. – 2016. – Т.6. №2. – С. 339-344.

СОДЕРЖАНИЕ

Ilmiy-ommabop nashr

Muharrir: Ulug‘ BEK
Texnik muharrir: Bobur Hamroyev
Sahihalovchi dizayner: Ulug‘bek Urunov
Musahhih: Sevinch Ahmedova

Litsenziya raqami: AI 310. 2017-yil 24-noyabr sanasida berilgan.
Bosishga 2018-yil 30-dekabr sanasida ruxsat etildi.
Bichimi: 60 x 84 1/8; Shartli bosma tabog‘i: 38,00.
Nashriyot hisob tabog‘i: 38,37. 01-sonli buyurtma.

ISBN 978-9943-5635-1-3

Original maket «NAVOIY UNIVERSITETI» nashriyot-matbaa
uyida tayyorlandi va www.Turklang.uz saytiga pdf shaklida joylashtirildi.

Nashriyot manzili: Toshkent shahri, Yusuf Xos Hojib ko‘chasi, 103-uy.
Tel.: +998 (94) 639-0344; (97) 344-0241; (90) 909-5401.
Web: navoiy-uni.uz E-mail: navoiyuniversiteti@mail.ru

₪ 388 ₪

С 23
УДК 811.512.1 (063)
ББК 81.2 ТЮРК (я43)

Шестая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2018». (Труды конференции) –Ташкент: Издательско-полиграфический дом «NAVOIY UNIVERSITETI», 2018. – 390 с.

Научно-популярное издание

Редактор: Улуг БЕК
Технический редактор: Бобур Хамраев
Дизайнер-верстальщик: Улугбек Урунов
Корректор: Севинч Ахмедова

Регистрация лицензии: AI 310. Выдана 24 ноября 2017 года.

Разрешено в печать 30.12.2018.
Формат: 60 x 84 1/8; Усл. печ. лист: 38,00.
Издат. лист: 38,37. Заказ №01.

ISBN 978-9943-5635-1-3

Оригинал макет изготовлен в издательско-полиграфическом доме «NAVOIY UNIVERSITETI» и размещен на сайте www.turklang.uz

Адрес издательства: г. Ташкент, ул. Юсуф Хос Хожиб, дом 103.
Tel.: +998 (94) 639-0344; (97) 344-0241; (90) 909-5401.
Web: navoiy-uni.uz E-mail: navoiyuniversiteti@mail.ru

Отдел маркетинга ИПД ГУП «NAVOIY UNIVERSITETI»:
+998 (97) 701–5401.

₪ 389 ₪