Maksutkhan N.,
Tleugabylova Z.A.,
Rakhimbayeva R.M.

Automated kazakh language morphological analyser

This article analyzes the morphemic structure in the corpus of Kazakh Language, and especially studies stem extraction and affix segmentation. This analysis will be useful to develop programs that relates to Kazakh language. There are two examples of the appliance of this analysis:

1) Multipurpose dictionary is a program that will help users to translate more complicated words in Kazakh which consist of different affixes into English. Furthermore, it will make a morphological analysis of translated word. In this case Kazakh morphology analysis helps to find the root of the word and show affixes and suffixes. The main algorithm will be like this:

- It extracts the root from the given word
- translates the root word and returns morphological analysis of word

2) Kazakh Language Spell Checker is tool that will help you to check spelling in Kazakh. Moreover, If the spelling is incorrect it will offer its own variants.

Key words: the morphemic structure, Kazakh Language, to develop programs, the root of the word.

Максутхан Н.,
Тлеугабылова З.А.,
Рахимбаева Р.М.

Қазақ тілінің автоматтандырылған морфологиялық талдағышы

Бұл мақалада қазақ тілінің морфемалық құрылымына, әсіресе түбірлердің шығу тегі саласындағы зерттеулерге талдау жасалынады. Бұл талдау қазақ тіліне байланысты бағдарламаларды әзірлеу үшін өте пайдалы. Осы талдауды қолдану үшін екі мысал бар:

1) Әмбебап сөздігі – әр түрлі аффикстерден тұратын күрделі қазақша сөздерді ағылшын тіліне аударуға пайдаланушыларға көмектесетін бағдарлама. Сонымен қатар, ол аударылған сөздерге морфологиялық талдау жасайды. Бұл жағдайда морфологияны қазақша талдау сөздің түбірін тауып және аффикстер мен суффикстерді көрсетуге мүмкіндік береді. Негізгі алгоритм төмендегідей болып табылады:

- ол берілген сөзден түбірді шығарады
- сөздің түбірін аударады және морфологиялық талдау жасайды

2) Қазақ тілінің орфографиясын (емле) тексеру – бұл құрал қазақ тілінде дұрыс жазуды тексеруге көмектеседі. Сонымен бірге, егер бұл емле дұрыс болмаса, онда бағдарлама өз нұсқаларын ұсынады.

Түйін сөздер: морфемалық құрылым, қазақ тілі, бағдарламалар әзірлеу, сөздің түбірі.

Максутхан Н.,
Тлеугабылова З.А.,
Рахимбаева Р.М.

Автоматизированный морфологический анализатор казахского языка

Эта статья анализирует морфемные структуры казахского языка, и особенно исследования в области происхождения корня. Этот анализ будет полезен для разработки программ, относящихся к казахскому языку. Есть два примера для применения этого анализа:

1. Универсальный словарь – программа, которая поможет пользователям перевести сложные казахские слова, состоящие из разных аффиксов, на английский язык. Кроме того, он будет делать морфологический анализ переведенных слов. В этом случае казахский анализ морфологии позволит найти корень слова и показать аффиксы и суффиксы. Основной алгоритм будет такой:

- он извлечет корень из заданного слова,
- переведет корень слова и сделает морфологический анализ.

2. Проверка орфографии казахского языка – это инструмент, который поможет проверить правописание на казахском языке. Кроме того, если это правописание будет неправильным, программа предложит собственные варианты.

Ключевые слова: морфемная структура, казахский язык, разработка программ, корень слова.

UDC 373.1.02:372.8

Maksutkhan N.[1], Tleugabylova Z.A.[2], Rakhimbayeva R.M.[3],

[1]master student, [2]senior lecture, [3]senior lecture
of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
e-mail: nazgul.kaz@gmail.com, tleugabylova@mail.ru, roza_madjarova68@mail.ru

AUTOMATED KAZAKH LANGUAGE MORPHOLOGICAL ANALYSER

The morphological analysis of language depends on language type. Kazakh Language belongs to Turkish Language group of Altaic Language Family, which are «agglutinative languages». In such languages words are formed by combining root words and morphemes. There are roots and several suffixes and affixes, when they are combined, the word modifies or extends its meaning.

Word formation rules.

The Kazakh alphabet consists of 42 letters, 37 sounds of which 12 (*А, Ә, Е, И, О, Ө, У, Ұ, Ү, Ы, І, Э*) are vowels, 25 (*Б, В, Г, Ғ, Д, Ж, З, Й, К, Қ, Л, М, Н, Ң, П, Р, С, Т, Ф, Ш, Щ, Х, Ц, Ч, Һ*) are consonants and others are soundless or combination of two sounds (*Ь, Ъ, Ю, Я, Ё*). In Kazakh language words are formed by adding to root words suffixes and affixes. Suffixes are always added before affixes. There are many rules concerning combination of root and suffixes/affixes. The root words are classified into two groups: nominal and verbal. Their affixes differ from each other. We focus only on nominal root words. The rules about verbal root words will be considered in the future researches. [1]

The affixes, which can be added to nominal roots in Kazakh language are divided into the following four types:

1) Plural: Kazakh language has six various affixes to express the plural form of words.

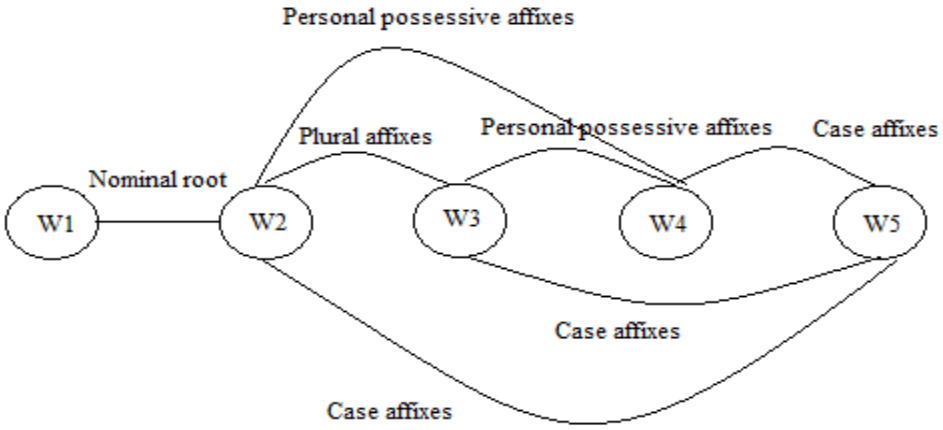
2) Personal possessive: Kazakh language has six various affixes to express the possessive forms of personal pronouns.

3) Case: Kazakh language has seven various affixes to express the different cases.

4) Predicative Person: The first, second and third personal pronouns are usually followed by the words with additive predicative personal elements. [1]

The above-mentioned four types of affixes can be used separately or linked together. Suffixes in Kazakh are complex, especially when a root is linked with many suffixes. There are some rules we can follow to add affixes to word roots (figure) [3]

Kazakh word formation uses a number of phonetic harmony rules. The vowel harmony rules require that vowels in a suffix/affix must be hard or soft according to the last syllable when they are affixed to a root. [2]



Figure

Hard vowels	*а, о, ұ, ы, у*
Soft vowels	*ә, е, и, ө, ү, і, э, у*

For example,

адам{person} + plural affix[*дар*, *дер*?], syllable *-дам* is hard, so we must add hard affix *-дар*. The result word is *адамдар*{people}.

But, there are other two plural affixes for this case: *-лар*,*-тар*. We didn't choose them. Why?

Another basic aspect of Kazakh phonology is consonant harmony. There are 3 groups of consonants:

Unvoiced consonants	*п, ф, к, қ, т, с, ш, щ, х, ц, ч, һ*
Voiced consonants	*б, в, г, ғ, д, ж, з*
Sonorant consonants	*р, л, й, у, м, н, ң*

The affixing rule according to consonant harmony is shown in table below:

The last letter of stem	Affixes
vowels or sonorant consonants *-р, -й, -у*	*-лар, -лер*
voiced consonants or sonorant consonants *-м, -л, -н, -ң*	*-дар, -дер*
unvoiced consonants or voiced consonant *-б, -в, -г, -д*	*-тар, -тер*

Here we considered only plural affixes. The other affixes conform to these rules too. But there can be some differences, which we are going to consider during the implementation of this project.

The algorithm

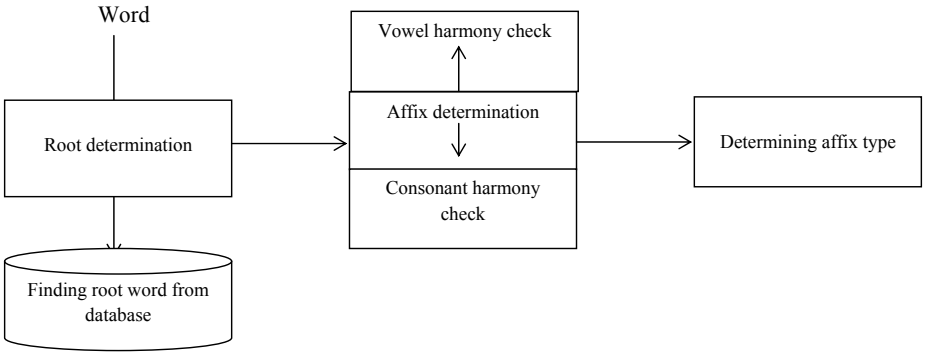


Figure 1 – Morphological analysis

Kazakh language database consists of all root words, which are classified by part of speech.

Root determination algorithm

1. The inputted word is «candidate root». After each step, next «portion» will be «candidate root».

2. To start from the right, by deleting one letter, if the «candidate root» is not found from Kazakh language database. If there is word that is equal to our «portion», we will stop and this «portion» will be the root of word.

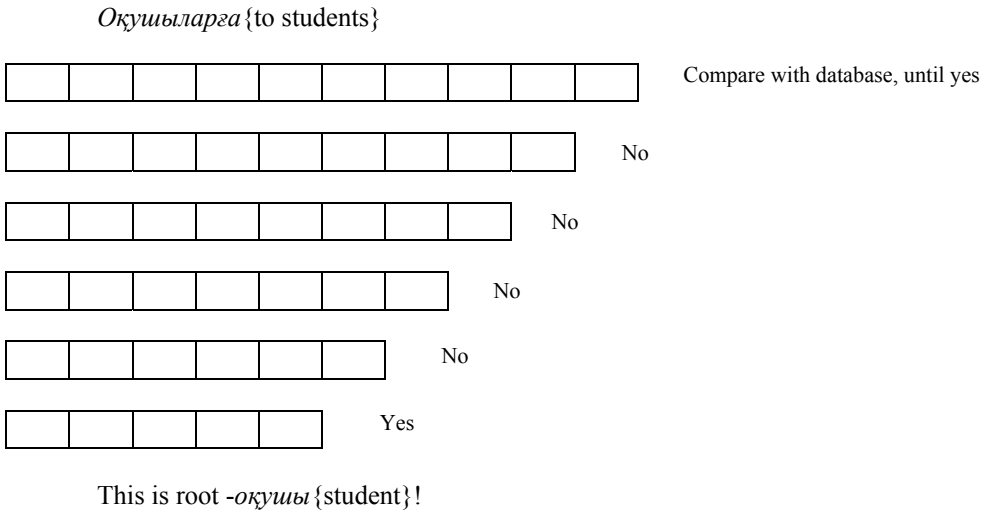


Figure 2 – Root determination

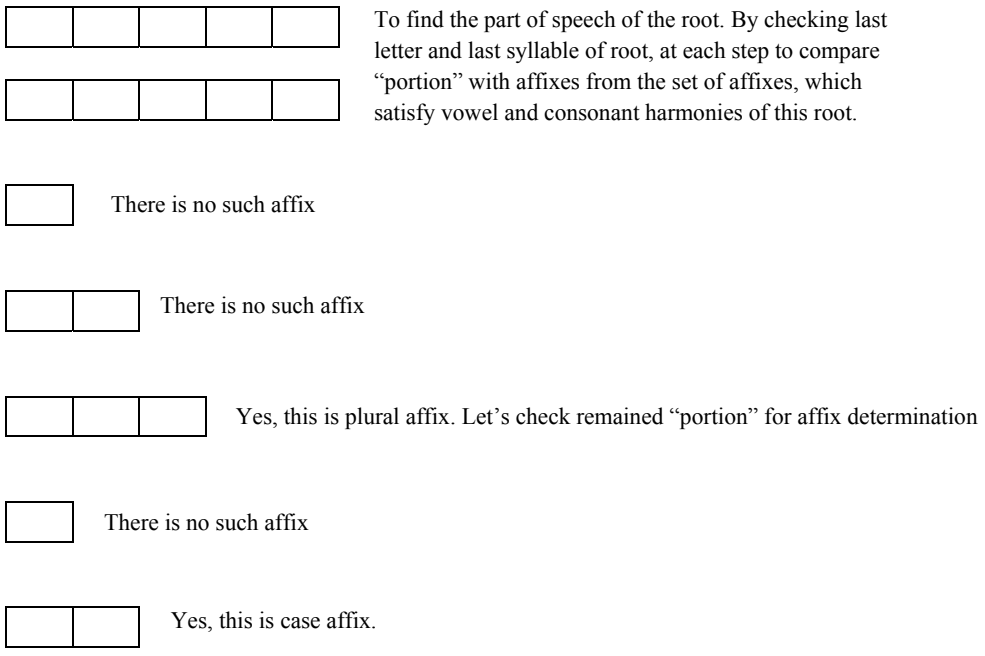


Figure 3 – Affix determination

The result of morphological analysis:

Оқушы [noun, hard syllable, vowel] + *лар*[plural affix, hard syllable] + *ға* [case affix, hard syllable]

References

1 Tujmebaev ZH.K. Kazahskij YAzyk. Grammaticheskij spravochnik. – Almaty, 1996. – S. 34.

2 SHaripbaev A., Bekmanova G. T. Sintez form slov tyurkskogo yazyka s pomoshch'yu semanticheskih nejronnyh setej. Sovremennye problemy prikladnoj matematiki i informacionnyh tekhnologij. – Tashkent: Al'-Horezmi, 2009. – 145 s.

3 SHaripbaev A. A., Bekmanova G. T. Logicheskaya semantika slov v kazahskom yazyke // Materialy Vserossijskoj konferencii. Znanie-ontologiya-teoriya. – Novosibirsk: Zont, 2009. – S. 246-249.

Литература

1 Туймебаев Ж.К. Казахский Язык. Грамматический справочник. – Алматы, 1996. – С. 34.

2 Шарипбаев А., Бекманова Г. Т. Синтез форм слов тюркского языка с помощью семантических нейронных сетей. Современные проблемы прикладной математики и информационных технологий. – Ташкент: Аль-Хорезми, 2009. – 145 с.

3 Шарипбаев А. А., Бекманова Г. Т. Логическая семантика слов в казахском языке // Материалы Всероссийской конференции. Знание-онтология-теория. – Новосибирск: Зонт, 2009. – С. 246-249.