У.А. Тукеев, Ж.М. Жуманов, Д.Р. Рахимова

*Казахский национальный университет им. аль-Фараби, Алматы, Казахстан; e-mail: Ualsher.Tukeyev@kaznu.kz, z.zhake@gmail.com, Diana.Rakhimova@kaznu.kz*

# Моделирование семантических ситуаций времен казахского языка при машинном переводе

Данная статья предлагает формальное описание времен глаголов в казахском языке с использованием семантических ситуаций и регулярных выражений. Описываются особенности системы времен казахского языка, семантические ситуации и их запись в форме регулярных выражений. С использованием семантических ситуаций моделируется система времен казахского языка. Приводится пример практического использования предлагаемого решения. Приведены практические результаты.
*Ключевые слова*: казахский язык, семантика, регулярные выражения, машинный перевод.

У.А. Тукеев, Ж.М. Жуманов, Д.Р. Рахимова

## Машиналық аударма үшін қазақ тілі шақтарының семантикалық жағдайларын модельдеу

Осы мақала семантикалық жағдайларды және тұрақты сөйлемшелерді пайдаланумен қазақ тіліндегі етістіктер шақтарының формальды сипаттамасын ұсынады. ?азақ тілінің шақтар жүйенің ерекшеліктері, семантикалық жағдайлар және олардың тұрақты сөйлемшелер түрінде жазылуы сипатталған. қазақ тілінің шақтар жүйесі семантикалық жағдайларды пайдаланумен модельденген. ?сынылатын шешімнің практикалық пайдаланудың мысалы жасалынды. Практикалық нәтижелер келтірілген.
*Түйін сөздер*: {қазақ тілі, семантика, тұрақты сөйлемшелер, машиналық аударма.}

U.A. Tukeyev, Zh.M. Zhumanov, D.R. Rakhimova

## Modeling of semantic situations for Kazakh language's tenses for machine translation

This paper proposes a formal description of Kazakh language verbs' tenses using semantic situations and regular expressions. The features of Kazakh language tenses system, semantic situations and their record in the form of regular expressions are described. With the use of semantic situations Kazakh language tenses system is modeled. An example of proposed solution's practical use is given. Practical results are presented. *Key words*: {Kazakh language, semantics, regular expressions, machine translation.}
*Key words*: {Kazakh language, semantics, regular expressions, machine translation.}

### Введение

Категория времени является одной из важных грамматических категорий естественных языков. В различных языках данная категория имеет разное грамматическое представление, однако общее назначение ее всегда одно и то же – показать то, как относится

текст к временной шкале, принятой в языке. За исключением очевидных способов отсчета времени (настоящее, прошлое, будущее) имеются еще несколько выражений времени, которые в некоторых языках бывают неочевидны. Формальное описание категории времени затрудняется тем, что одно и то же время может выражаться в языке разными грамматическими конструкциями. Это приводит в тому, что при переводе предложений одного естественного языка в другой может оказаться затруднительным найти соответствие времен.

Для решения этой проблемы можно использовать тот факт, что выражение времени в естественных языках включает две составляющих: грамматическую и смысловую. Причем вторая составляющая, на взгляд авторов, имеет более важное значение. Так, например, при автоматизированном переводе очень часто времена одного и того же предложения в разных языках могут не соответствовать друг другу.

**Категория времени в лингвистике**

Согласно [1], в грамматике время – это категория, которая определяет положение действия на временной шкале. Категория времени использует указание времени, привязанное к какому-то моменту. Например, перед текущим моментом (прошлое), в текущий момент (настоящее) и после текущего момента (будущее). Привязку времени к текущему моменту иногда называют абсолютным временем, привязку к моменту, отличному от текущего, - относительным временем. Первый случай порождает так называемые «простые» временные конструкции (past simple в английском языке, жедел ?ткен ша? в казахском языке). Использование относительного времени можно наблюдать в таких грамматических конструкциях как «будущее в будущем» или «будущее в прошлом» (например времена группы perfect в английском языке). Относительное время может иметь смысловую окраску предшествования одних событий другим, их одновременности или порядка их следования.

Существуют языки в которых категория времени реализуется не через грамматику, а с использованием явного указания момента совершения действия (например китайский). В некоторых языках система времен не «трехсоставная» (прошлое-настоящее-будущее), а «двусоставное». В японском языке есть категории прошедшего времени и непрошедшего времени (настощее и будущее время в одной глагольной форме), в гренландском языке имеются категории будущего и небудущего времени. В «четырехвременных» языках одна из временных категорий (прошедшее время или будущее время) уточняется – недавнее и отдаленное. Встречаются даже языки, использующие до 6 временных моментов.

В общем, в лингвистике можно выявить следующие виды времен:

Будущие времена:

- немедленное будущее: прямо сейчас;
- ближайшее будущее: скоро;
- сегодняшнее будущее: позже сегодня;
- вечернее будущее: в этот вечер;
- пост-сегодняшнее будущее: после сегодняшнего дня;

- завтрашнее будущее: завтра;
- отдаленное будущее или далекое будущее;
- относительное будущее;
- предположительное будущее;
- будущее с намерением;

Настоящие времена:

- продолжающееся настоящее: «все еще»;

Прошедшие времена:

- немедленное прошедшее: совсем недавнее прошлое, только сейчас;
- недавнее прошедшее: в последние несколько дней/недель/месяцев;
- давнее прошедшее: контрастирует с недавним прошлым;
- сегодняшнее прошедшее: ранее сегодня;
- утреннее прошедшее: сегодня утром;
- пред-сегодняшнее прошедшее: до сегодняшнего дня;
- вчерашнее прошедшее;
- пред-вчерашнее прошедшее: позавчера;
- далекое прошедшее: более чем несколько дней/недель/месяцев назад;
- недалекое прошедшее: контрастирует с далеким прошлым;
- историческое прошедшее: показывает, что действие/состояние было частью события в прошлом;
- древнее прошедшее: легендарное прошлое;
- пред-прошедшее (относительное прошедшее время).

Время - грамматическая категория глагола [2]. Следовательно, грамматические преобразования необходимые для выражения описанных значений применяются к глаголам. Для этого используются синтетические и аналитические формы глаголов. [2] В синтетических формах несколько морфем объединяются в пределах одного слова (например, «прочитаю», «напишет»). В аналитических формах основное и дополнительное значения слова выражаются раздельно (например «буду читать», «будет писать»).

**Описание системы времен казахского языка**

Таблица 1. Система времен в казахском языке.

| **Название времени** | **Образование** | **Значение** | **Пример** |
|---|---|---|---|
| Прошедшее время | | | |
| Жедел өткен шақ | Основа глагола + ды/ді/ты/ті + личные окончания | Недавнее прошедшее | Мен барғанмын. |
| Бұрыңғы өткен шақ | Основа глагола + қан/кен/ған/ген + личные окончания | Давнее прошедшее | Мен барыппын. |
| Бұрыңғы өткен шақ | Основа глагола + ып/іп/п + личные окончания | Давнее прошедшее | Сен келіпсін. |
| Бұрыңғы өткен шақ | Основа глагола + қан/кен/ған/ген + еді/екен + личные окончания | Давнее прошедшее | Сен кеткен екенсің. |
| Ауспалы өткен шақ | Основа глагола + атын/етін/йтын/йтін + личные окончания | Давнее прошедшее, иногда относительное прошедшее | Мен бұрын хат жазатынмын. |
| Настоящее время | | | |
| Нақ осы шақ | Отыр, түр, жүр, жатыр + личные окончания | Продолжающееся настоящее | Мен жатырмын. |
| Нақ осы шақ | Основа глагола + ып/іп/а/е/и + служебный глагол + личные окончания | Продолжающееся настоящее | Сен ойлап жүрсің. |
| Ауспалы осы шақ | Основа глагола + а/е/й + личные окончания | Настоящее или ближайшее будущее | Мен ойнаймын. |
| Будущее время | | | |
| Болжалды келер шақ | Основа глагола + ар/ер/р + личные окончания | Предположительное будущее | Мен келермін. |
| Мақсатты келер шақ | Основа глагола + мақ/мек/бақ/бек/-пақ/пек/шы/ші + личные окончания | Будущее с намерением | Сен жазбақсын |

В казахском языке можно выделить 7 времен. Три из них представляют прошедшее время, одно – настоящее время, два – будущее время и одно может выражать как настоящее, так и будущее время [2]. Все они представлены в таблице 1.

Как можно видеть из таблицы при образования времен в казахском языке в основном используются синтетические формы глагола. Аналитическую форму образования имеют варианты бұрыңғы өткен шақ и нақ осы шақ.

### Описание семантических ситуаций

В каждом естественном языке одним из ключевых компонентов является смысл, вкладываемый во фразы и выражения. Независимо от используемого языка в аналогичных контекстах используются выражения с аналогичным смыслом, но, зачастую, с различными грамматическими конструкциями. Примерами таких контекстов может быть «приветствие», описание родственных связей, использование порядковых числительных и т.п. Возможно составить набор семантических «ситуаций», которые имеют одинаковое смысловое значение во всех языках. Сравнительное описание пар языков, используемых в машинном переводе может быть дополнено набором соответствий грамматических конструкций различных языков одним и тем же семантическим ситуациям.

Пример такого набора для казахского и английского языков представлен в таблице 2.

Таблица 2. Пример набора соответствий грамматических конструкций английского и казахского языков одним и тем же семантическим ситуациям.

| **Казахский язык** | **Семантическая ситуация** | **Английский язык** |
|---|---|---|
| Обстоятельство места + подлежащее + «бар» | Наличие чего-либо, где-либо | «there is» + дополнение + обстоятельство места |
| Слово + «және» + слово<br>Слово + «да», «де», «та», «те»<br>Слово + «мен», «бен», «пен» + слово | Сочинительный союз «и» | Слово + «and» +слово |
| Основа + «-йін», «-йын», «-айін», «-айын» «-йік», «-йы?», «-айік», «-айы?» | Повелительное наклонение 1-го лица | Let me + «основа»<br>Let us + «основа» |
| Инфинитив + притяжательные окончания + «керек» | Долженствование, необходимость | Must + основа<br>have + инфинитив<br>ought + инфинитив<br>to be + инфинитив |
| Сказуемое (основа) + «емес» + личное окончание | Отрицательная форма именного сказуемого | Подлежащее + to be + «not» + сказуемое |

Формально семантические ситуации можно описать следующим образом:

<язык> ::= <предложение> | <язык><предложение>

<предложение> ::= <устойчивое выражение>

<предложение> ::= <сочетание слов> | <сочетание слов><предложение>

<сочетание слов> ::= <грамматическая форма слова> | <сочетание слов><грамматическая форма слова>

<сочетание слов> ::= <слово> | <сочетание слов><слово>

<грамматическая форма слова> ::= <слово><грамматические признаки>

<грамматическая форма слова> ::= <грамматические признаки><слово>

<грамматические признаки> ::= <аффикс> | <предлог> | <...>

<слово> ::= <существительное> | <прилагательное> | <...>

<семантическая ситуация> ::= <устойчивое выражение> | <сочетание слов> | <грамматическая форма слова>

Условия использования тех или иных времен глагола также могут являться семантическими ситуациями, которые имеют свое значение (смысл) и которые выражаются в синтетической или аналитической форме. Систему времен языка можно представить как набор семантических ситуаций. Но для этого помимо словесного описания этих ситуаций необходим формат записи, который позволит использовать полученный набор в дальнейшей компьютерной обработке.

**Регулярные выражения и их использование в обработке естественных языков**

В программных технологиях, работающих с текстами, существует инструмент, называемый регулярными выражениями, значительно облегчающий обработку текстовых

данных. Регулярные выражения — это часть технологической области программирования, широко используемой в огромном диапазоне программ. Регулярные выражения можно представить себе как мини-язык программирования, имеющий одно специфическое назначение: находить подстроки в больших строковых выражениях. Это не новая технология; изначально она появилась в среде UNIX и обычно используется в языке программирования Perl. Однако, в каждом развитом языке программирования существует своя реализация регулярных выражений, в общих чертах соответствующая варианту из Perl.

Язык регулярных выражений предназначен специально для обработки строк. Он включает два средства:

- набор управляющих кодов для идентификации специфических типов символов (метасимволы);
- система для группирования частей подстрок и промежуточных результатов таких действий (квантификаторы).

Перечень метасимволов регулярных выражений:

- . – любой символ за исключением конца строки;
- [abd] – один из символов, находящихся в квадратных скобках;
- [âbd] – один символ, который не присутствует в скобках;
- [0-9a-fA-F] – один символ из указанных диапазонов;
- \t – символ табуляции;
- \r – символ возврата каретки;
- \n – новая строка;
- \e – символ escape;
- \w – большие и маленькие латинские буквы, цифры, знак подчеркивания;
- \s – пробел;
- \d – любая цифра;
- \b - граница слова.

Перечень квантификаторов регулярных выражений:

- *- предыдущий символ может повторяться 0 или более раз;
- + - предыдущий символ может повторяться 1 или более раз;
- ? - Предыдущий символ может повторяться 0 или 1 раз.

Для разбиения регулярных выражений на группы можно использовать скобочки. Символ '|' можно использовать для перебора нескольких вариантов. Использование этого символа совместно со скобками – '(...|...|...)' – позволяет создать группы вариантов.

Приведем несколько примеров. Регулярное выражение для имени пользователя ИС (логина), которое может содержать буквы, цифры, подчеркивания, дефисы и быть длиной от 3 до 16 символов может выглядеть следующим образом: [a-z0-9_-]{3,16}.

Регулярное выражение для формата даты dd MMM yy может выглядеть так: [0-3]{1}[0-9]{1}[ ]{1}(Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec){1}[ ]{1}[0-9]{2}.

Семантические ситуации, которые используют синтетическую форму образования, можно записать в виде регулярного выражения. Составление регулярных выражений для семантических ситуаций с аналитической формой образования может иметь определенные сложности, если в соответствующем грамматическом правиле отсутствуют закономерности.

**Моделирование системы времен с использованием семантических ситуаций**

Используем семантические ситуации и их запись в виде регулярных выражений, описанные выше, для описания системы времен глаголов в казахском языке. (таблица 3) В данной статье используется синтаксис регулярных выражений из языка программирования C#. В других языках программирования могут быть незначительные отличия.

Таблица 3. Система времен казахского языка в виде семантических ситуаций.

| | |
|---|---|
| Жедел өткен шақ | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ді\|ды\|ті\|ты)(м\|н\|ныз\|ніз\|к\|қ\|ндар\|ндер\|ныздар\|ніздер)?\b" |
| Бұрынғы өткен шақ 1 | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ып\|іп\|п)(пын\|пін\|сын\|сін\|сыз\|сіз\|ты\|ті\|пыз\|піз\|сындар\|сіндер\|сыздар\|сіздер)\b" |
| Бұрынғы өткен шақ 2 | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ған\|ген\|қан\|кен)(мын\|мін\|сын\|сін\|сыз\|сіз\|мыз\|міз\|сындар\|сіндер\|сыздар\|сіздер)?\b" |
| Бұрынғы өткен шақ 3 | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ған\|ген\|қан\|кен) екен(мін\|сін\|сіз\|біз\|сіндер\|сіздер)?\b" |
| Ауыспалы өткен шақ | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(атын\|етін\|йтін\|ытын) еді(м\|н\|ніз\|к\|ндер\|ніздер)?\b" |
| Нақ осы шақ | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ып\|іп\|п\|а\|е\|я)(отыр\|тұр\|жүр\|жатыр)(мын\|мін\|сын\|сін\|сыз\|сіз\|мыз\|міз\|сындар\|сіндер\|сыздар\|сіздер)?\b" |
| Ауыспалы осы шақ | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(а\|е\|й)(мын\|мін\|сын\|сін\|сыз\|сіз\|ды\|ді\|мыз\|міз\|сындар\|сіндер\|сыздар\|сіздер)?\b" |
| Болжалды келер шақ | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ар\|ер)(мын\|мін\|сын\|сін\|сыз\|сіз\|мыз\|міз\|сындар\|сіндер\|сыздар\|сіздер)?\b" |
| Мақсатты келер шақ | @"\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(мақ\|мек\|бақ\|бек\|пақ\|пек)(пын\|пін\|сын\|сін\|сыз\|сіз\|пыз\|піз\|сындар\|сіндер\|сыздар\| сіздер)?\b" |

Каждое из записанных выражений содержит элемент [аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+. Буквально подобная запись обозначает произвольная последовательность символов в квадратных скобках длиной от 1 символа. В формах грамма-

тических времен казахского яззыка используются основы слов, а именно глаголов. По этой причине к указанному элементу должно быть добавлено дополнительное требование – последовательности символов должны соответствовать основам глаголов казахского языка.

**Практическая реализация**

Предлагаемое представление системы времен казахского языка в виде семантических ситуаций можно использовать для решения задач программирования, связанных с естественными языками. В качестве подобного использования можно привести пример программы машинного перевода, которая переводит формы времен казахского языка в соответствующие им формы времен английского языка.

Код в листинге 1 производит поиск глагольных временных форм (семантических ситуаций описанных в предыдущем разделе) в некотором тексте. Переменная regex содержит описание семантической ситуации в виде регулярного выражения. Переменная text содержит анализируемый текст. Операция regex.Match() осуществляет анализ соответствия текста регулярному выражению. В случае нахождения соответствия, анализируемый текст содержит форму глагольного времени. Результат работы кода можно видеть на рис. 1.

Листинг 1 – Анализ текста на наличие в нем глагольных форм.

```
this.verb_situation_patterns_1.Length/3; i++) {
    Regex regex = new Regex(this.verb_situation_patterns_1[i, 1]);
    Match match = regex.Match(text);
    if ((match.Success) && (match.Length == text.Length))
        return new  Verb_Situation();
}
```

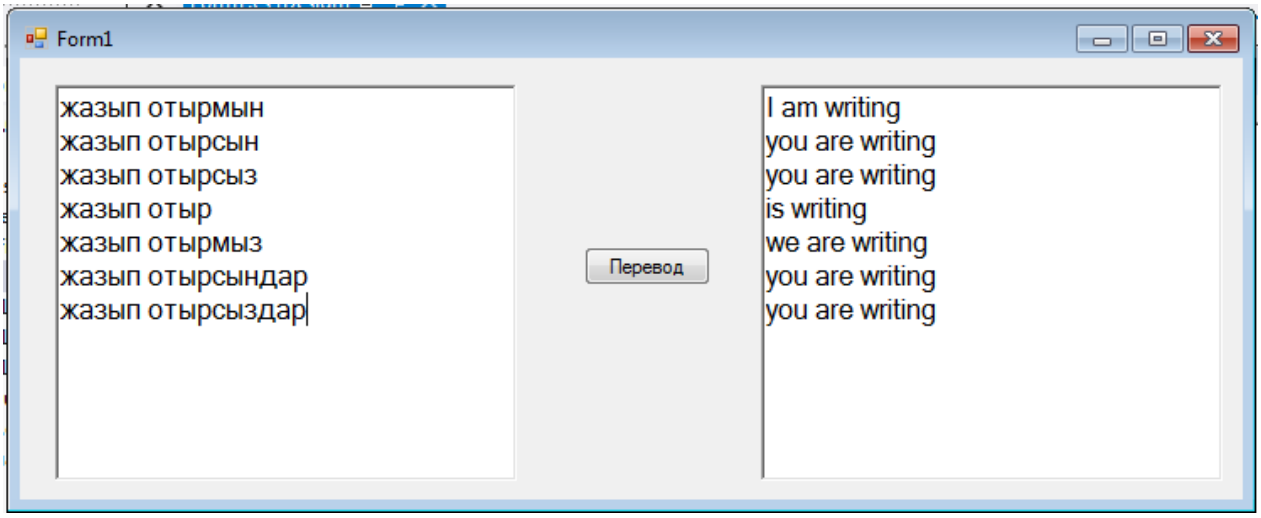


Рис. 1. Результат анализа и перевода глагольных форм.

Как можно видеть по рисунку, была корректно определена ситуаций «нақ осы шақ» и подобрано соответствие этой ситуации в английском языке – present continuous tense.

Все возможные вариации нақ осы шақ по лицам и числам были найдены и распознаны обработчиком регулярных выражений.

**Заключение**

Данная статья посвящена проблеме моделирования лингвистической категории времени. Для решения этой проблемы предлагается использовать две составляющих выражения времени в естественных языках: грамматическую и смысловую. Смысловая составляющая представлена набором семантических ситуаций, грамматическая – соответствующим набором регулярных выражений.

В статье описаны особенности системы времен казахского языка, представлены семантические ситуации и их запись в форме регулярных выражений, с использованием которых смоделирована система времен казахского языка. Приведен пример практического использования предлагаемого решения.

Использование предлагаемого аппарата регулярных выражений в разработке программ обработки естественных языков позволяет многократно сократить объем программного кода.

Развитие предлагаемого решения связано с дальнейшей проработкой описания семантических ситуаций для того, чтобы они могли лучше отражать особенности грамматических конструкций с аналитической формой образования.

## Список литературы

[1] C. Fabricius-Hansen, “Tense.” Encyclopedia of Language and Linguistics, 2nd ed. Amsterdam: Elsevier, 2006. pp. 566-573.

[2] Лингвистический энциклопедический словарь. Главный редактор В.Н. Ярцева Москва. «Советская энциклопедия». 1990

[3] Мурзагельдинова О.И. Грамматический справочник-шпаргалка по казахскому языку. Келешек-2030, 2009 г.- 44 с.

[4] Дж. Фридл Регулярные выражения. Издательство «Питер», 2003. 460 с.

*Поступила в редакцию 23 ноября 2012*