

Построение логической семантики слов казахского языка

Шарипбаев А.А.¹, Беманова Г. Т.²

¹Евразийский национальный университет им. Л. Н. Гумилева, ул.
Мунайтпасова 5, г. Астана, 010000, Казахстан

²Евразийский национальный университет им. Л. Н. Гумилева, ул.
Мунайтпасова 5, г. Астана, 010000, Казахстан

sharalt@mail.ru, gulmira-r@yandex.ru

Аннотация. В работе рассматриваются представление словаря словоформ и их семантических признаков казахского языка в виде нейронной сети. С целью построения логической семантики словоформ казахского языка

Ключевые слова: естественные языки, нейронные сети, распознавание речи.

Казахский – типичный тюркский язык, сохранивший большинство черт, общих для этой группы и обладающий рядом характерных кыпчакских особенностей.

Структурно-типологическая характеристика казахского языка, в основном, связана с его принадлежностью к агглютинативным языкам. Для описания агглютинативного типа применяется, как правило, набор признаков, учитывающих не только фонетические, но также морфологические и синтаксические особенности.

В казахском языке строго определен порядок расположения окончаний. Например, для имен существительных к основе слова вначале добавляется окончание множественного числа затем притяжательное окончание (означает принадлежность предмета тому или иному лицу) далее следует падежное окончание и последним окончание формы спряжения (добавляется только к одушевленным существительным). [1]

Таким образом, правило добавления окончаний можно представить в виде:

$$W=S\&Kg\&Tg\&Sg\&Gg \quad (1),$$

где W – словоформа,
S – основа слова,
Kg- окончание множественного числа,
Tg- притяжательное окончание,
Sg- падежное окончание,
Gg- окончание формы спряжения.

Казахский язык является флективным, то есть словоформы образуются путем соединения (конкатенации) буквенных цепочек – основы и флексий, с каждой из которых связан набор морфологических характеристик (см. пример 1)

Пример 1. Оқушы-лар-ыңыз-быз : основа «оқушы» (ученик), «лар» - окончание множественного числа, «ыңыз»- притяжательное окончание, «быз»- окончание формы спряжения 1-го лица. Таким образом, это казахское слово употребляется в контексте «Мы, ваши ученики...».

При добавлении более чем одного окончания анализируется, чем оканчивается основа слова, выделяется признак, по которому добавляется первое окончание, далее анализируется добавленное окончание, выделяется признак, по которому добавляется следующее окончание и т. д. В случае отсутствия одного из окончаний (нулевое окончание) последующее окончание добавляется к тому, что ему предшествует.

Аналогично изменяются другие изменяемые части речи казахского языка.

Таким образом, из словаря начальных форм представляется возможным сформировать словарь всех существующих словоформ различных частей речи казахского языка.

В качестве структуры семантической нейронной сети, выполняющей морфологический анализ, можно выбрать синхронизированное линейное дерево. В семантической нейронной сети каждый символ соответствует отдельному нейрону. Поскольку одинаковые символьные последовательности разных слов хранятся как один фрагмент линейного дерева, то грамматический словарь, реализованный в виде семантической нейронной сети, будет иметь объем меньший, нежели, чем если хранить каждую словоформу в виде отдельной записи.

Можно представить словарь всех словоформ в виде модели нейрона .

Начальную форма слова, из которой образуются все его словоформы путем склонения, спряжения и т.д. , назовем леммой. Пусть лемма - это группа нейронов, или один нейронный субавтомат в слое извлечения смысла. Пусть общее число субсостояний словарной статьи равно числу словоформ этой статьи. Пусть каждое субсостояние такого субавтомата представляет собой один возбужденный нейрон. При этом, в случае одновременного возбуждения двух разных нейронов одного субавтомата будем говорить что субавтомат имеет одновременно два разных субсостояния. Каждая словарная статья имеет главный нейрон, соответствующий лемме. Главный нейрон леммы возбужден всегда, когда распознано слово, принадлежащее лемме. Каждой словоформе соответствует отдельный нейрон. Он возбуждается в случае, если словоформа распознана.

В слое извлечения смысла существуют нейроны, не принадлежащие отдельным леммам. Эти нейроны соответствуют признакам словоформ общим для многих словарных статей, таким как род, падеж, число, время ... Они возбуждаются при возбуждении словоформ, обладающих соответствующими признаками. Будем считать, что состояния этих нейронов, соответствующие признакам словоформ, так же принадлежат к субавтоматам лемм, с которыми эти нейроны образуют связи. Тогда, несколько лемм могут одновременно находиться в одном и том же состоянии. Задача классификации или определения словарной статьи и словоформы по заданной символьной последовательности сводится к прохождению волны возбуждения через слой извлечения смысла и возбуждению соответствующего субавтомата для соответствующей словарной статьи. Задача словоизменения сводится к изменению состояния такого субавтомата из начального состояния - соответствующего словоформе из которой начинается словоизменение в конечное состояние - соответствующее словоформе в которую требуется преобразовать исходную словоформу.

Для решения задач словоизменения/словообразования так же можно использовать синхронизированное линейное дерево. В этом случае оно будет выступать как переключательная цепь, коммутирующая возбуждение, переводящее субавтомат из одного состояния в другое. Переключение состояний субавтомата будет происходить при подаче на вход синхронизированного линейного дерева специальных команд. Эти команды будут распознаны синхронизированным линейным деревом и преобразованы в градиентное значение на выходе соответствующих им нейронов-эффекторов, что вызовет возбуждение или торможение нейронов, соответствующих состояниям леммы.

Внутренняя структура словоформы в синхронизированном линейном дереве будет выглядеть следующим образом.

Таблица1. Обозначения во внутренней структуре словоформы.

Символ	Обозначение
#	пробел между словами
!	начало слова
	конец слова
\$	признак словоформы (падеж, число, спряжение и т. д.)

Рассмотрим пример для слова «бала» (лемма) - «ребенок» и двух его словоформ «балам»- «мой ребенок», «балаң»- «твой ребенок» (в казахском языке одушевленные существительные изменяются по лицам с помощью личных окончаний). Рецептор возбуждается на символ начала слова «!». Далее переходит в состояние «б». " при подаче символа «б», далее последовательно «!ба», «!бал», «!бала» , и затем одновременно два субсостояния «!балам|»и «!балаң|»(рис.1).

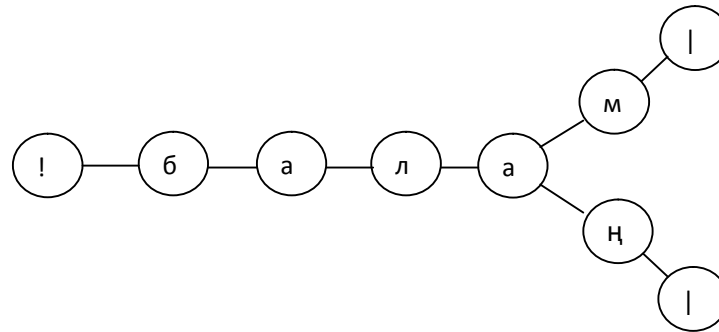


Рис.1. Представление словоформ в виде синхронизированного линейного дерева.

Рассмотрим признаки словоформ. Символом «*» обозначим первый спецсимвол признака словоформы, выступающий в той же роли, что и символ «!» для словоформ. Начало словоформы и ее признака обозначаем различными спецсимволами для уменьшения размера дерева поиска, это может увеличить скорость работы последовательной вычислительной системы. Однако, для решения задачи словоизменения на параллельной вычислительной системе, было бы достаточно ограничиться различием спецсимволов «\$» и «¶». На рисунке 2 приведен пример структуры связей леммы, определяющей следующие признаки: зат есім (имя существительное) – «*зе\$», жіктік жалғау (личное окончание), бірінші бет (1 б на рисунке, переводится как первое лицо), жіктік жалғау (личное окончание), екінші бет (2 б на рисунке, переводится как второе лицо). При подаче на лемму слова "балам_" она переходит в возбужденные субсостояния: «!балам|», «*зе\$», «*ж1бже\$», а при подаче слова «балаң» в возбужденные субсостояния: «!балаң|», «*зе\$», «*ж2бже\$». [2].

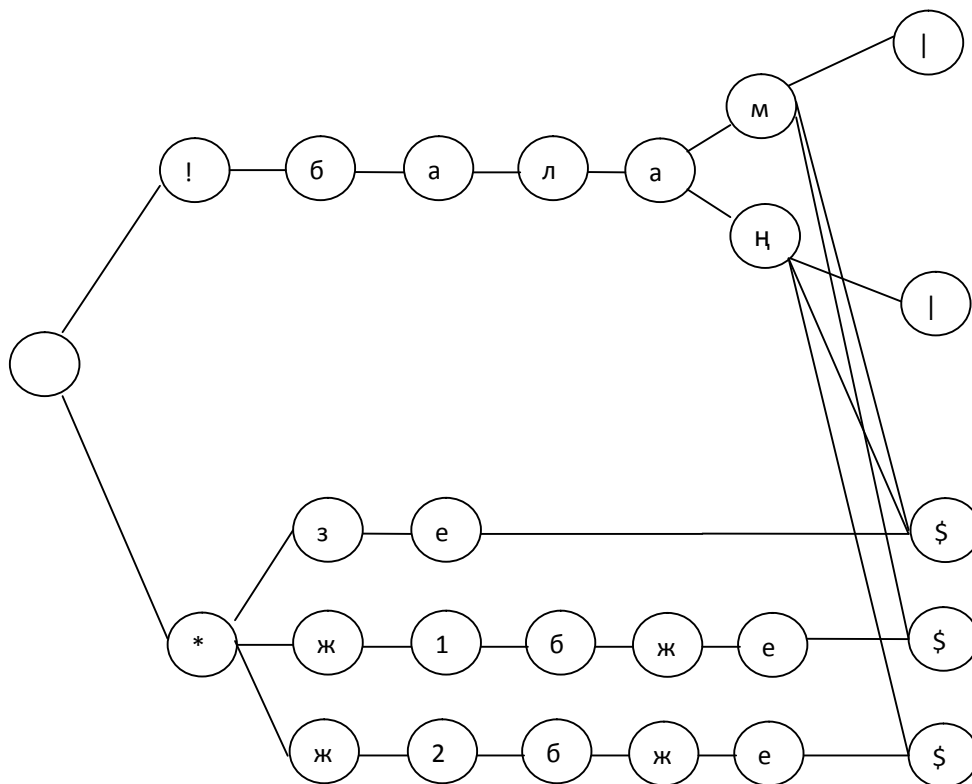


Рис. 2. Представление словоформ и их морфологической информации в виде синхронизированного линейного дерева.

Данная модель может использоваться как в задачах словообразования, так и в задачах морфологического анализа, а также для синтаксического разбора.

Данное дерево строится на основе таблицы семантических признаков, которая содержит признаки для разных частей речи. Например, одушевленность-неодушевленность для

существительных определяет разные траектории словообразования для одушевленных и неодушевленных существительных.

Пример 2. «Мен Сашамын» можно перевести в контексте «Я Саша». Добавляется жіктік жалғау личное окончание первого лица, поскольку Саша – это одушевленное существительное.

К неодушевленным существительным личное окончание не добавляется.

Также в качестве семантических признаков можно выделить возможность сочетания глаголов для сложной формы настоящего времени, образование степеней прилагательных и т. д. Все эти признаки показаны в семантической таблице .

semantyc : таблица									
	word	zat_es	syn_es	san_es	esim	ust	etis	z_e_1	z_e_2
►	Абай	☒	☐	☐	☐	☐	☐	☐	☐
	абайла	☐	☐	☐	☐	☐	☒	☐	☐
	автобус	☒	☐	☐	☐	☐	☐	☐	☐
	автокөлік	☒	☐	☐	☐	☐	☐	☐	☐
	автотұрақ	☒	☐	☐	☐	☐	☐	☐	☐
	ағаш	☒	☐	☐	☐	☐	☐	☐	☐
	ағылшын	☒	☐	☐	☐	☐	☐	☐	☐
	адам	☒	☐	☐	☐	☐	☐	☐	☐
	адас	☐	☐	☐	☐	☐	☒	☐	☐
	азаматтық	☐	☒	☐	☐	☐	☐	☐	☐
	аздап	☐	☐	☐	☐	☒	☐	☐	☐
	азық-түлік	☒	☐	☐	☐	☐	☐	☐	☐
	ай	☒	☐	☐	☐	☐	☐	☐	☐
	айлық	☐	☒	☐	☐	☐	☐	☐	☐
	айналыс	☐	☐	☐	☐	☐	☒	☐	☐
	айналыса	☐	☐	☐	☐	☐	☒	☐	☐
	айтқанның	☐	☐	☐	☐	☐	☒	☐	☐
	айтпақшы	☐	☐	☐	☐	☐	☐	☐	☐

Рис. 3. Семантическая таблица

Данная таблица содержит логические поля, принимающие только два возможных значения «истина», «ложь». Эта таблица является таблицей истинности для описанной нейронной модели.

Список литературы

[1] Туймебаев Ж. К. Казахский язык. Грамматический справочник. – Алматы: Ғылым, 2004.- 108 с.

[2] Шуклин Д.Е. Структура семантической нейронной сети, реализующей морфологический и синтаксический разбор текста