OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.89

ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ ПРЕДСТАВЛЕНИЯ МОРФОЛОГИЧЕСКИХ ПРАВИЛ КАЗАХСКОГО ЯЗЫКА В ВИДЕ СЕМАНТИЧЕСКИХ ГИПЕРГРАФОВ

Шарипбаев А.А., Бекманова Г.Т., Муканова А.С., Ергеш Б.Ж.

[*] *Евразийский национальный университет имени Л.Н. Гумилева, г. Астана, Казахстан*

sharalt@mail.ru

gulmira-r@mail.ru

asel_ms@bk.ru

saturn_banu@mail.ru

В данной работе построены онтологические модели морфологических правил казахского языка в виде семантических гиперграфов. В этих графах вершины представляют семантические признаки (морфологические понятия), а ребра – отношения между этими признаками. С помощью такого гиперграфа описываются структуры словоформ представлены в виде деревьев, которые преобразуются в линейные скобочные записи. Линейные скобочные записи являются формальными моделями морфологических правил. Программная реализация линейных скобочных записей позволили автоматизировать синтез всевозможных словоформ морфологический анализ казахского языка.

Ключевые слова: онтология, морфологические правила, семантический гиперграф, морфологический анализ.

Введение

Агглютинативные языки (от лат. Agglutinatio — приклеивание) — языки, имеющие строй, при котором доминирующим типом словоизменения является агглютинация («приклеивание») различных формантов (суффиксов или префиксов), причём каждый из них несёт только одно значение.

Агглютинативные языки — тюркские, финно-угорские, монгольские, тунгусо-маньчжурские, корейский, японский, часть индейских и некоторые африканские языки.

Казахский язык относится к тюркской группе языков и характеризуется большим числом словоформ для каждого слова, образованных путем добавления к его концу суффиксов и окончаний. Для него определен строгий порядок аффиксов. Вначале к корню слова прибавляются суффиксы затем окончания множественности, притяжательные окончания, падежные окончания, окончания спряжения [Қазақ грамматикасы, 2002].

В настоящее время онтология является мощным и распространенным инструментом моделирования отношений между объектами различных предметных областей. Принято классифицировать онтологии по степени зависимости от задач или прикладной области, по модели представления онтологических знаний и его выразительным возможностям и другим параметрам [Gruber,1993][Gruber1995]. Прикладные онтологии описывают концепты, которые зависят как от онтологии задач, так и от онтологии предметной области.

Прикладная онтология разрабатывается на основе общих принципов построения онтологий, но с учетом использования в качестве модели представления знаний семантических гиперграфов. Данный формализм позволят определить онтологию O в виде тройки: (V, R, K) где V – множество понятий проблемной среды (вершины гиперграфа), R – множество отношений между понятиями (дуги и ребра гиперграфа), а K – множество имен понятий и отношений в данной предметной области.

Одним из формальных средств представления знаний является язык семантических гиперграфов, в котором можно в зависимости от типов связей реализовывать классифицирующие, функциональные, ситуационные, структурные сети и сценарии. Семантический гиперграф является расширением семантических сетей, где естественным образом представляются n-арные отношения, которые позволяют задавать не только

атрибуты объектов, но и представлять их структурные, «целостные» описания [Berge, 1985] [Vizing, 2007].

Гиперграф H определяется парой (V,R), где $V=\{v_i\}-$ множество вершин; $R=\{r_j\}-$ множество ребер; $i=1,2,3..n$; $j=1,2,3..m$; каждое ребро представляет собой пару из элементов множества V, т.е. $r=\{(v_{j_t},v_{j_s})\}$, $j_s\neq j_t, s,t-$ натуральные числа.

1. Онтологические модели морфологических правил

1.1. Имя существительное

Для имен существительных в качестве семантических признаков начальных форм выступают одушевленность (jand) и неодушевленность (jans) имен существительных. В зависимости от этого признака и определяется траектория словоизменения имени существительного.Имя существительное в казахском языке спрягается (jikt) и изменяется по падежам (sept), а также числам (kopt) и имеет притяжательную форму (taul). На рисунке 1 показана онтологическая модель имени существительного с учетом семантических признаков.

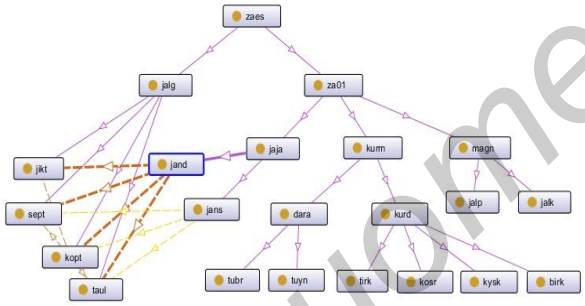

Рисунок 1 – Онтологическая модель имени существительного

Если представить онтологию в виде гиперграфа, то его вершины и ребра будут:

V={jand, jans},

E ={(jand, jikt), (jand, sept), (jand, kopt), (jand, taul), (jans, sept), (jans, kopt), (jans, taul)}.

Из указанного семантического гиперграфа можно получить формальные правилас помощью скобочной записи. Количество формальных правил для имени существительного 4500. Далее для одного одушевленного существительного с помощью формальных правил автоматически генерируется 93 словоформы (словарных статей), а для неодушевленного существительного генерируется 82 словоформы. Также имя существительное возможно образовать из других частей речи.

Пример словоизменения одушевленного существительного «бала» - «ребенок» содержит все словоформы данного существительного и их морфологическую информацию, которая содержит в сокращенном обозначении информацию о том в каком числе, падеже находится существительное, от какого лица происходит действие и его принадлежность тому или иному лицу. В таблице 1 приведено изменение существительного «бала» в дательном падеже.

Таблица 1 – Изменение существительного «бала» в родительном падеже.

Правила	Пример	Объяснение
((зежа01)ға)!ба	((бала)ға)	((**зе-** существительное,**жа-** одушевленное, **01**- признак твердости, **ға-** окончание дательного падежа))**! ба** – родительный падеж

1.2. Имя прилагательное

Для имен прилагательных в качестве семантических признаков начальных форм выступают возможности образования из него сравнительной и/или превосходной степени прилагательного, а также выступают такие признаки как относительные (kats) и качественные (sapa),простые(dara), сложные (kurd), производные (tuyn) и т.д. Онтологическая модель имени прилагательного представлена на рисунке 2.

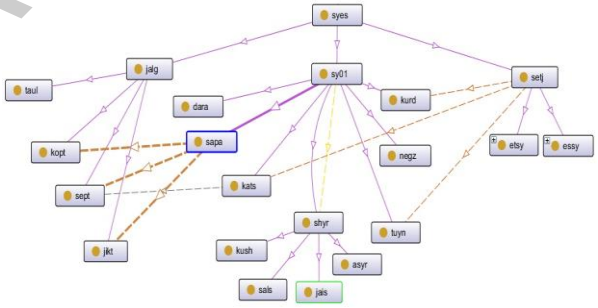

Рисунок 2 – Онтологическая модель имени прилагательного

Определить возможно ли из данного прилагательного образовать сравнительную степень прилагательного и с помощью каких конкретно суффиксов, может только эксперт. Это касается и возможности использования вспомогательных слов при образовании превосходных степеней прилагательных. В данном случае разметку семантических признаков в базе знаний осуществлял специалист-лингвист.

Если представить онтологию в виде гиперграфа, то его вершины и ребра будут:

V={kats, sapa},

E ={(kats, sept), (sapa, sept), (sapa, kopt), (sapa, jikt)}.

Разряды прилагательного субстантивируясь, изменяется по падежам, спрягается по лицам и принимает аффиксы принадлежности. В таблице 2

приведено изменение прилагательного «ақылды» - **«умный».**

Таблица 2 – прилагательное «ақылды» - «умный».

Правила	Пример	Объяснение
((сы01)ның)!ілл	(ақылды)ны- «умного»	((**сы-** имя прилагательное, **01-** признак твердости), **ның-** окончание родительного падежа))! **ілл-** родительный падеж
((сы01)мын)!жі11	(ақылды)мын -«я умный»	((**сы-** имя прилагательное, **01-** признак твердости), **мын-** окончание спряжения ед.ч., 1 лица)! **жі11-** спряжение. ед.ч. 1 лицо

С помощью добавления 135 суффиксов образуется имя прилагательное из других частей речи. В результате из 40000 слов словаря генерируется 66000 прилагательных.

1.3. Имя числительное

Имена числительные по составу в казахском языке разделяют на простые (dara) и сложные (kurd). Например, простые: бір - один, он- десять, жүз-сто, мың-тысяча;сложные: он бес- пятнадцать, бес жүз- пятьсот, елу мың - пятьдесят тысяч.

Словообразование сложных имен числительных возможно реализовать автоматически, поскольку в большинстве случаев они образуются из простых числительных путем всевозможных сочетаний разряда числительного и простых числительных.

Пример. Образование сложного числительного «он бір» - «одиннадцать» происходит путем соединения числительных «он» - «десять» и «бір» - « один». Сложное числительное «жүз он бір» -«сто одиннадцать» происходит путем соединения числительных «жүз» - «сто», «он» - «десять» и «бір» - «один» .

По значению имена числительные делятся на шесть групп, которые образуются из простого или сложного числительного путем присоединения соотвествующих окончаний или суффиксов. В большинстве случаев все группы числительных автоматически образуются из количественных путем присоединения суффиксов. Онтологическая модель имени числительного представлена на рисунке 3.

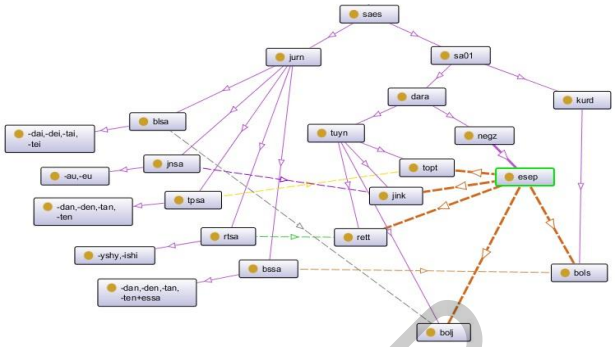

Рисунок 3 – Онтологическая модель имени числительного

Если представить онтологию в виде гиперграфа, то его вершины и ребра будут:

V={esep, jurn},

E={(esep, topt), (esep, jink), (esep, rett), (esep, bolj), (esep,bols)}.

1.4. Глагол

Глагол наряду с именем существительным сложная для словообразования и словооизменения часть речи. Словообразование и словооизменение происходит как автоматически, так и по результатам заполненной лингвистом базы знаний.

Необходимо отметить, что глаголы словоизменяются по лицам и числам, а также происходит словообразование новых видов глаголов из других частей речи. Онтологическая модель глагола представлена на рисунке 4.

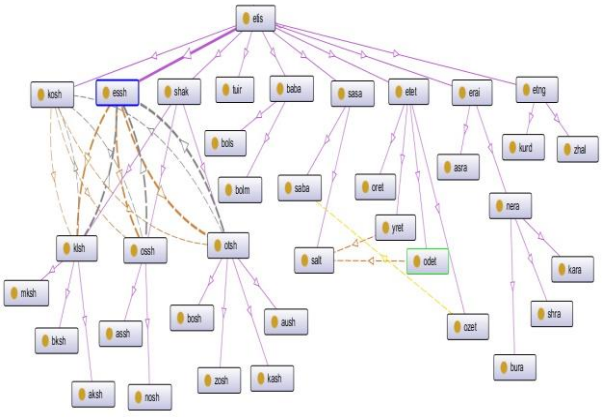

Рисунок 4 – Онтологическая модель глагола

Если представить онтологию в виде гиперграфа, то его вершины и ребра будут:

V={essh,kosh,saba,salt},

E={(essh, klsh), (essh, ossh), (essh, otsh), (kosh, klsh), (kosh, ossh), (kosh, otsh),(saba,oset), (salt,yret), (salt,odet)}.

Ниже приведены формальные правила словоизменения и словообразования глагола:

((((етот01)п) отыр)мын)!окжі11

((((етот01)п) отыр)мыз)!окжі11
((((етот01)п) отыр)сың)!окжі22
((((етот01)п) отыр)сыңдар)!окжі22
((((етот01)п) отыр)сыз)!окжі2*
((((етот01)п) отыр)сыздар)!окжі2*
((((етот01)п) отыр))!окжі33
((((ет01)т)қыз)дыр)!өг
((((ет01)т)қыз)ғыз)!өг
(((((ет01)т)тыр)т)қыз)!өг
((((((ет01)т)тыр)т)қыз)дыр)!өг
(((((ет01)т)қыз)дыр)т)!?г
(((((((ет01)т)тыр)т)қыз)дыр)т)!өг
(((ет01)т)ты)!өе
(((ет01)т)қан)!өе
(((ет01)т)атын)!өе
(((ет01)т)ар)!ке

С помощью формальных правил образуются новые глаголы и отглагольные формы из других частей речи. В результате из 40000 слов словаря генерируется 395000 глаголов.

1.5. Местоимение

В качестве семантических признаков начальных форм для местоимения является деление в зависмости от его значения на 7 групп: личные, указательные, вопросительные, возвратные, неопределенные, отрицательные, определительные.

Онтологическая модель местоимений представлена на рисунке 5.

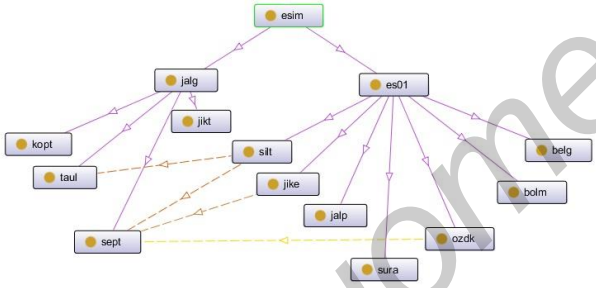

Рисунок 5 – Онтологическая модель местоимений

Личные, указательные и возвратные местоимения склоняются по правилам, определенным для каждой группы. Местоимения других групп склоняются лишь частично, некоторые местоимения вообще не склоняются. Для них определен семантический признак склонения или несклонения. Притяжательная форма (тәуелдік жалғау (taul)) и форма спряжения (жіктік жалғау (jikt)) существует только для некоторых местоимений.

Таким образом, в качестве семантических признаков местоимений можно выделить склонение местоимений, существование притяжательной формы, формы спряжения и принадлежность к той или иной группе местоимений.

Если представить онтологию в виде гиперграфа, то его вершины и ребра будут:

V={silt,jike,ozdk},

E ={(silt, taul), (silt, sept), (jike, sept), (ozdk, sept)}.

ЗАКЛЮЧЕНИЕ

Построены онтологические модели морфологических правил казахского языка, что позволило записать формальные правила словоизменения и словообразования каждой части речи. Программная реализация этих правил позволила из 40 000 начальных форм слов с размеченными семантическими признаками автоматически генерировать более 3 200 000 словоформ (словарных статей). Тем самым создан морфологический анализатор казахского языка, который сохраняет в памяти только небольшое количество начальных форм слов с размеченными семантическими признаками, а все возможные правильные словоформы получаются автоматически в соответствии с формальными правилами. Результаты работы будут использованы при создании всевозможных программ обработки казахского языка (трансляторов, семантических поисковиков, речевых технологий и др.).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

[Қазақ грамматикасы, 2002] Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис// Астана-2002., С. 20-25

[Gruber, 1993] Gruber T.R. A Translation Approach to Portable Ontology Specifications / Gruber T.R.// Knowledge Acquisition, 1993, P.199-220

[Gruber, 1995] Gruber T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing / Gruber T.R. // International Journal Human-Computer Studies. – 1995, - Vol. 43 - P.907-928

[Berge, 1985] Graphs and Hypergraphs / Berge C.C.; Elsevier Science Ltd., 1985

[Vizing, 2007] Vizing V.G. About a coloring of intsidentor in the hypergraph / Vizing V.G. // Diskretn. Anal. Issled. Oper., Ser., 2007, № 1, P. 40-45

ONTOLOGICAL MODELS OF MORPHOLOGICAL RULES OF KAZAKH LANGUAGE IN THE FORM OF SEMANTIC HYPERGRAPHS

Sharipbayev A.A., Bekmanova G.T., Mukanova A.S., Yergesh B.Zh.

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

sharalt@mail.ru

gulmira-r@mail.ru

asel_ms@bk.ru

saturn_banu@mail.ru

This paper illustrates construction of ontological models of morphological rules of Kazakh language in the form of semantic hypergraphs. In these graphs the nodes represent the semantic features, and the edges represent the relationship between these features. With such a hypergraph structure of word forms are described in the form of trees, which are converted into linear parenthesis notation. Linear parenthesis notations are the formal models of morphological rules.