КАЗАХСКО-АНГЛИЙСКИЙ ПЕРЕВОД С ИСПОЛЬЗОВАНИЕМ ГРАММАТИКИ СВЯЗЕЙ И ПРАВИЛ СЕМАНТИЧЕСКИХ СИТУАЦИЙ

Тукеев У.А., Жуманов Ж.М., Картбаев А.Ж.
Казахский Национальный Университет им. аль-Фараби, Алматы, Казахстан,
ualsher.tukeyev@kaznu.kz, z.zhake@gmail.com, a.kartbayev@gmail.com

Введение

Основными подходами к компьютерному переводу являются: подход, основанный на правилах (rule-based approach), статистический подход (statistical approach), подход на основе примеров (example-based approach), подход «память переводов» (translation memory approach), гибридный подход (hybrid approach). Подход, основанный на правилах, предполагает следующие этапы обработки текста: морфологический анализ, синтаксический анализ, перевод, разрешение неоднозначности. Морфологический анализ казахского языка описывается регулярными грамматиками, синтаксический анализ можно описывать контекстно-свободными грамматиками, перевод осуществляется с использованием правил соответствий. Однако это не единственные инструменты обработки естественных языков, которые могут быть использованы в компьютерного перевода. В данной работе предлагается для синтаксического анализа использовать грамматику связей, а правила соответствий перевода дополнить моделью семантической онтологии. Помимо описания предлагаемых инструментов приводится пример их использования в казахско английском переводе .

Синтаксический анализ с использованием грамматики связей

Грамматика связей — это теория синтаксического анализа, созданная Д. Темперли (Davy Temperley), Д. Слитором (Daniel Sleator) и Дж. Лаферти (John Lafferty) из Университета Карнеги-Мелон. Данная грамматика определяет связи между парами слов предложения, но в отличие от традиционного подхода к синтаксису не пытается построить их в полное дерево разбора. Основными параметрами для данной грамматики являются направленность связей и расстояние между связанными парами слов. [1]

В лингвистике существует классификация языков на основе типологии порядка слов в предложении. [2][3] Она основывается на порядке, в котором подлежащее (subject), сказуемое (verb) и прямое дополнение (object) стоят в предложении. Согласно данной классификации существует 6 возможных типов языков:

SVO — Subject Verb Object
SOV — Subject Object Verb
VSO — Verb Subject Object
VOS — Verb Object Subject
OSV — Object Subject Verb
OVS — Object Verb Subject

Казахский язык относится к типологии SOV. Это значит, что в абсолютном большинстве предложений казахского языка подлежащее и дополнение будут связаны со сказуемым справа, а сказуемое будет связано и с подлежащим, и с дополнением слева. Расстояние связи между парой подлежащее-сказуемое будет больше, чем расстояние между парой дополнение-сказуемое.

Разбор предложения с использованием грамматики связей лучше всего пояснить на примере. Возьмем предложение: «**Марат жаңа кино көрді»**. Для этого предложения можно составить следующие правила построения связей:

```
<subject>:   S+;
<adjective>: A+;
<object>:    A- & O+;
<verb>:      S- & O-;
```

Они означают, что в предложениях данного типа элемент <subject> может иметь 1 связь типа S, направленную вправо, <object> может иметь 1 связь типа A слева и 1 связь типа O справа и т.д.

Результат разбора предложения имеет следующий вид:

```
   +-------S-------+
   |      +--A-+--O-+
   |      |    |    |
Марат жаңа кино көрді.
```

Архитектура парсера грамматики связей достаточно проста. Он состоит из грамматического файла, описывающего слова анализируемого языка, и непосредственно анализатора. Часто грамматический файл разбивают на несколько файлов для удобства сопровождения. Анализатор в настоящее время разрабатывается в рамках проекта OpenSource текстового редактора AbiWord. Он доступен под открытой лицензией. Однако команда проекта поддерживает разработку грамматического файла только для английского языка. Таким образом, главной задачей при использовании реализации грамматики связей для остальных языков является формализация грамматической модели языка и представление ее в формате, принятом в грамматике связей.

Разработка грамматики связей для нового языка состоит из следующих этапов [4]:

- установка и настройка анализатора грамматики связей;
- создание грамматического файла для нового языка;
- расширение и отладка грамматических данных в файле.

Пример элементарного грамматического файла для казахского языка, отвечающего за разбор предложения «Адамдар жазады» представлен ниже.

```
"адамдар.nnp" "аттар.nnp" "балапандар.nnp":      % Сущ-е во множественном числе
S+;

"жазады.vb" "барады.vb" "келеді.vb":       % ukfujks
S- & {W-};

LEFT-WALL: W+;
```

Связи, направленные вправо обозначаются знаком «+» (плюс), влево - «-» (минус). Необязательные связи окружаются фигурными скобками **{…}**. Нежелательные связи окружаются квадратными скобками **[]**. Множественные связи сочетаются между собой с помощью конъюнкции **«&»** или дизъюнкции **«or»**.

Помимо вышеописанного к парсеру грамматики связей предъявляется ряд дополнительных требований:

- *Соблюдение условия проективности:* связи между словами не должны пересекаться. Правильным считается результат, в котором линии, обозначающие связи между словами, не пересекаются между собой.

- *Соблюдение условия связности:* в разобранном предложении не должно быть изолированных слов или групп слов.

- *Соблюдение условия полноты требований:* в результате разбора для каждого слова в предложении выполнены все условия на связи, с учетом дизъюнкций и конъюнкций.

Результатом работы парсера грамматики связей являются связи между словами предложений. Выбор названий этих связей зависит от разработчика. В том случае если задать в качестве возможных связей роль слов в предложении, то после работы парсера мы получим разобранное предложение.

Использование правил семантических ситуаций в переводе казахского языка на английский

В процессе компьютерного перевода, не зависимо от используемых подходов, происходит преобразование текста из одного языка в другой. Это преобразование подразумевает преобразование лексики, синтаксиса, контекста и семантики исходного текста. Описание этих преобразований представляет собой соответствия конструкции исходного языка конструкциям целевого языка. Создание подобных описаний предполагает создание онтологии языка.

В информационных технологиях и компьютерных науках под онтологией подразумевается явная, спецификация концептуализации, где в качестве концептуализации выступает описание множества объектов и связей между ними. [5] Онтология — это попытка всеобъемлющей и детальной формализация некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из структуры данных, содержащей все релевантные классы объектов, их связи и правила (теоремы, ограничения), принятые в этой области.

Формально на наиболее общем уровне онтология определяется как: O = <X,R,F>,
где: X — конечное множество понятий предметной области,
R — конечное множество отношений между понятиями,
F — конечное множество функций интерпретации.

Онтологию естественного языка можно представить как сочетание грамматики языка с правилами соответствия грамматических конструкций. Определение грамматики языка традиционно и чаще всего имеет структурный характер. Употребление грамматических конструкций любого языка неотделимо от смысла, который в них вкладывается и контекста использования. Однако в каждом языке можно выделить более или менее стандартный набор семантических ситуаций, которые имеют одинаковое смысловое значение во всех языках. В различных ситуациях используются различные грамматические конструкции. Поэтому описание онтологии языка следует дополнить набором правил соответствий грамматических конструкций различных языков одним и тем же семантическим ситуациям.

Таким образом, мы можем качестве понятий предметной области использовать грамматику языка, в качестве функций интерпретации — описания семантических ситуаций, а в качестве отношений между понятиями определим связи между правилами грамматики и семантическими ситуациями.

В итоге онтологию языка мы можем представить следующим набором:

$$O = <G, Sit, L>,$$

где: G — описание грамматики языка;
Sit — набор семантических ситуаций;

L — набор связей между G и Sit.

G представляет собой набор структурных грамматик, описывающих лексику, фразовую структуру и синтаксис языка.

$$G = <G_L, G_P, G_S>$$

Поэтому онтологию языка можно представить в следующем виде:

$$O = (G_L, G_P, G_S, Sit, L).$$

Для задания каждого из составляющих онтологии предлагается использовать следующие аппараты: G_L, G_P, G_S: структурные грамматики, описанные в форме Бэкуса-Науэра. Их использование широко освещено в литературе. В дополнение к общепринятому описанию каждому правилу вывода грамматик предлагается назначить уникальный идентификатор (номер или индекс) с помощью которого его можно будет связывать с семантическими ситуациями.

$$G = <T, N, S, (R\_ID, R)>,$$

где: T — множество терминальных символов;
N — множество нетерминальных символов;
S — начальный символ грамматики;
R_ID — идентификаторы правил вывода;
R — правила вывода.

Sit: множество, задаваемое набором своих членов и дополненное соответствующим ему множеством идентификаторов. Подробнее они описаны ниже.

L: множество пар идентификаторов ситуаций (S_ID) и идентификаторов правил (R_ID). Причем каждый идентификатор ситуаций может входить в множество L один и только один раз.

В каждом естественном языке одним из ключевых компонентов является смысл вкладываемый во фразы и выражения. Независимо от используемого языка а аналогичных контекстах используются выражения с аналогичным смыслом но, зачастую, с различными грамматическими конструкциями. Примерами таких контекстов может быть «приветствие», описание родственных связей между людьми, использование порядковых числительных и т.п. Вполне возможно составить набор семантических ситуаций, которые имеют одинаковое смысловое значение во всех языках. Поэтому описание онтологии языка следует дополнить набором соответствий грамматических конструкций различных языков одним и тем же семантическим ситуациям. Пример такого набора для казахского и английского языков представлен в следующей таблице.

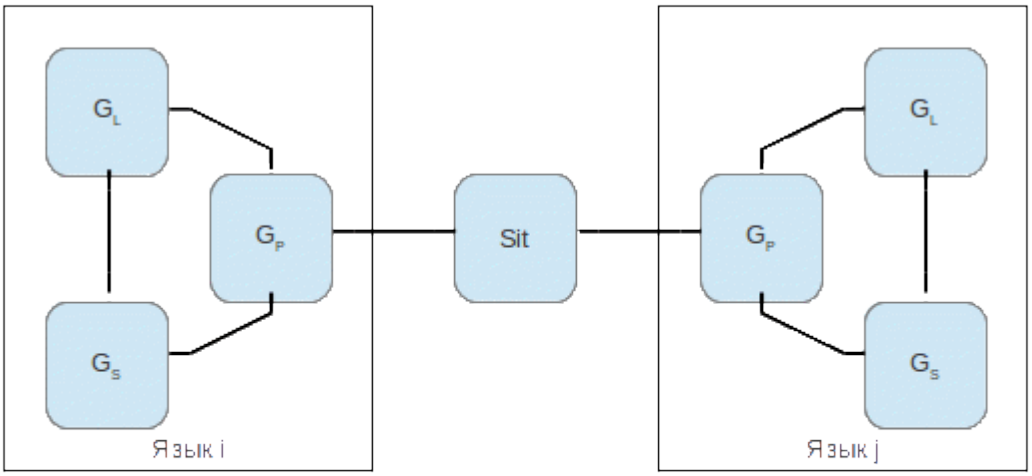
Таблица 1 — Примерный набор семантических ситуаций для казахско-английского перевода

Казахский язык	**Семантическая ситуация**	**Английский язык**
Основа + «-у» *(оқ + у)*	Неопределенная форма глагола (инфинитив)	«To» + основа *(to read)*
Основа *(ал)*	Повелительная форма глагола	Основа *(take)*

Основа + «-лар», «-лер», «-дар», «-дер», «-тар», «тер» *(кітаптар)*	Множественное число существительных	Основа + «-s», «-es», «-ies» *(books)* Имеются исключения (child - children)
Количественное числительное + существительное в единственном числе	Выражение множественности (числительное + существительное)	Количественное числительное + существительное во множественном числе
Основа + «-ма», «-ме», «-ба», «-бе», «-па», «-пе»	Отрицательная форма глагола	Not + основа
Обстоятельство места + подлежащее + «бар»	Наличие чего-либо где-либо	«there is» + дополнение + обстоятельство места
Слово + «және» + слово Слово + «да», «де», «та», «те» Слово + «мен», «бен», «пен" + слово	Сочинительный союз «и»	Слово + «and» +слово
Сказуемое + «ма», «ме», «ба», «бе», «па», «пе»	Общие вопросы (касаются всего предложения в целом)	Образование вопросительных предложений по правилам времен
Основа + «-йін», «-йын», «-айін», «-айын» «-йік», «-йық», «-айік», «-айық»	Повелительное наклонение 1 лица	Let me + «основа» Let us + «основа»
Основа + притяжательные окончания Притяжательное местоимение + основа + притяжательные окончания	Притяжательность	Притяжательное местоимение + основа

Пример перевода предложения с казахского языка на английский язык

Применительно к компьютерному переводу грамматику связей и семантические ситуации можно использовать для осуществления преобразования из одного языка в другой. (Рисунок 1). Грамматика связей на рисунке представлена блоком G_S.



G_L — грамматика, описывающая лексику
G_P — грамматика, описывающая фразовые структуры
G_S — грамматика, описывающая синтаксис
Sit — набор семантических ситуаций

Рисунок 1 — Применение онтологии в процессе компьютерного перевода

Особенности процесса перевода представлены на рисунках 2 и 3. Морфологический анализ осуществляет разбор слов предложений. Синтаксический анализ, выполненный с использованием грамматики связей. Результат работы грамматики связей (в частности определенная связь между словами сложного сказуемого («тапсыру қажет») используется для более точного перевода и использованием семантической ситуации, выражающей долженствование. (рисунок 3)

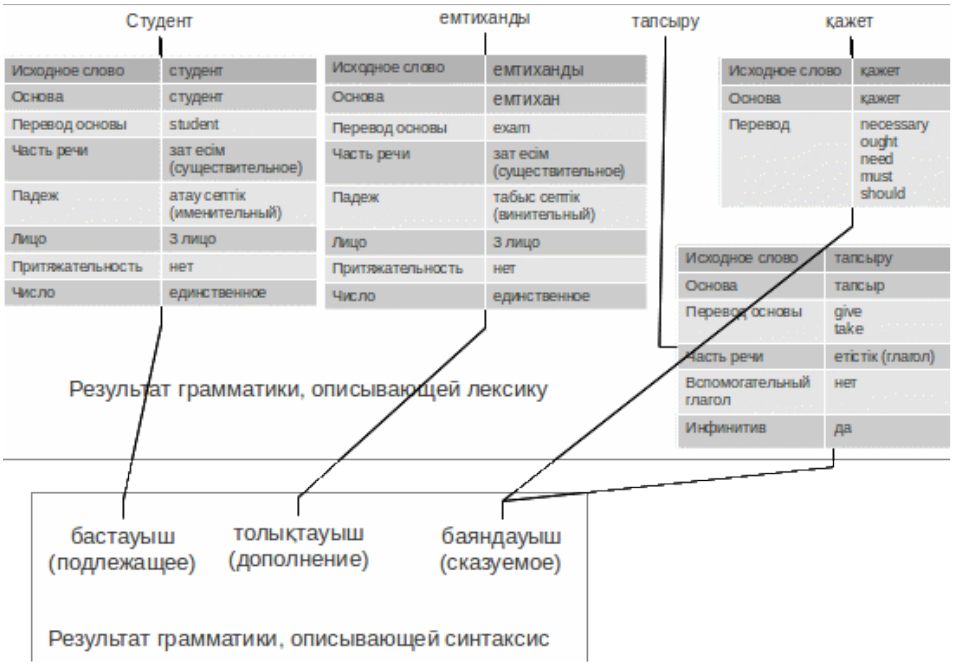


Рисунок 2 — Процесс компьютерного перевода: лексический, фразовый и синтаксический анализ

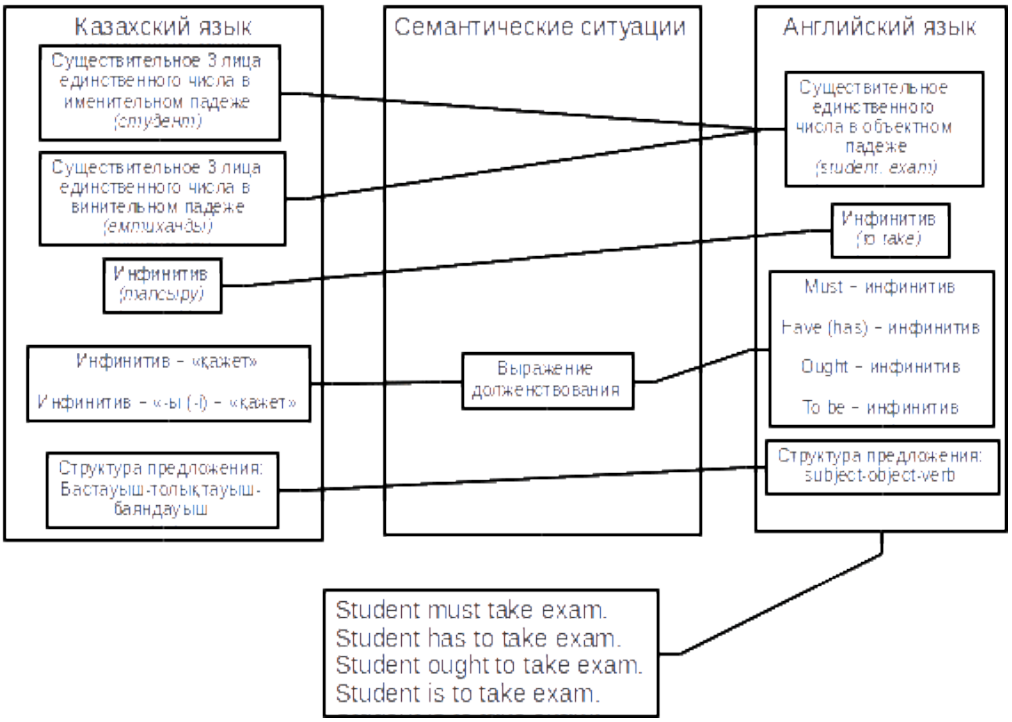


Рисунок 3 - Процесс компьютерного перевода: преобразование с использованием семантических ситуаций

Заключение

В данной статье показано каким образом возможно развить методы казахско-английского перевода с использованием грамматики связей и правил ёсемантических ситуаций. Из основных подходов к компьютерному переводу был выделен подход, основанный на правилах (rule-based approach). Далее к его этапам применяются указанные инструменты: для синтаксического анализа предлагается использовать грамматику связей, а правила соответствий перевода дополнить моделью семантической онтологии. Были описаны основные особенности предлагаемых инструментов и представлен пример осуществления перевода с их помощью.

Список использованных источников

1. Daniel Sleator and Davy Temperley. 1991. Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.
2. Matthew S. Dryer. Order of Subject, Object, and Verb .TWAC. 28 January 2005 .
3. Типология порядка слов. http://ru.wikipedia.org/wiki/SOV
4. Jon Dehdari.A Primer for Localizing Link Grammar.http://www.ling.ohio-state.edu/~jonsafari/link-grammar/primer.html
5. Raul CorazzonThe Place of ontology in modern philosophy. An overview.http://www.ontology.co