УДК 004.43

Жуманов Ж.М.
Казахский Национальный университет им. аль-Фараби, Алматы, Казахстан
e-mail: z.zhake@gmail.com

Разработка грамматики связи для синтаксического анализа казахского языка[1]

В данной статье кратко рассмотрены существующие модели описания синтаксиса естественных языков и подробно описана грамматика связей. Дано ее общее описание, описан пример применения грамматики связей казахского языка. Подробно описан процесс реализации парсера грамматики связей, и предложен способ автоматизации процесса формирования грамматического файла.

Ключевые слова: синтаксис естественных языков, грамматика связей казахского языка, неглубокий анализ, аффиксная грамматика, грамматика составляющих, грамматика зависимостей, статистический синтаксический анализ.

Синтаксический анализ — это процесс сопоставления последовательности слов языка с его грамматикой. Программа или часть программы, выполняющая синтаксический анализ, называется синтаксическим анализатором. При анализе исходный текст преобразуется в структуру данных, которая отражает синтаксическую структуру предложения и пригодна для дальнейшей обработки. Качество дальнейшей обработки исходного текста напрямую зависит от качества синтаксического анализа. Поэтому детальное изучение существующих методов его осуществления является важной задачей.

Для описания синтаксиса естественных языков можно использовать следующие теории:

- неглубокий анализ (shallow parsing);
- аффиксная грамматика;
- грамматика составляющих;
- грамматика зависимостей;
- грамматика связей;
- статистический синтаксический анализ (statistical parsing).

Краткое их описание и сравнение представлено в таблице 1.

Данная работа посвящена грамматике связей. Дается общее ее описание, описывается пример применения грамматики связей для предложения казахского языка и подробно описывается реализация парсера грамматики связей. Из-за особенностей лексики

[1]Работа выполнена при поддержке грантового финансирования научно-технических программ и проектов Комитетом науки МОН РК, грант № 709/ГФ, 2012г.-2014г.

казахского языка непосредственное использование парсера грамматики связей связано с большим объемом работы по заполнению грамматического файла. Поэтому также в данной работе предлагается способ автоматизации процесса формирования грамматического файла.

Таблица 1: Модели синтаксиса естественных языков

№	Модель синтаксиса	Характеристики
1	Неглубокий анализ (shallow parsing)	Идентифицирует компоненты предложения; проста в реализации; не определяет внутреннюю структуру компонентов предложения; не определяет роль компонентов в предложении; эффективна только для родственных пар языков, которые имеют похожий синтаксис.
2	Аффиксная грамматика	Используется для описания синтаксиса языков (в основном, языков программирования); основывается на интуитивном описании языка; может описывать некорректные предложения естественных языков.
3	Грамматика составляющих	Делит предложение на группы составляющих; используется в формальных моделях языка; при классификации групп учитывает части речи.
4	Грамматика зависимостей	Структура предложения рассматривается в терминах вершин и зависимых; отсутствуют фразовые узлы; не подходят для языков со строгим порядком слов в предложении.
5	Грамматика связей	Определяет связи между парами слов предложения; учитывает направленность связей и расстояние между связанными парами слов; основана на типологии порядка слов в предложении.
6	Статистический синтаксический анализ (statistical parsing.)	Правила грамматики связываются с вероятностью.

Общее описание 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]

Преимуществами грамматики связей являются:

- разработана для описания и анализа только естественных языков;

- не включает описание некорректных предложений;
- осуществляет разбор внутренней структуры предложений;
- сочетает характеристики грамматик составляющих и грамматик зависимостей;
- несложна в реализации.

Основным элементом грамматики связи является соединение (коннектор). Коннектор состоит из имени типа связи (например, S – подлежащее, O – дополнение и т.д.), в которую может вступать анализируемое слово, и суффикса, определяющего вектор направления соединения («+» право-направленный коннектор и «-» лево-направленный коннектор). Лево-направленный и право-направленный коннекторы одного типа образуют связь (соединение).Одному слову может быть приписана целая формула коннекторов, составленная с помощью следующих связок:

- & - несимметричная конъюнкция. Например, если слову W приписана формула A+ & B+ (W: A+ & B+), то некоторое слово X, с которым слово W образует связь A, должно стоять раньше по тексту, чем слово Y, с которым слово W образует связь B;
- or - дизъюнкция. Если W: A+ or B-, то слово W может образовывать либо связь A вправо, либо связь B влево.
- {} - факультативность. Если W: A+ & {B+}, то после того, как слово W образовало правую связь A, оно может образовывать или не образовывать связь B.
- @ - неограниченность означает, что связь может строиться неограниченное число раз.

Для разнонаправленных коннекторов конъюнкция симметрична: формулы W: A- & B+ и W: B+ & A- эквивалентны.

Пример разбора предложения казахского языка с использованием грамматики связей

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. [2]

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 одну связь типа S, направленную вправо, <object> может иметь одну связь типа A слева и одну связь типа O справа и т.д.

Результат разбора предложения имеет следующий вид (рисунок 1):

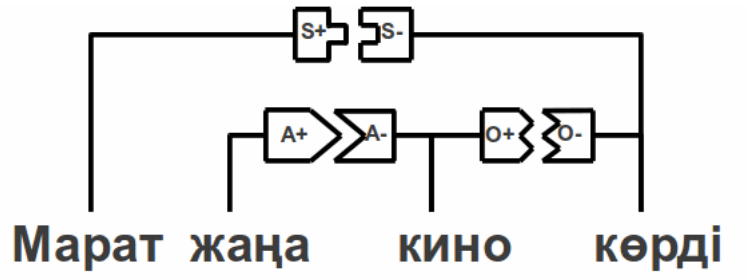

Рисунок 1. Пример разбора предложения грамматикой связей

Разработка грамматики связей для казахского языка

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: [3]

- 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": % глаголы
S- & { W- };
```

Большой проблемой при создании грамматических словарей для языков со сложной морфологией является учет всех возможных словоформ. Это особенно критично в казахском языке, который является агглютинативным языком. Грамматика связей не предназначена для морфологического анализа. При создании ее реализации для казахского языка необходимо использовать морфологические анализаторы на этапах создания и расширения грамматического файла.

Использование дополнительного морфологического анализатора помимо ускорения процесса создания файла позволяет также дополнить описание слов метками (тегами) грамматической информации. Они указываются сразу же после слова в грамматическом файле и отделяются точкой. Это повышает потенциал дальнейшего использования разработанного парсера грамматики связей в будущем в различных областях области обработки естественных языков. Например, при решении задачи машинного перевода.

Помимо использования морфологического анализатора, для облегчения процесса создания грамматического файла можно использовать методы корпусной лингвистики. Заполнять словарь на начальном этапе реализации грамматики лучше всего словами с наибольшей частотой использования в рассматриваемом языке. Для этого можно применить статистический анализ лингвистического корпуса языка, при его доступности. Создание подобного корпуса не относится к проблеме реализации грамматики связей. Хотя его наличие способно значительно улучшить ее качество.

Таблица 2. Коннекторы грамматики связей для казахского языка

Тип коннектора	**Значение связи**
A	связь определение - определяемое
AA	связь определение - определяемое, когда определяемое само является определением
S	связь подлежащее - сказуемое
O	связь дополнение - сказуемое
P	связь обстоятельство - сказуемое
N	связь между словом и отрицательной частицей
Q	связь между словом и вопросительной частицей
F	связь со словом «үшін»
VV	связь вспомогательный глагол - глагол
C	связь с союзом
AS	превосходная степень («өте», «ең»)
E	связь между однородными членами предложения
W	связь со вспомогательными словами

Разработка собственного анализатора грамматики связей с учетом вышеописанных особенностей не является сложной задачей. Однако помимо указанного к нему добавляется ряд 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.*

Достаточно вероятна ситуация, когда для одного предложения возможно несколько вариантов разбора. В общем случае все эти варианты можно считать равнозначными и для простоты реализации в качестве результата выдавать первый найденный корректный вариант разбора. В тех редких ситуациях, когда варианты считать равнозначными нельзя, имеет смысл разработать собственный вариант анализатора грамматики связей с учетом этих дополнительных требований.

Как было указано выше, главным элементом грамматики связей является соединение (коннектор). Для казахского языка был составить следующий список коннекторов (таблица 2):

С использованием перечисленных коннекторов можно составить грамматический файл для осуществления синтаксического анализа казахского языка. Для составления грамматического файла проще всего воспользоваться предложениями с типовой грамматической структурой.

Приведем несколько примеров.

Предложение «Интернет мыңдаған корпоративті, үкіметтік, ғылыми және үй желілерінен құралған». Грамматический файл для его разбора будет иметь следующий вид:

```
«интернет» : {@A-} & S+;
«мыңдаған» : AA+;
«корпоративті» «үкіметтік» «ғылыми» «үй»: {@AA-} & {C-} & {C+} & A+;
«және» : C- & C+;
«желілерінен» : {@A-} & O+;
«құралған» : {@O-} & S-;
```

Предложение «Интернет — компьютерлік серверлердің бүкіләлемдік желісі». Грамматический файл для его разбора будет иметь следующий вид:

```
«интернет» : {@A-} & S+;
«компьютерлік» : {A+} & {AA+};
«серверлердің» : {AA-} & A+;
«бүкіләлемдік» : A+;
«желісі» : {A-} & S-;
```

Предложение «Ғаламтор — еренсілтемелерді пайдалана отырып Интернетті қарап шығу жүйесі». Грамматический файл для его разбора будет иметь следующий вид:

```
«ғаламтор» : {@A-} & S+;
«еренсілтемелерді» : O+;
«пайдалана отырып» : P-;
«интернетті» : O+;
«қарап шығу» : {@O-} & {P+} & {@O-} & A+;
«жүйесі» : {A-} & S-;
```

Объединяя эти файлы в один, мы получим общий грамматический файл.

«интернет» «ғаламтор» : {@A-} & S+;
«мыңдаған» : AA+;
«корпоративті» «үкіметтік» «ғылыми» «үй»: {@AA-} & {C-} & {C+} & A+;
«және» : C- & C+;
«желілерінен» : {@A-} & O+;
«құралған» : {@O-} & S-;
«компьютерлік» : {A+} & {AA+};
«серверлердің» : {AA-} & A+;
«бүкіләлемдік» : A+;
«желісі» «жүйесі» : {A-} & S-;
«ерексілтемелерді» : O+;
«пайдалана отырып» : P-;
«интернетті» : O+;
«қарап шығу» : {@O-} & {P+} & {@O-} & A+;

Разработка алгоритма и программы формирования грамматического файла казахского языка

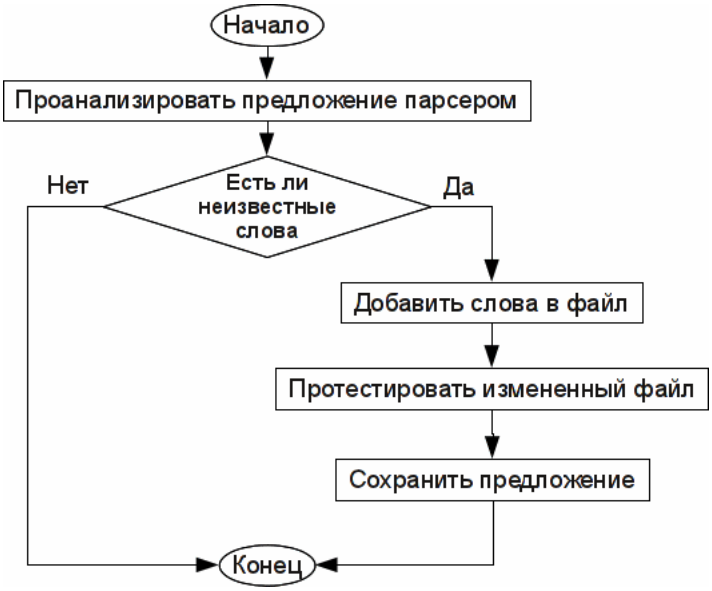

Рисунок 2. Алгоритм ручного заполнения грамматического файла

Как уже было описано в предыдущем разделе, грамматическая информация для парсера грамматики связей хранится в грамматическом файле. В «традиционной» реализации это текстовый файл с расширением «.dict». Структура грамматического файла имеет следующий вид:

"слово[.тэг части речи]": формула коннекторов [% комментарии]

Указание тега части речи и комментариев не является обязательным. Однако рекомендуется эти записи использовать для удобства дальнейшего сопровождения файла. Слова, обладающие одинаковыми формулами коннекторов, можно группировать следующим образом:

"слово1.тэг части речи"["слово2.тэг части речислово3.тэг части речи"...] : формула коннекторов [% комментарии]

На начальном этапе грамматический файл заполняется человеком. Рекомендуется придерживаться следующего алгоритма (рисунок 2):

- проанализировать предложение парсером;
- если есть слова, отсутствующие в словаре, добавить слова в файл;
- протестировать измененный файл;
- сохранить предложение для будущего тестирования.

Очевидно, что подобная процедура способна свести на нет удобство использования грамматики связей. Кроме того, при использовании в системе машинного перевода содержание грамматического файла должно быть согласовано со словарем, используемым для перевода.

Результат разбора предложения имеет вид (рисунок 2).

Для многих естественных языков (и для казахского языка в том числе) характерно следующее утверждение. Слова, имеющие одинаковые грамматические характеристики, играют одинаковую роль в предложении. К примеру имена существительные (зат есім) в именительном падеже (атау септік) в абсолютном большинстве предложений будут являться подлежащими. По этой причине, всем таким словам будет соответствовать одна и та же формула коннекторов грамматики связей. Соответственно, при добавлении подобного слова нет необходимости составлять новое правило. Достаточно внести это слово в соответствующую группу. Грамматические характеристики слов можно определять используя лексический анализатор. Это позволяет автоматизировать процесс формирования грамматического файла. Предлагаемый алгоритм имеет следующий вид (рисунок 3):

- проанализировать предложение парсером;
- если есть слова, отсутствующие в словаре, произвести их лексический анализ;
- сравнить грамматические характеристики с имеющимися в грамматическом файле;
- если имеются совпадающие характеристики, добавить новое слово в соответствующую группу;
- если совпадающих характеристик нет, сохранить новое слово для ручной обработки;
- протестировать измененный файл;
- сохранить предложение для будущего тестирования.

Использование текстового файла для хранения грамматической информации удобно с точки зрения создания. Но при использовании в системе машинного перевода имеет смысл реализовать хранение грамматической информации для грамматики связей в базе данных. Однако и в этом случае использование предложенного алгоритма позволяет автоматизировать заполнение подобной баз данных.

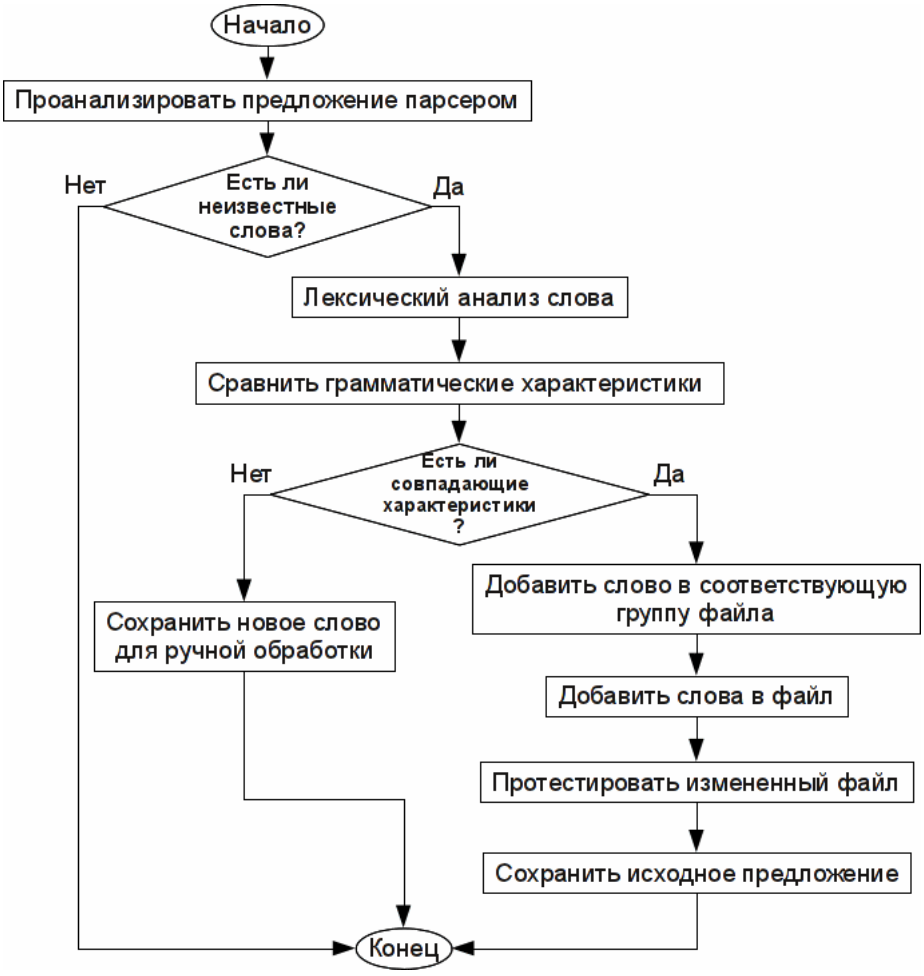

Рисунок 3. Алгоритм автоматизированного заполнения грамматического файла

В работе кратко рассмотрены существующие модели описания синтаксиса естественных языков и подробно описана грамматика связей. Дано ее общее описание, описан пример применения грамматики связей казахского языка. Подробно описан процесс реализации парсера грамматики связей, и предложен способ автоматизации процесса формирования грамматического файла.

Конечно, грамматика связей — не единственный способ осуществления синтаксического анализа естественных языков, в общем, и казахского языка, в частности. Однако описанные в данной работе преимущества этой грамматики оправдывают то внимание, которое ей уделено. Остальные представленные теории синтаксического анализа не могут быть отброшены. Они также заслуживают пристального изучения. Однако это задача выходит за рамки данной статьи.

Список литературы

[1] *Sleator D.D.K., Temperley D.* Parsing English with a Link Grammar // Third International Workshop on Parsing Technologies 64. – 1995. – 91 p.

[2] *Dryer Matthew S.* Order of Subject, Object, and Verb // In The World Atlas of Language Structures / edited by Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. – Oxford University Press, 2005. – P. 51 – 55.

[3] *Dehdari Jon.* A Primer for Localizing Link Grammar. http://www.ling.ohio-state.edu /jonsafari/link-grammar/primer.html 14.06.12

Zhumanov Zh.M. Development of link grammar for Kazakh language's syntactic analysis.
The paper briefly discusses existing models of natural languages's syntax and describes link grammar in detail. Its general description is given, an example of link grammar application for Kazakh language is described. Implementation of a link grammar parser is described in detail, and a method for automating the process of grammatical file forming is proposed.

Жуманов Ж.М. Қазақ тілінің синтаксистық талдау үшін байланыстар грамматикасын әзірлеу.
Мақалада кәдімгі тілдердің синтаксисін сипаттайтын қазіргі үлгілері қысқаша қаралған, және байланыстар грамматикасы толық сипатталған. Оның ортақ сипаттамасы берілген, қазақ тілі үшін байланыстар грамматиканы қолдану мысалы сипатталған. Байланыстар грамматикасының парсерін іске асыру процессі толық сипатталған, және грамматикалық файлдың құру процессін автоматтандыру әдісі ұсынылған.