работы интевьюеров (как низового звена, собирающего исходные данные в ходе социологических опросов) на текущем этапе нуждается в улучшении [4].

В третьих, невозможно недооценить важность и необходимость поддержания личных контактов со всеми исследовательскими коллективами, ведущими работы в области, даже в малейшей степени смежной с областью проводимого исследования.

В общем же размышления о данной проблеме приводят нас к дискуссии о необходимости переустройства отечественной социологической науки в новом тысячелетии и о степени соответствия используемых в ней информационных сетей и систем потребностям времени.

Заключение. Данное сообщение обращено прежде всего к студентам, планирующим развиваться в области математического и информационного сопровождения социальных исследований и математического моделирования масштабных социальных процессов. Проблемы, затронутые статье существуют уже давно, и неоднократно обсуждались, но до сих пор не решены, поэтому как минимум их обсуждение остаётся по-прежнему актуальным.

ЛИТЕРАТУРА

1. Толстова Ю. Н. Социология и компьютерные технологии / Социологические исследования, № 8, Август 2015, С. 3-13

2. Информационная система "Модернизация" ЦИСИ ИФРАН [Электронный ресурс] URL: http://mod.vscc.ac.ru/ требуется авторизация

3. Лапин Н.И., Беляева Л.А. Программа и типовой инструментарий «Социокультурный портрет региона России» (Модификация – 2010). М.: ИФРАН, 2010

4. Романчуков С.В., Берестнева Е.В., Маклакова Т.Г., Шухарев С.О., Информационная технология оценки качества работы интервьюеров. Труды Конгресса по интеллектуальным системам и информационным технологиям (IS-IT’ 16) - Дивноморское, 2-9 сентября 2016. – Таганрог: ЮФУ, 2016. – Т. 1 - С. 275-278

ОШИБКИ И НЕДОСТАТКИ СИНТАКСИЧЕСКОГО АНАЛИЗА СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА ТЕКСТА SEMSIN

В. В. Чемерилов, А.С. Фадеев.
(г. Томск, Томский политехнический университет)
e-mail: vchemerilov@gmail.com

THE SYNTAX ERRORS AND DISADVANTAGES OF THE SEMANTICS AND SYNTAX ANALISIS OF THE SEMSIN PARSER.

V.V. Chemerilov, A.S. Fadeev
(Tomsk, Tomsk Polytechnik University)

Annotation. This article describes algorithm of the Semsin parser. During the research of this tool for its adaptation to the text-to-speech system, errors in syntactic analysis were found**.**

Keywords: text-to-speech systems, text parser, semantic text analysis, syntax text analysis.

Введение. В процессе компьютерного анализа текстовых данных, текст проходит через несколько этапов обработки. На одном из таких этапов происходит синтаксический анализ, задача которого состоит в определении синтаксических связей между словами предложения. Результаты синтаксического анализа могут быть использованы для различных целей, например, для автоматического разрешения омонимии [1]. В данной работе рассмотрены ошибки, совершаемые семантико-синтаксическим анализатором Semsin при синтаксическом анализе предложений.

Анализатор Semsin. Semsin - это семантико-синтаксический анализатор русскоязычного текста, который выполняет следующие задачи [2]:

- Снятие частеречной и морфологической омонимии;
- Построение синтаксического дерева зависимостей;
- Частичное снятие лексической неоднозначности.

Принцип работы анализатора Semsin состоит в следующем: на первом этапе работы системы выделяются токены - минимальные, линейные и неделимые компоненты текста. На втором этапе происходит обработка отдельных слов и разрешение морфологической омонимии на основе анализа ближайшего окружения. На третьем этапе выделяются группы с фамилиями, названиями, числами. На четвертом этапе происходит подключение прилагательных, причастий и снятие неоднозначностей прилагательное–существительное. Далее идет поиск предлогов на основе анализа предложных групп. На последнем этапе работы анализатора происходит синтаксический анализ каждого предложения в тексте.

Синтаксический анализ в компьютерной лингвистике. Синтаксическим анализ – это процесс сопоставления линейной последовательности лексем естественного языка с его формальной грамматикой. Целью синтаксического анализа является построение синтаксического дерева зависимостей между словами в предложении. При удачном синтаксическом анализе предложение отображается в виде полносвязного дерева с единственной корневой вершиной. Так как одна словоформа может соответствовать нескольким грамматическим формам, при анализе необходимо произвести свертку предложения для всех случаев. Из них надо выбрать наиболее достоверные - те случаи, которые обладают минимальным числом висячих вершин.

Работа системы синтаксического анализа предложения начинается с применения правил разбора. С этой целью используется специальный алгоритм, который на каждом шаге проверяет возможность использования правила к отдельному фрагменту фразы (обычно двум-трем словам). Если правило удается применить к фрагменту, то он сворачивается – во фрагменте выделяется главное слово, остальные отбрасываются. Если дальнейшие применение правил над фрагментом невозможно, происходит откат (на любом из шагов). Последний фрагмент восстанавливается и предпринимается попытка использования другого правила. Окончательным результатом разбора следует считать такую последовательность применения правил, при которой происходит максимальное сворачивание предложения.

В процессе выполнения операции применения правил разбора стоит обратить внимание на обработку именных групп - устойчивых словосочетаний, состоящих из существительных и связанных с ними прилагательных [3]. Такие группы, как правило, описывают содержание текста и служат для автоматической рубрикации, тематического индексирования, уточнения запроса при поиске и т.д. Синтаксические отношения именных групп могут быть описаны большим количеством правил контекстно-независимой грамматики, которые учитывают только согласование грамматических форм. Например, если перед существительным стоит прилагательное, то они должны быть согласованы между собой по числу, падежу и роду (если единственное число).

При полном синтаксическом разборе фразы можно установить синтаксические роли именных групп в предложении, что позволит ранжировать их по степени важности (для автора). Это соответствует пониманию ключевых идей текста. Наиболее значимыми являются слова из групп подлежащего, затем следуют слова из групп сказуемого, прямого и косвенного дополнений, обстоятельства.

Иногда возникает ситуация, при которой появляются несколько равноправных вариантов разбора – явление синтаксической омонимии. Разрешение омонимии возможно только при привлечении семантики, а иногда и прагматики. Смысловую связь между понятиями предложения описывается с помощью глагола-предиката, аргументами, выступающими в

роли данных понятий. На основе синтактико-семантических связей можно сформировать логическую схему ситуации, описываемой во фразе.

Однако для выполнения данной операции требуется словарь управления моделей глаголов [3]. В таком словаре должно содержаться около 20 тыс. русских глаголов (почти все глаголы русского языка), а также должно быть указано с какими предлогами и падежами производится это управление. Для каждой модели обозначаются семантические роли аргументов глагола – это позволяет разделить связи по смыслу.

Ошибки синтаксического анализа Semsin. В результате исследования выходных данных дерева зависимостей синтаксического анализа семантико-синтаксического анализатора Semsin, удалось выявить следующие ошибки и недостатки.

1. Semsin неправильно определяет синтаксические связи и их типы между словами в предложении (рис 1).

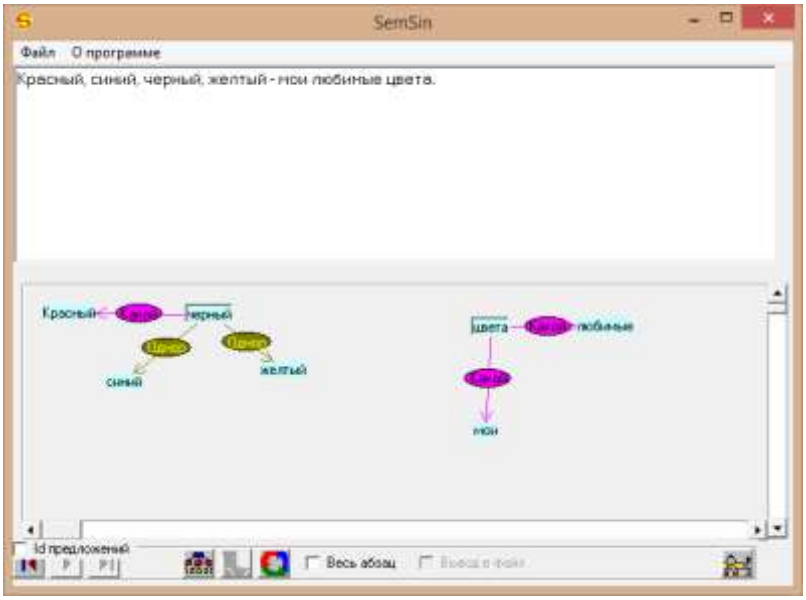

Рис. 1. Ошибка синтаксического анализа Semsin при разборе предложения «Красный, синий, черный, желтый - мои любимые цвета.»

В данном примере происходит ошибка в предложении, содержащее в себе однородные члены (прилагательное “черный” должно иметь синтаксическую связь “Однор” с другими частями предложения). Синтаксическое дерево зависимостей должно быть целым (не делиться на сегменты) – пропущены синтаксические связи.

2. Скудный список вопросов для определения типа синтаксических связей между словами. Особенно это заметно при анализе обстоятельства (рис. 2).

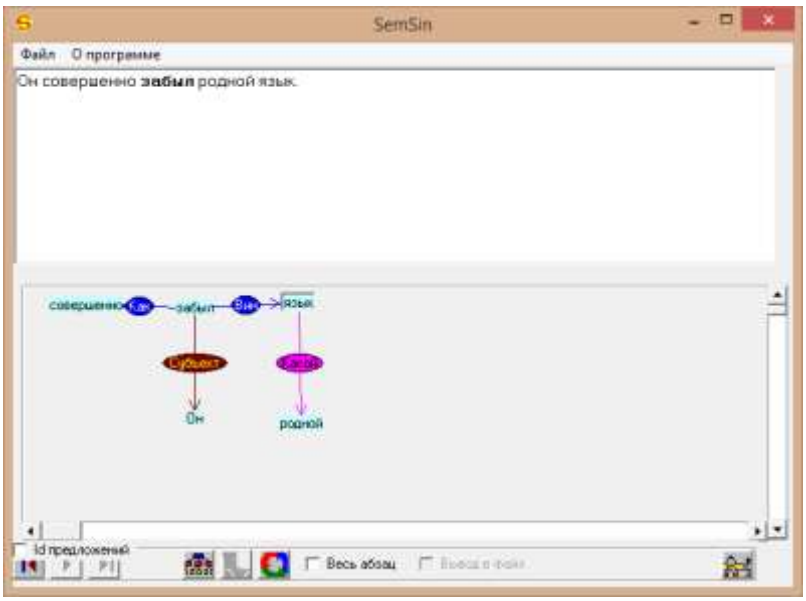

Рис. 1. Недостаток данных синтаксического анализа Semsin при разборе предложения «Он совершенно забыл родной язык»

Синтаксическая связь между словами «забыл» и «совершенно» определена правильно, однако с помощью этой связи сложно определить вид обстоятельства. Если бы синтаксическая связь между ними была «в какой степени? насколько?», тогда можно было установить, что слово «совершенно» является обстоятельством образом действия, меры и степени (наличие в предложении обстоятельства меры и степени является признаком акциональной доминанты, при которой логическим ударением выделяются все слова, обозначающие действия).

Заключение. Описанные выше ошибки происходят далеко не во всех случаях, но все же они имеют место быть. Они тесно связаны с ошибками морфологического и семантического анализа, поэтому при их исправлении необходимо обратить внимание на данные морфологии и семантики.

ЛИТЕРАТУРА

1. Лобанов Б. М., Житко В. А. О решении задач снятия омонимии при распознавании и синтезе речи [Электронный ресурс]. - URL: https://libeldoc.bsuir.by/handle/123456789/4372.

2. Каневский Е.А., Боярский К.К Предсинтаксический модуль в анализаторе SemSin [Электронный ресурс]. - URL: http://ojs.ifmo.ru/index.php/IMS/article/viewFile/46/47.

3. Ермаков А.Е Компьютерная лингвистика и анализ текста [Электронный ресурс]. - URL: https://www.osp.ru/pcworld/2002/09/16396.