Автоматическая классификация эллиптических конструкций в русской спонтанной речи

Automatic Classification of Elliptical Constructions in Russian Spontaneous Speech

Алина Путинцева
Alina Putintseva
aaputintseva@niuitmo.ru

Любовь Ковригина
Liubov Kovriguina
lyukovriguina@corp.ifmo.ru

Иван Шилин
Ivan Shilin
shilinivan@corp.ifmo.ru

Университет ИТМО, Санкт-Петербург, Российская Федерация

ITMO University, Saint-Petersburg, Russian Federation

Abstract

The paper describes experimental results on ellipsis detection in Russian spontaneous speech. The algorithm, which detects certain types of elliptical constructions in the input data, uses morphological features and features based on linear adjacency of utterance fragments, showing competitive results alongside with the algorithms exploiting parsing results.

Keywords: *ellipsis detection in spoken Russian, ellipsis type classification, linear adjacency features for ellipsis detection*

Аннотация

В статье описывается эксперимент по автоматическому обнаружению и восстановлению эллиптических конструкций в корпусе спонтанной русской речи. Алгоритм, классифицирующий высказывания по признаку наличия или отсутствия эллипсиса определенного типа, использует признаки, вычисляемые на основании морфологической разметки, что является существенным преимуществом в случаях, когда в рассматриваемых корпусах отсутствует синтаксическая разметка. Качество работы алгоритма сопоставимо с алгоритмом восстановления эллипсиса, использущим информацию о частоте синтаксических связей.

Ключевые слова: *автоматическое обнаружение эллиптических конструкций в русской спонтанной речи, автоматическая классификация эллиптических конструкций*

1 Введение. Эллипсис как лингвистическое явление

Лингвистическая теория трактует эллипсис как феномен отсутствия: см., напр., определение эллипсиса как "невыраженности тех фрагментов предложения, значение которых может быть восстановлено из контекста"[Testelets, 2011]. С.В.Чебанов при описании представлений о форме в работе [Chebanov, 1984] относит стерезис к средствам описания и постижения формы ("Стерезис (лишенность) - свое иное формы"(Аристотель), схожие взгляды на явление эллипсиса встречаются и у других исследователей: "Эллипсис есть значение без формы[1]"[Merchant, 2016].. Можно предположить, что нелинейность речи[2] и стерезис лежат в основе механизмов реализации эллипсиса. Эллипсис, по-видимому, ни в одном языке не подчиняется единому правилу и реализуется в виде нескольких конструкций, которые подчиняются разным правилам и по-разному вписываются (или не вписываются) в принятую исследователем теоретическую модель [Testelets, 2011]. Возникновение эллипсиса обусловлено стремлением к экономии языковых средств, времени и усилий и повышению эффективности коммуникации: во многих случаях мысль может быть выражена в сокращенной форме и при этом оставаться понятной для слушателя [Bogdanova-Beglaryan, 2014]. Эллипсис позволяет избавиться от избыточности в речи, и, вероятно, разнообразие эллиптических конструкций (гэппинг, псевдогэппинг, эллипсис вершины именной группы, слусинг и др.) обусловлено прагматическим компонентом высказывания.

Дополнительная сложность заключается в том, что правила эллипсиса зависят от языка [Testelets, 2011]. Кроме того, в русском языке существуют схожие с эллипсисом синтаксические явления, и важной задачей является разграничение эллипсиса и таких понятий, как неполные предложения, парцелляция, синтаксический нуль и зевгма [Kasaeva, 2014], при этом И.В.Камагина уточняет, что «между эллипсисом и сходными явлениями неполноты, имплицитности, зевгмы может не быть четкой границы» [Kamagina, 2015].

2 Типология эллиптических конструкций

Общепринятой типологии эллиптических конструкций на данный момент не существует, а в позициях учёных наблюдаются расхождения. В данном разделе собраны типы эллиптических конструкций, выделяемые большинством исследователей.

В работе [Paducheva, 2004] предложено разделение эллипсиса на анафорический и ситуативный. Первый тип предполагает отсылку к другим фрагментам текста и связь опущенных элементов высказывания с контекстом слов текущего предложения (1)[3] или соседних предложений (2):
(1) Второй способ заключается в том что ORACLE поднимает одну таблицу на од на один

[1]Далее приведена расширенная цитата в оригинале:"Ellipsis continues to be of central interest to theorists of language exactly because it represents a situation where the usual form/meaning mappings, the algorithms, structures, rules, and constraints that in non-elliptical sentences allow us to map sounds and gestures onto their corresponding meanings, break down. In fact, in ellipsis, the usual mappings seem to be entirely absent. In ellipsis, there is meaning without form"[Merchant, 2016]

[2]С.В.Чебанов, личное сообщение, 2005 г.

[3]Примеры эллиптических конструкций приведены из корпуса PARS [Kovriguina et al., 2018], за исключением примеров из литературных произведений, приведенных в цитируемых работах. Части высказывания, подвергшиеся эллипсису, восстановлены в скобках курсивом.

инстанс, вторую таблицу (*поднимает*) - на второй инстанс и пытается как-то по сети это всё рассинхронизовать.
(2) Ты читал замечательную книжку Стентона Глэнца про медицинскую статистику? Она написана очень доходчиво, там (*объясняются ошибки применения статистических методов*) на марсианах, на юпитериан на людях с Юпитера, людях с Венеры.

При ситуативном эллипсисе пропущенный элемент не упоминается в речи эксплицитно и не может быть восстановлен однозначно. Интерпретация таких высказываний, поверхностно-синтаксическая структура которых позволяет квалифицировать их только как фрагменты, возможна лишь при знании ситуации (3-6):
(3) *Покажите*!
(4) Мне чёрный (*хлеб, чай*).
(5) А я (*зайду*) вон туда.
(6) "Три билборда"лежат в очереди на просмотр, но, говорят, мрачный, депрессивный (*фильм*) и так далее.

Задача восстановления ситуативного эллипсиса является наиболее сложной и сводится к анализу прагматического компонента содержания высказывания, дискурс-анализу и анализу экстралингвистического контекста.

Анафорический эллипсис, в свою очередь, разделяют на несколько видов [Testelets, 2011], [Testelets et al., 2005], которые перечислены ниже.

2.1 Эллипсис составляющих именной группы

Эллипсис составляющих именной группы "с сохранением представителя"(англ. *noun ellipsis*, *N-ellipsis*, *noun phrase ellipsis*, *NPE*) может возникнуть в предложении (предложениях) с повторяющимися существительными или именными группами и реализуется как опущение одной из составляющих (часто - существительного-вершины именной группы, см.примеры (7-9). При этом в эллиптических предложениях сохраняется «представитель» пропущенной составляющей, обладающий теми же грамматическими категориями, что и она сама. В качестве «представителя» могут выступать:

- прилагательное, местоимение или причастие, согласующееся в числе, роде и падеже с опущенной составляющей:
  (7) "Не будет никакого цветового кодирования, а если и будет, то минимальное (*кодирование*)" ;

- числительное, согласующееся в падеже и/или роде (8-9) с опущенной составляющей:
  (8) "теоретически, две (*недели*), может, три недели";
  (9) "ну, я заплатил за самолет около двадцати тысяч, а зимой он не стоит и восьми (*тысяч*) ";

- генитивная конструкция[4] (или ее фрагмент), входящая в более крупную именную группу, в данном случае восстановлению подлежит именная группа, содержащая генитивную конструкцию:
  (10) "смотри, в этой папке (находятся) документы прошлого года, а в этой - (*документы*) текущего (*года*) ";

[4]Об эллипсисе в генитивных конструкциях смотри подробнее труды Е.В.Падучевой[Paducheva, 2013].

- существительное, при котором элиминируется синтаксическая группа или целая определительная клауза, при этом существительное-представитель имеет то же число и падеж, что и существительное-вершина в полной конструкции:
(11) "Слева вот во втором ряду на фотографии - племянница брата, а в первом по центру - дочь (*брата*)".

2.2 Слусинг

Слусинг (англ. sluicing) — эллиптическая конструкция, в которой зависимая клауза со специальным косвенным вопросом заменяется вопросительно-относительным местоимением (или местоимением с предлогом) [Kibrik et al., 2017]. "Предложение «стирается» целиком, остается только вопросительное местоимение"[Testelets et al., 2005], см. примеры (12, 13) ниже:
(12) Кто-то тебе звонил, но я не знаю, кто (*тебе звонил*).
(13) Он вроде как собирался уехать в командировку, только вот когда (*он собирался уехать*)...

2.3 Эллипсис глагольной группы, стриппинг, гэппинг и псевдогэппинг

Четыре перечисленных типа эллиптических конструкций (VP-эллипсис, стриппинг, гэппинг и псевдогэппинг) связаны с исключением глагола, как изолированного, так и в составе глагольных групп и целых клауз (о различительных тестах между этими конструкциями, см., напр., работы Т.Графа [5]). Стриппинг наблюдается в тех случаях, когда из высказывания удаляется неначальный конъюнкт сочинительной конструкции (в общем случае, все неначальные конъюнкты) [Merchant, 2016] [Kolokonte, 2008], при этом сочинительная конструкция может состоять как из клауз (14), так и из синтаксических групп меньшего объема:
(14) Я готовился всю ночь, и они тоже (готовились всю ночь).

В английском языке некоторые исследователи не различают стриппинг и эллипсис глагольной группы (VP-ellipsis) [Kolokonte, 2008].
При гэппинге (англ. gapping) из высказывания удаляется "вершина одной или нескольких неначальных сочиненных составляющих"[Bogdanov, 2012], см. пример (15). Псевдогэппинг реализуется схожим образом, но место опущенного фрагмента занимает вспомогательный глагол или связка. В английском языке к гэппингу предъявляется более сильное требование: из высказывания удаляется глагол - вершина неначальной клаузы [Kolokonte, 2008]. Я.Г.Тестелец указывает, что в русском языке это требование смягчено, и к гэппингу относятся все случаи "внутренних пробелов"независимо от того, опущен предикат или один из аргументов [Testelets, 2011], см. пример (16):
(15) Мы, например, записали десять мегабайт час назад и (*записали*) один мегабайт только что.
(16) Он открывает окно, я закрываю (*окно*).

[5] `http://thomasgraf.net/doc/other/ellipsis.pdf`

2.4 Фрагментирование

Фрагментирование (эллипсис при ответе на вопрос, англ. answer ellipsis) — тип эллипсиса, возникающий в диалогах при ответах на вопросы.
(17) — Кому ты звонил?
— (*Я звонил*) Александру Евгеньевичу.

2.5 Сравнительный эллипсис

Сравнительный эллипсис (англ. comparative deletion) подразумевает пропуск слов в конструкциях со сравнительной степенью (18) и в конструкциях со сравнительными союзами и частицами (19):
(18) В первом зале людей было больше, чем во втором (*зале было людей*).
(19) Сидя он спит так же крепко, как и лёжа (*он спит*).

3 Представление эллиптических конструкций в синтаксически размеченных корпусах

Методы автоматического восстановления эллипсиса, как правильно, используют корпусы текстов с морфологической, синтаксической и, при наличии, семантической разметкой. Ниже приведены подходы к представлению эллиптических конструкций в корпусе текстов.

3.1 Синтаксически размеченный подкорпус Национального корпуса русского языка

Национальный корпус русского языка (НКРЯ) включает в себя ряд подкорпусов, различающихся составом, видами разметки и способами её получения. В частности, СинТагРус представляет собой синтаксически размеченную часть НКРЯ[6]. Синтаксическая структура каждого предложения корпуса имеет вид дерева зависимостей, узлами которого являются слова, а рёбрами — синтаксические отношения между ними. В НКРЯ применяются сходные процедуры для представления эллиптических конструкций, в которых опущенный фрагмент однозначно восстановим из контекста, и эллиптических конструкций, в которых отсутствующий фрагмент может быть восстановлен лишь гипотетически.

В эллиптичных клаузах опущенные слова восстанавливаются, и им приписывается дополнительный признак фантом: "Например, в случае предложения *Я купил чемодан, а он сумку* между *он* и *сумку* вставляется узел

```
КУПИТЬ [V, сов, изъяв, прош, ед, муж, фантом]
```

с пустым текстовым элементом. От этих фантомных слов проводятся все необходимые связи. Леммы в таких словах совпадают с теми, которые уже встретились в предложении, а отдельные морфологические характеристики могут меняться (так, в предложении

[6] `http://ruscorpora.ru/instruction-main.pdf`

Я купил чемодан, а она сумку характеристика *муж* в новом, «фантомном» глагольном узле *купила* заменяется на *жен*). Подчеркнем, что фантомные слова вводятся только в синтаксическую структуру предложения. Текстовый вид предложения остается неизменным"[7].

В случаях, "когда в предложении «опущен» глагол некоторой размытой семантики, как в следующем тексте: Парочку морей бы еще в Сибирь. Африку можно бы ниже. Индия пусть. (Т. Толстая) добавляется узел, аналогичный «фантому», ему приписываются наиболее естественные характеристики, а в качестве леммы пишется НЕОПР-ГЛАГОЛ (неопределенный глагол) и затем в скобках глагол, который является «естественной гипотезой». Так, в последнем примере после пусть добавляется узел с леммой НЕОПР-ГЛАГОЛ (ОСТАВАТЬСЯ)"[8].

3.2 Хельсинкский аннотированный корпус русского языка

Подход к синтаксической разметке, применяемый в хельсинкском аннотированном корпусе русских текстов ХАНКО, подразумевает параметризацию клауз [Frolova, 2012]. Каждая клауза имеет такие характеристики, как:

- роль (самостоятельная, главная, подчинённая);
- структура (двусоставная, односоставная, фразеологизированная);
- наличие эллипса (эллиптическая клауза).

Таким образом, положение пропущенных элементов высказываний в текстах ХАНКО не указано явно, а наличие эллиптической конструкции характеризует клаузу в целом.

3.3 Представление эллиптический конструкций в стандарте Universal Dependencies

Основные принципы стандарта Universal Dependencies, касающиеся разметки предложений, содержащих эллиптические конструкции[9], можно сформулировать следующим образом: 1) Если исключённая вершина синтаксического дерева не имеет зависимых вершин, никакая дополнительная разметка не осуществляется. 2) Если исключённая вершина имеет зависимые вершины, одна из них принимает на себя роль опущенного элемента. 3) Если исключённая вершина является предикатом, а заменившая её вершина — одним из её актантов или адъюнктов, для связи с остальными элементами используется зависимость *orphan*. Стандарт регламентирует представление конструкций с эллипсисом в именных группах и эллипсис в клаузах. Согласно стандарту, при опущении существительного его роль в дереве передаётся в следующем порядке: $amod \rightarrow nummod \rightarrow det \rightarrow nmod \rightarrow case$. Другими словами, наиболее приоритетным среди потенциальных замещающих узлов является узел, связанный с опускаемым отношением *amod*, наименее приоритетным — *case*. Механизм замещения вершины при N-эллипсисе изображён на рисунке 1. В случае опущения предиката наиболее вероятный претендент на его роль — его субъект (nominal subject — *nsubj*). Далее в порядке убывания приоритетов идут оставшиеся аргументы —

[7] http://www.ruscorpora.ru/instruction-syntax.html

[8] \T2A\CYRT\T2A\cyra\T2A\cyrm\T2A\cyrzh\T2A\cyre.

[9] http://universaldependencies.org/v2/ellipsis.html

obj (object) и *iobj* (indirect object). Дальнейшая очерёдность выглядит следующим образом: *obl* → *advmod* → *csubj* → *xcomp* → *ccomp* → *advcl*.

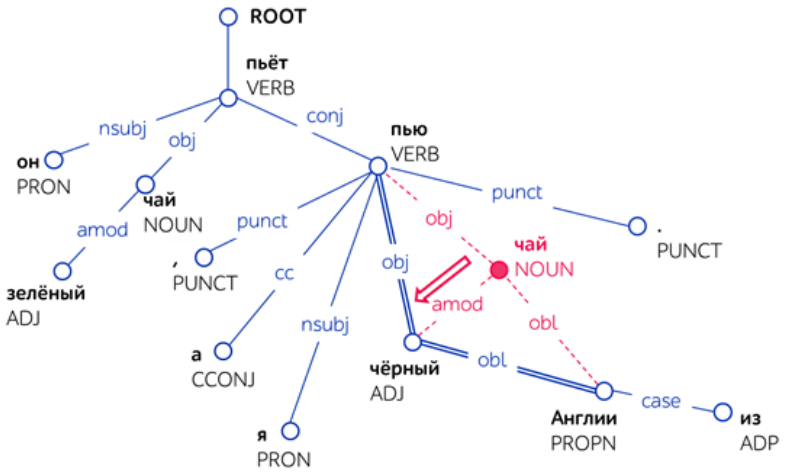

Рис. 1: Иллюстрация принципа переноса связей вершины именной группы, подвергшейся эллипсису, на зависящее от нее прилагательное в предложении *Он пьет зеленый чай, а я пью черный чай из Англии.*

Кроме того, стандарт UD включает в себя так называемое расширенное представление (enhanced dependencies), допускающее ввод дополнительных (вспомогательных) узлов на местах пропущенных (процедура аналогична подходу, принятому в НКРЯ).

4 Обзор подходов к автоматическому обнаружению и восстановлению эллипсиса

Задача восстановления эллипсиса в большинстве случаев не представляет особой трудности для человека, но достаточно сложна при машинной обработке. Однако, в обоих случаях используется один и тот же приём — восстановление из контекста. При этом восстановление разных видов эллипсиса подразумевает привлечение различных видов контекста, а сложность поставленной задачи варьируется в зависимости от вида контекста. Выделяют следующие виды контекста [Mal'kovskij et al., 2014]:

- контекст слов текущего предложения;
- контекст соседних предложений;
- контекст целого текста;
- прагматический контекст.

При анализе спонтанной речи традиционные представления о предложении оказываются неприменимы или применимы с трудом, поэтому при разработке алгоритма были использованы контекстные признаки следующих единиц анализа спонтанной речи:

- контекст слов элементарной дискурсивной единицы;

- контекст соседних элементарных дискурсивных единиц;

- контекст псевдопредложения;

- контекст целого текста (монолога, полилога), моделируемый с помощью векторных представлений.

Эллипсис относится к темам, активно обсуждаемым научным сообществом. Среди работ авторства российских учёных стоит выделить труды М.Г.Мальковского и его коллег с кафедры алгоритмических языков МГУ им. М.В.Ломоносова, связанные с разработкой системы морфологического и синтаксического анализа Treeton [Mal'kovskij et al., 2014], [Mal'kovskij et al., 2012]. В работе[Mal'kovskij et al., 2012] описан подход, опирающийся на понятие синтаксической валентности: часть речи словоформы и, при наличии, набор значений её морфологических характеристик определяют, с какими другими словоформами она может быть связана в предложении. Облигаторные валентности, не заполненные в высказывании, авторы предлагают считать показателем эллипсиса. Задача восстановления пропущенных элементов и, таким образом, заполнения свободных валентностей сводится к составлению набора шаблонов и правил, а результатом анализа является совокупность возможных вариантов синтаксического дерева рассматриваемого предложения, каждому из которых поставлено в соответствие некоторое значение вероятности. В работе [Minyajlov, 2016] рассмотрено обнаружение и восстановление эллиптических конструкций с помощью различных алгоритмов классификации. При этом ведётся разделение анализируемых предложений на два класса — «нулевые» (эллиптические) и «обычные». В качестве материала для работы используется синтаксически размеченная часть НКРЯ, содержащая предложения обоих классов. Классификация ведётся на основе как морфологических, так и синтаксических признаков. В качестве значений метрик для каждого предложения рассчитывается:

– частота слов каждой части речи;

– частота связей каждого из возможных типов связей;

– частота каждой из возможных пар частей речи среди биграмм;

– частота каждой из возможных пар частей речи среди узлов синтаксического дерева, непосредственно связанных друг с другом;

– для каждой тройки (p_1, p_2, p_3) частота троек (v_1, v_2, v_3), где p_i — часть речи, v_i — узел синтаксического дерева, имеющий частеречный тэг (POS-тэг) p_i, v_2 непосредственно зависит от v_1, а v_3 — от v_2 ($v_1 \rightarrow v_2 \rightarrow v_3$); 6) для каждой пары (r_1, r_2) частота троек (v_1, v_2, v_3), где r_i — тип синтаксической связи, v_i — узел синтаксического дерева, v_1 и v_2 связаны отношением r_1 , v_2 и v_3 — отношением r_2.

Классификация производилась без кросс-валидации, корпус был разделен на обучающую и тестовую выборки в соотношении 70% : 30%, использовались следующие алгоритмы:

– метод k ближайших соседей (k-nearest neighbors algorithm, kNN),

– метод опорных векторов (support vector machine, SVM),

– деревья принятия решений (decision tree),

– метод AdaBoost (adaptive boosting),

– бэггинг (bootstrap aggregating).

Лучшие значения точности классификации автору удалось получить при использовании метода опорных векторов: при разных наборах признаков, указанных выше, точность составила от 0.666 до 0.722 [Minyajlov, 2016]. В.С.Миняйлов указывает, что "существенно повысить эту точность не удалось, даже за счет добавления признаков или использования различных алгоритмов классификации"[Minyajlov, 2016].

Другой подход к решению проблемы исследования описан в работе [Giuliani et al., 2014]. Авторы справедливо указывают, что алгоритмы, основанные на использовании результатов парсинга, могут быть неэффективны при обработке спонтанной речи, если были использованы модели синтаксического анализа, разработанные на текстах, появившихся в результате подготовленного рече- / текстопорождения. Это связано в общем случае с нелинейностью речи и отличающимся набором правил порождения высказываний в спонтанной речи. Ключевая идея предлагаемого ими подхода основывается на том, что псевдопредложение спонтанной речи можно разделить на несколько грамматически правильных фрагментов, стыки которых являются потенциальными позициями для эллипсиса. Таким образом, поиск пропущенных слов сводится к выделению с начала и конца псевдопредложения таких подстрок максимальной длины, которые могут быть корректно интерпретированы парсером.

5 Алгоритм обнаружения эллиптических конструкций спонтанной русской речи

5.1 Описание алгоритма и данных

Разработанный алгоритм состоит из двух основных модулей: модуля извлечения признаков и модуля классификации. Вспомогательные модули включают модули загрузки и выгрузки данных, модуль обмена данными. На вход алгоритму поступает размеченное псевдопредложение из корпуса PARS [Kovriguina et al., 2018] в формате CoNLL-U[10]. Используя только контекстную и морфологическую информацию, метод извлечения признаков (об используемых признаках см. подраздел 5.2) анализирует входные данные и записывает соответствующие значения в вектор признаков. Метки класса в датасете были проставлены вручную. Датасет включает 60 примеров с гэппингом, 40 с N-эллипсисом и 200 - без эллипсиса.

Следует дать некоторые пояснения о псевдопредложении, принятом в качестве операциональной структуры в корпусе PARS и некоторых других корпусах [Shitaoka et al, 2004], [Kovriguina et al., 2018]. Фразовая структура отдельной элементарной дискурсивной единицы может быть достаточно примитивной, но при синтаксическом анализе спонтанной речи часто бываает необходимо установить связи синтаксических групп текущей ЭДЕ с синтаксическими группами соседних ЭДЕ. Уровень разметки, соответствующий псевдопредложению, был введен с целью адекватного представления связей на уровне микросинтаксиса (в смысле И.А.Мельчука) и соблюдения требований, предъявляемых к формату CoNLL-U. Элементарные дискурсивные единицы a и b объединяется в псевдопредложение, если в a и b существуют словоформы, связанные синтаксическим отношением из

[10] `http://universaldependencies.org/format.html`

списка Universal Dependencies[11], или перечня дополнительных синтаксических отношений из корпуса PARS[12].

5.2 Признаковое пространство

Признак 1 — Среднее расстояние между адъективом и его потенциальной вершиной

Атрибутивные слова (или адъективы: прилагательные, местоимения-прилагательные, числительные-прилагательные, причастия глаголов, анафорические местоимения типа *он*, *который*[13]) в русском языке, как правило, линейно близки к их вершинам. Тогда можно предположить, что отсутствие потенциальной вершины в пределах определённого окна с некоторой долей вероятности свидетельствует о её опущении и, таким образом, наличии в предложении (псевдопредложении) эллипсиса вершины именной группы.

При этом, согласно результатам анализа текстов корпуса СинТагРус, в предложениях без эллиптических конструкций адъективы, для которых в стандарте Universal Dependencies используются синтаксические отношения *amod* и *nummod:gov*, в подавляющем большинстве случаев зависят от существительных (см. таблицу 1).

Таблица 1: Наиболее распространённые POS-тэги атрибутов и их вершин

POS-тэг зависимого слова (адъектива)	Синтаксическое отношение	Частота встречаемости вершин	
		POS-тэг вершины	f, относительная частота
1	2	3	4
ADJ	amod	NOUN	93,6%
		PROPN	2,4%
DET	amod	NOUN	93,3%
		PRON	2,5%
NUM	nummod:gov, nummod	NOUN	88,5%
		NUM	3,7%

При расчете относительной частоты f в столбце 4 количество связей N рассчитывается как количество случаев, когда адъектив, имеющий POS-тег, указанный в столбце 1, связано синтаксическим отношением из столбца 2, с вершиной X, имеющей один из POS-тегов из стандарта Universal Dependencies.

Распределения расстояний между адъективом и его вершиной, как в корпусе SynTagRus,так и в корпусе спонтанной речи PARS позволяют утверждать, что адъективы чаще всего находятся на расстоянии одного или двух словоупотреблений от своей вершины (см. рисунки 2, 4).

На основании данных, приведённых выше, были сформулированы шаблоны именных групп с адъективами (см. таблицу 2).

[6pt]article

Значение признака рассчитывается как среднее расстояние от каждого слова, относящегося к одной из частей речи, перечисленных в столбце 1 таблицы 2, до его потенциальной вершины. Под потенциальной вершиной понимается ближайшее слово, часть речи

[11] `http://universaldependencies.org/u/dep/`

[12] `https://github.com/MANASLU8/PARS/blob/master/additional_relations_spoken_syntax`

[13] См. Русская корпусная грамматика `http://rusgram.ru/`

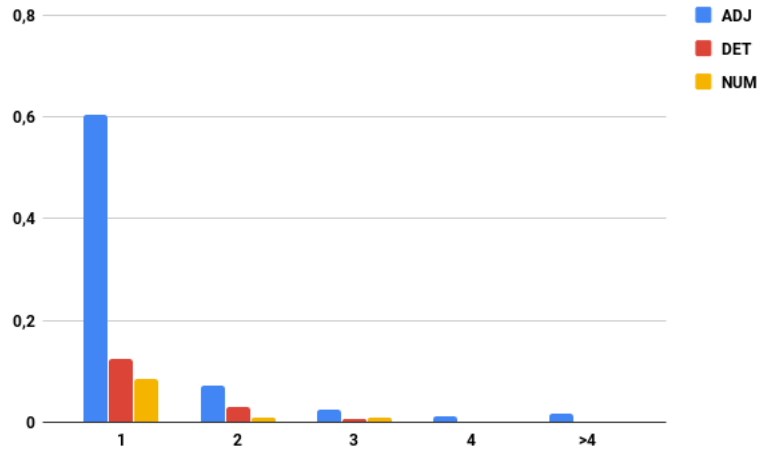

Рис. 2: Распределение расстояний между определениями и их вершинами в корпусе SynTagRus

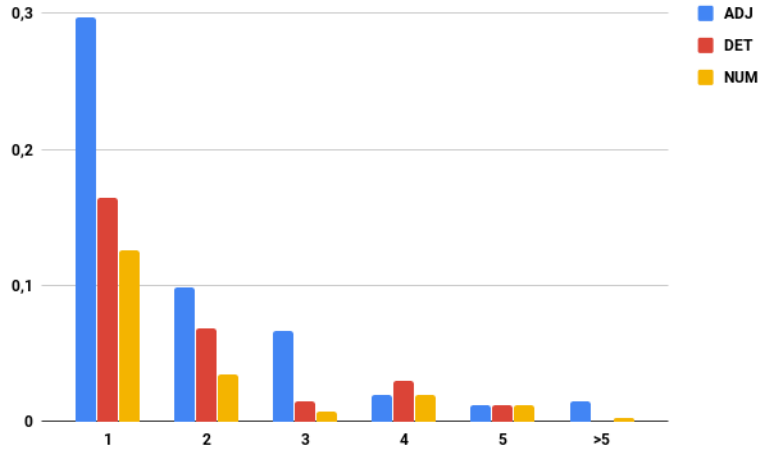

Рис. 3: Распределение расстояний между определениями и их вершинами в корпусе PARS

которого приведена в соответствующей строке столбца 2 таблицы 2, такое, что перечисленные в столбце 3 таблицы 2 словоизменительные признаки обоих слов совпадают. Чем больше будет полученное значение, тем выше вероятность N-эллипсиса. Очевидно, что поиск потенциальной вершины имеет смысл только в рамках ближайшего контекста. В ходе работы было сделано предположение, что его протяжённость в словоупотреблениях в большинстве случаев ограничивается числом Ингве-Миллера (7 ± 2), характеризующим объем кратковременной памяти человека [Miller, 1956].

Для экспериментальной проверки предположения было построено распределение расстояний между фантомными (восстановленными) узлами и их вершинами в корпусе СинТагРус (Рисунок 4).

Как видно из рисунка 4, большая часть фантомных узлов находится на расстоянии 3–7 слов от вершин, от которых они зависят.

Признак 2 — Среднее расстояние между повторяющимися словами При расчёте значений рассмотренного выше признака было замечено, что иногда отсутствие по-

Таблица 2: Шаблоны для вычисления значения признака

Метка части речи и морфологические признаки атрибута	**Метка части речи потенциальной вершины**	**Общие слово-изменительные категории атрибута и вершины**	**Примеры предложений с эллипсисом вершины именной группы**
1	2	3	4
Прилагательное, Причастие, порядковое числительное, указательное местоимение	Существительное, собственное существительное	Gender, Number, Case	Я предпочитаю яркие цвета, он чаще одевается в чёрный. Вчера мы шли по этой улице, а сегодня идём по той.
Количественное числительное	Существительное, собственное существительное, порядковое числительное	Gender, Number, Case	Чаще за занятие удаётся разобрать одну тему, но иногда успеваем изучить две или три.

тенциальной вершины в ближайшем окружении наблюдается при повторе — речевом сбое. При этом такое предложение не обязательно содержит эллипсис. В примере на рисунке 5 первое словоупотребление «две» относится к речевому сбою - повтору (возможно, говорящий хотел перестроить высказывание, но вербализованная часть позволяет говорить только о повторе), поэтому, несмотря на наличие формальных маркеров, N-эллипсис в данном примере отсутствует. В корпусе PARS повторы рамечаются с помощью дополнительного синтаксического отношения $repair : repetition$.

Для учёта таких случаев было решено ввести признак, отвечающий за повторы адъективов. Значение признака в корпусе без синтаксической разметки определяется как усреднённое расстояние от каждой словоформы, подходящей под критерии из столбца 1 таблицы 2, до ближайшей к ней словоформы с такой же леммой.

Признак 3 — Среднее расстояние между корневыми актантами и их потенциальными вершинами Для идентификации предложений с гэппингом было решено ввести признак, основанный на удалённости аргументов (core arguments) от предиката. Стандарт UD предусматривает для них три синтаксических отношения: субъект ($nsubj$), объект (obj) и косвенное дополнение ($iobj$)[15]. Значение признака рассчитывается как усреднённое расстояние от существительных и местоимений до ближайших глаголов.

Признак 4 — Расстояние от наречия «тоже» до ближайшего глагола Одна

[15] `http://universaldependencies.org/u/dep/`

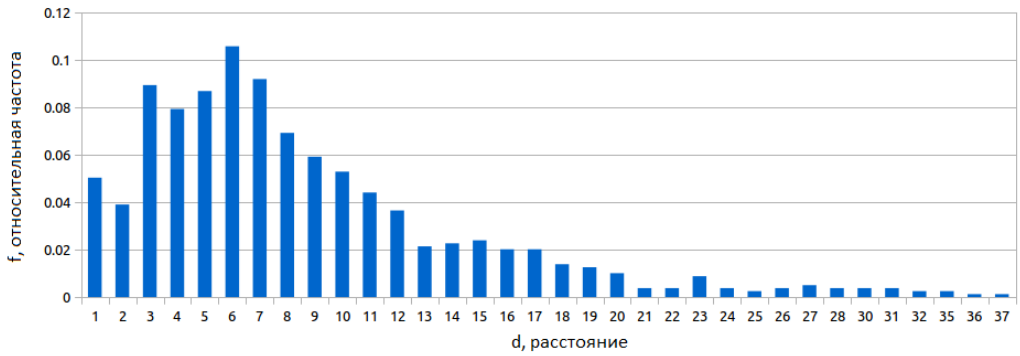

Рис. 4: Расстояние между фантомным узлом и его вершиной в корпусе SynTagRus

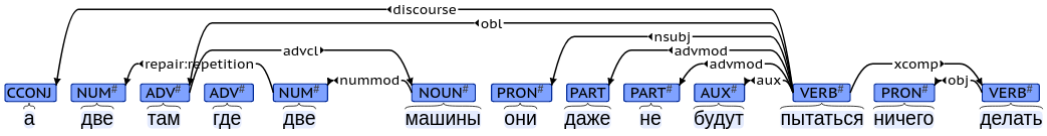

Рис. 5: Пример разметки речевого сбоя - повтора из корпуса спонтанной русской речи PARS

из наиболее распространённых маркеров стриппинга — пропуск повторяющегося глагола/глагольной группы, за которым следует наречие «тоже». Рисунок 6 демонстрирует соотношение POS-тэгов в рамках множества вершин слова «тоже» корпуса СинТагРус. Стоит отметить, что все перечисленные на нём части речи, кроме глагола, по-видимому, могут быть вершиной этого наречия только в эллиптических конструкциях.

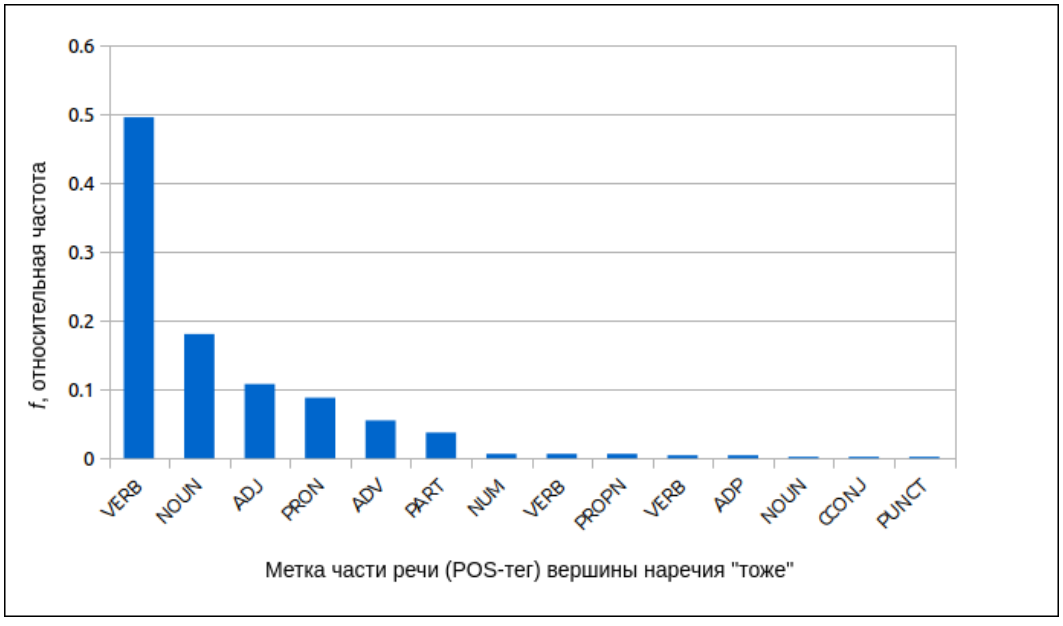

Рис. 6: Распределение вершин наречия "тоже"по меткам частей речи

Анализ расстояний между наречием «тоже» и его вершиной в неэллиптических предложениях (т. е. в предложениях, где она выражена глаголом) свидетельствует о том, что

в большинстве случаев «тоже» — соседнее с глаголом слово.

Группа признаков — Частота различных частей речи в псевдопредложении Метки частей речи используются во многих работах в качестве маркера эллипсиса [Minyajlov, 2016],[Mal'kovskij et al., 2014]. Аналогичную группу признаков было решено ввести и в разрабатываемое признаковое пространство. Частеречная структура псевдопредложения кодируется в виде вектора, в котором указываются частоты частей речи, встретившихся в данном псевдопредложении (если в нем не встретились слова определенной части речи, то в векторе на их месте записываются нули).

5.3 Классификация

Каждая строка в обучающей выборке соответствует псевдопредложению из корпуса PARS и представляет собой вектор признаков, описанных в разделе 5.2, и метку класса (N-эллипсис, VP-эллипсис, отсутствие эллипсиса).

Для обучения классификатора и оценки признакового пространства были использованы алгоритмы классификации, ранее применявшиеся для решения задач идентификации и восстановления эллипсиса в работах [Minyajlov, 2016], [Nielsen, 2004]. Обучение классификаторов проводилось с использованием процедуры скользящего контроля по 10 блокам[16]. Для классификации использовались следующие пакеты из языка R: rpart, class, e1071, adabag, kernlab [17].

- метод k-ближайших соседей (k-nearest neighbors algorithm, kNN),
- метод опорных векторов (support vector machine, SVM),
- деревья принятия решений (decision tree),
- метод AdaBoost (adaptive boosting),
- бэггинг (bootstrap aggregating).

Для оценки качества работы классификатора и применимости созданного признакового пространства и собранного набора данных использовались метрики точности (1), полноты (2) и F-меры (3):

$$Precision = \frac{TP}{TP + FP}, \quad (1)$$

$$Recall = \frac{TP}{TP + FN}, \quad (2)$$

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \quad (3)$$

5.4 Результаты и обсуждения

Результаты оценки качества автоматической классификации для каждого из используемых алгоритмов приведены в таблице 3. Лучшие результаты были получены с помощью методов, строящих композиции классификаторов (бустинг, бэггинг).

[16] `http://www.machinelearning.ru/wiki/index.php?title=CV`

[17] `https://cran.r-project.org/web/packages/available_packages_by_name.html`

Таблица 3: Результаты классификации типов эллиптических конструкций в русской спонтанной речи

Вид эллипсиса / Классификатор	*VP*-эллипсис			*N*-эллипсис			Без эллипсиса		
	P	R	F_1	P	R	F_1	P	R	F_1
Деревья решений	0,89	0,71	0,79	0,65	0,86	0,74	0,70	0,71	0,70
k-ближайших соседей, k=3	0,89	0,74	0,81	0,76	0,69	0,72	0,71	**0,92**	0,80
Метод опорных векторов (SVM)	0,89	0,68	0,77	0,50	0,33	0,40	0,51	0,81	0,63
AdaBoost	0,90	0,81	0,85	**0,81**	0,86	**0,83**	0,73	0,79	**0,76**
Бэггинг	**0,92**	**0,82**	**0,87**	0,75	**0,92**	**0,83**	**0,76**	0,71	0,73

Как видно из таблицы 3, результаты классификации превосходят результаты, полученные другими авторами ([Minyajlov, 2016], см. тж. раздел 4). Однако, прямое сравнение провести невозможно, так как предыдущие эксперименты проводились на материале корпуса СинТагРус, который значительно больше и разнообразнее по своей структуре. Учитывая это, можно говорить о том, что основным результатом данной работы является пробная оценка качества классификации эллиптических конструкций на материале спонтанной речи с помощью признаков, вычисляемых исключительно на морфологических и линейных характеристиках высказываний.

Acknowledgments

Статья выполнена по результатам проекта № 16-36-60055, поддержанного РФФИ в 2016-2018 гг.

Список литературы

[Rozental' et al., 1985] Rozental', D., Telenkova, M. (1985) Slovar'-spravochnik lingvisticheskih terminov: Posobie dlya uchitelya [Dictionary-reference of linguistic terms. Manual for teachers.] M.: Prosveshchenie, 1985. pp. 357. (In Russian) = Розенталь Д., Теленкова М. Словарь-справочник лингвистических терминов: Пособие для учителя. М.: Просвещение, 1985. С. 357.

[Bogdanova-Beglaryan, 2014] Bogdanova-Beglaryan, N. (2014) Eshche raz o zakone ekonomii v povsednevnoj spontannoj rechi // Kommunikativnye issledovaniya [Once again on the economy law in everyday spontaneous speech // Communicative Research] – 2014. – №. 1. pp. 241–251. (In Russian) = Богданова-Бегларян Н.В. Еще раз о законе экономии в повседневной спонтанной речи // Коммуникативные исследования. – 2014. – №. 1. С. 241–251.

[McShane et al., 2015] McShane, M., Babkin, P. (2015) Automatic Ellipsis Resolution: Recovering Covert Information from Text // AAAI. – 2015. – pp. 572-578.

[Kasaeva, 2014] Kasaeva, Z. (2014) Ob osnovnyh chertah ustnoj spontannoj rechi (obshchie zamechaniya) //Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Russkaya filologiya [On the main features of spontaneous speech (general observations) // Vestnik of the Moscow State Regional University. Series: Russian philology.] – 2014. – №. 1. pp. 241–251. (In Russian) = Касаева З. Об основных чертах устной спонтанной речи (общие замечания) //Вестник Московского государственного областного университета. Серия: Русская филология. – 2014. – №. 2. – С. 60-68.

[Testelets, 2011] Testelets, Ya. (2011) Ellipsis v russkom yazyke: teoreticheskij i opisatel'nyj podhody //Konferenciya «Tipologiya morfosintaksicheskih parametrov»: doklad [Ellipsis in Russian: theoretical and descriptive approaches // Conference «Typology of morphosyntactic parameters»: report.] —M.: MGGU. – 2011. (In Russian) = Тестелец Я.Г. Эллипсис в русском языке: теоретический и описательный подходы //Конференция «Типология морфосинтаксических параметров»: доклад.—М.: МГГУ. – 2011.

[Merchant, 2005] Merchant, J. (2005) Fragments and ellipsis // Linguistics and philosophy. – 2005. – Т. 27. – №. 6. – pp. 661-738.

[Kamagina, 2015] Kamagina, I. (2015) Ellipsis i skhodnye sintaksicheskie yavleniya v sovremennom russkom yazyke //Vestnik RUDN. Seriya: Teoriya yazyka. Semiotika. Semantika [Ellipsis and similar syntactic phenomena in the modern Russian language //Vestnik RUDN. Series: Theory of Language. Semiotics. Semantics.] — 2015. – №. 1. 168–174. (In Russian) = Камагина И. Эллипсис и сходные синтаксические явления в современном русском языке //Вестник РУДН. Серия: Теория языка. Семиотика. Семантика. – 2015. – №. 1. С.168–174.

[Kovriguina et al., 2018] Kovriguina, L., Shilin, I., Putintseva, A., and Shipilo, A. (2018) Multilevel Annotation in the Corpus for Parsing Russian Spontaneous Speech // 20th International Conference, SPECOM 2018, Leipzig, Germany. — 2018, Proceedings - 2018, pp. 311-320

[Paducheva, 2013] Paducheva, E.V. (2013) Russkoe otricatel'noe predlozhenie. M.: Yazyki slavyanskoj kul'tury. [Russian negative sentence. M.: Languages of Slavic Culture] — 2013. pp. 166. (In Russian) = Падучева Е. Русское отрицательное предложение. М.: Языки славянской культуры, 2013. с. 166.

[Paducheva, 2004] Paducheva, E.V. (2004) Dinamicheskiye modeli v semantike leksiki. M.: Yazyki slavyanskoj kul'tury. [Dynamical models in lexical semantics. M.: Languages of Slavic Culture] — 2004. 608 p. (In Russian) = Падучева Е.В. Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004. - 608 с.

[Chebanov, 1984] Chebanov, S. (1984) Predstavleniya o forme v estestvoznanii i osnovaniya obshchej morfologii [Notion of the forma in natural science and the foundations of general morphology//Orgaanilise vormiteooria Tartu. TRU] – 1984. – pp. 25-41. (In Russian) = Чебанов С.В. Представления о форме в естествознании и основания общей морфологии //Orgaanilise vormiteooria Tartu. TRU. – 1984. – С. 25-41.

[Testelets et al., 2005] Testelets, Ya., Bylinina, E. (2005) O nekotoryh konstrukciyah so znacheniem neopredelennyh mestoimenij v russkom yazyke: amal'gamy i kvazirelyativy //Doklad k seminaru «Teoreticheskaya semantika» IPPI RAN [On some constructions with the meaning of indefinite pronouns in the Russian language: amalgams and quasi-relativities // Report to the Seminar "Theoretical semantics"IPPI RAN] – 2005. (In Russian) = Тестелец Я., Былинина Е. О некоторых конструкциях со значением неопределенных местоимений в русском языке: амальгамы и квазирелятивы //Доклад к семинару «Теоретическая семантика» ИППИ РАН. – 2005.

[Merchant, 2016] Merchant, J. (2016) Ellipsis: A survey of analytical approaches // Ms., University of Chicago, Chicago, IL. – 2016.

[Kibrik et al., 2017] Kibrik, A., Podlesskaya, V. (2017) Rasskazy o snovideniyah: Korpusnoe issledovanie ustnogo russkogo diskursa [Stories about dreams: Case study of spoken Russian discourse] – Litres, 2017. (In Russian) = Кибрик А., Подлесская В. Рассказы о сновидениях: Корпусное исследование устного русского дискурса. – Litres, 2017.

[Kolokonte, 2008] Kolokonte, M. (2008) Bare argument ellipsis and Information Structure. – 2008.

[Bogdanov, 2012] Bogdanov, A. (2012) Opisanie geppinga v sisteme avtomaticheskogo perevoda //Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog" [Gapping representation in the machine translation system // Proceedings of the annual International Conference "Dialogue"] – 2012. – V. 2. pp. 61–70. (In Russian) = Богданов А. Описание гэппинга в системе автоматического перевода //По материалам ежегодной Международной конференции "Диалог". – 2012. – Т. 2. С. 61–70.

[Frolova, 2012] Frolova, T. (2012) Opisanie sintaksicheskogo otsutstviya v korpusah SinTagRus i HANKO. [Ellipsis description in SinTagRus and HANCO corpora] – 2012. (In Russian) = Фролова Т. Описание синтаксического отсутствия в корпусах СинТагРус и ХАНКО. - 2012.

[Mal'kovskij et al., 2014] Mal'kovskij, M., Starostin, A., Minyajlov, V. (2014) Vosstanovlenie ellipsisa kak zadacha avtomaticheskoj obrabotki tekstov //Programmnye produkty i sistemy. [Ellipsis recovery as a task of automatic text processing // Software products and systems.] – 2014. – №. 3 (107). pp.32–36. (In Russian) = Мальковский М., Старостин А., Миняйлов В. Восстановление эллипсиса как задача автоматической обработки текстов //Программные продукты и системы. – 2014. – №. 3 (107). С.32–36.

[Mal'kovskij et al., 2012] Mal'kovskij, M., Kuzina, L., Rodimova, P. (2012) Modifikaciya algoritma sintaksicheskogo analiza sistemy TREETON, orientirovannaya na analiz ellipticheskih predlozhenij //Sbornik nauchnyh trudov SWorld po materialam mezhdunarodnoj nauchno-prakticheskoj konferencii. [Elaboration of the parsing algorithm in the TREETON system, focused on the analysis of elliptic sentences //Proceedings of the SWorld international scientific and practical conference.] - V. 4,

№. 2. 2012. pp. 47–50. (In Russian) = Мальковский М., Кузина Л., Родимова П. Модификация алгоритма синтаксического анализа системы TREETON, ориентированная на анализ эллиптических предложений //Сборник научных трудов SWorld по материалам международной научно-практической конференции. - Т. 4, №. 2. 2012. С. 47–50.

[Minyajlov, 2016] Minyajlov, V. (2016) Obnaruzhenie ellipticheskih predlozhenij sredstvami algoritmov klassifikacii //Problemy fiziki, matematiki i tekhniki. Ser.: Informatika. [Detection of elliptic sentences using classification algorithms. //Problems of physics, mathematics and engineering. Ser.: Computer science.] - 2016. №. 3 (28). pp. 82–87. (In Russian) = Миняйлов В. Обнаружение эллиптических предложений средствами алгоритмов классификации //Проблемы физики, математики и техники. Сер.: Информатика. - 2016. №. 3 (28). С. 82–87.

[Giuliani et al., 2014] Giuliani, M., Marschall, T., Isard, A. (2014) Using Ellipsis Detection and Word Similarity for Transformation of Spoken Language into Grammatically Valid Sentences // Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). – 2014. – pp. 243-250.

[Nielsen, 2004] Nielsen, L. A. (2004) Verb Phrase Ellipsis detection using Automatically Parsed Text // Proceedings of the 20th international conference on Computational Linguistics. – Association for Computational Linguistics, 2004. – pp. 1093.

[Miller, 1956] Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information // Psychological review. – 1956. – V. 63. – №. 2. – pp. 81.

[Shitaoka et al, 2004] Shitaoka, K., Uchimoto, K., Kawahara, T., Isahara, H. (2004) Dependency structure analysis and sentence boundary detection in spontaneous Japanese // Proceedings of the 20th international conference on Computational Linguistics. – Association for Computational Linguistics, 2004. – pp. 1107.