Синтаксическая несовместимость как свойство линейной организации русского предложения[1]

Syntactic incompatibility as a property of the linear organization of a Russian sentence

Кобзарева Т. Ю. (stamstam@mtu-net.ru)

Российский государственный гуманитарный университет

Обсуждается одно из свойств организации линейной структуры русского предложения, важное для автоматического анализа, — синтаксическая несовместимость: невозможность одновременного появления некоторых компонент предложения в его фрагментах, заданных знаками препинания и сочинительными союзами. Рассматривается использование этого свойства на разных этапах автоматического анализа.

1. Введение

При синтаксическом анализе мы интерпретируем линейную структуру предложения (S): «В основе всего структурного синтаксиса лежит соотношение между структурным порядком и порядком линейным» [1]. Используя словарную информацию и информацию, которая закодирована порядком слов и знаков препинания (ЗП), мы ставим в соответствие анализируемым фрагментам линейной последовательности слов и ЗП некоторую грамматическую структуру.

Речь пойдет об особенностях линейной организации русского S, которые используются в системе поверхностно-синтаксического анализа русского S, разрабатываемой в настоящее время в РГГУ, и которые, как нам представляется, важны для любой системы, ориентированной на максимальное использование информации порядка слов и ЗП при минимизации лексико-семантической информации.

Специфику системы определяет впервые использованная при синтаксическом анализе русского предложения иерархия процедур анализа, рассмотренная и обоснованная в [2]. Самой важной ее особенностью является то, что сегментация — поиск границ сегментов (простых и придаточных S, деепричастных, причастных и др. обособленных оборотов) с одновременным элиминированием разрывов, возникающих при вложении сегментов в сегменты — предшествует моделированию внутренней структуры сегментов и отношений между ними, т. е. до построения большей части синтагматических связей.

Система состоит из 6 процедурно независимых модулей:

1) постморфология — несловарные проблемы морфанализа: обработка имен собственных и названий, окказиональных аббревиатур, числительных [13];
2) разрешение частичной омонимии совпадения форм разных частей речи [3];
3) предсегментация — построение проективных фрагментов именных и предложных групп [14], т.е. поиск хозяина необособленного согласованного определения, выраженного прилагательным или причастием, конструкций с числительными, сложных сказуемых и др., т.е. связей, определяющих единицы линейной структуры, необходимые для сегментации;
4) сегментация — построение сегментов [9];
5) моделирование структуры синтагматических связей внутри сегментов;
6) построение связей сегментов [15];.

На каждом этапе встают проблемы разрешения неоднозначности интерпретации S, которые часто порождаются потенциальной неоднозначностью некоторой компоненты его линейной структуры — отдельного слова [3,4], ЗП [5] или некоторого фрагмента предложения — последовательности слов и\ или ЗП [6,7]. Иногда это имплицирует истинную неоднозначность, то есть возможность нескольких правильных с т. зр. носителей языка пониманий всего S, но чаще правильным бывает только одно из гипотетических значений.

[1] Доклад подготовлен при частичной поддержке РФФИ грант № 09-06-00275-а

Ниже мы рассмотрим одно из свойств линейной структуры русского S — свойство синтаксической несовместимости — и покажем, что это свойство на разных уровнях анализа — в разных модулях — может помочь в определенных ситуациях найти правильную интерпретацию линейной структуры S.

2. Используемые понятия

В основе предлагаемого подхода лежат представления Люсьена Теньера о структуре ситуаций, представленных простыми S: «Глагольный узел, который является центром предложения в большинстве европейских языков, выражает своего рода маленькую драму. Действительно, как в какой-нибудь драме, в нем обязательно имеется действие, а чаще всего также действующие лица и обстоятельства» [1].

В каждом языке есть много способов объединения этих «маленьких драм» в одно S, где несколько ситуаций образуют синтаксическое единство. При этом в русском S исходные ситуации, каждая из которых могла бы быть выражена отдельным простым S с некоторой вершиной–предикатом, представлены, во-первых, простыми S, составляющие основу любого S: простые в составе сложносочиненных и простые–главные в сложноподчиненных. Во-вторых, — трансформами исходных простых S — придаточными S и разными оборотами — деепричастными, причастными и другими, которые подчинены простым-главным или таким же трансформам. В качестве обобщающего названия для всех таких частей S будем использовать уже упомянутый выше термин сегмент [8,9].

В русском письменном языке исторически сложилась традиция задавать границы сегментов при помощи ЗП. В настоящее время существуют правила пунктуации, которые регламентируют соответствующее использование ЗП. В некоторых ситуациях границами сегментов могут быть и сочинительные союзы (СС) или комбинации ЗП и СС [9]. ЗП, СС и морфологически автономные [12] комбинации ЗП и СС будем называть операторами (F).

При сочинении двух слов в тексте между ними тоже обязательно есть сочиняющий их F. Одна из проблем анализа связана с омонимией операторов: F могут не только быть границами сегментов, но и манифестировать сочинение слов и\или сегментов [5].

3. Проективность и линейная организация предложения

Свойство проективности, открытое Теньером применительно к подчинительным связям слов в простых S, хорошо исследовано [10]. Если, изображая граф связей слов, мы по горизонтали сохраняем порядок слов в S, а слова — узлы графа располагаются на разных уровнях соответственно иерархии подчинительных связей, то «для правильных синтаксических структур, изображенных в виде дерева, <…> перпендикуляры, опущенные из узлов дерева, не пересекают его ветвей».

Легко заметить, что в проективном фрагменте текста между двумя связанными словами могут находиться только слова, прямо или опосредованно им подчиненные. Из проективности графа зависимостей слов внутри сегмента вытекает проективность сегментов: непосредственные или опосредованные слуги предикативных вершин сегментов не могут находиться внутри линейного пространства сегмента, имеющего другую вершину. Это означает, что каждая из исходных ситуаций, которую манифестирует один сегмент, локализуется в своем участке линейного пространства S: S делится операторами на сегменты, и каждый сегмент представлен отдельной частью линейного представления S, не пересекающейся с другими сегментами.

Проективность связей внутри сегментов в каких-то случаях нарушается: нарушение внутрисегментной проективности обычно не ведет к непониманию или, во всяком случае, не очень сильно его осложняет. Но проективность сегментов практически никогда не нарушается. Соответственно, сегмент — часть линейной структуры предложения, где могут находиться только непосредственные или опосредованные слуги вершины этого сегмента.

У каждого сегмента обязательно есть левая и правая границы, задаваемые операторами (кроме левой границы сегмента в самом начале S). При сочинении двух слов в тексте между ними тоже обязательно есть сочиняющий их F. Эта функциональная омонимия операторов осложняет анализ [5].

Линейная структура сегмента — простого-главного или придаточного S может быть дополнительно осложнена сочинением предикатов, при этом в сегменте появляется несколько вершин.

Как было показано в [11], операторы, сочиняющие предикаты, являются границами зон влияния этих предикатов: Fk, «сочиняющий» два предиката, делит отрезок между ними на две зоны — части S, где могут находятся только непосредственные или опосредованные слуги одного из них и не могут — слуги другого.

Таким образом, в S

- между любыми двумя предикатами (будь то вершины разных сегментов или сочиненные сказуемые одного сегмента) **обязательно есть Fk,** разделяющий в линейной структуре зоны их влияния;
- в сегменте не могут находиться непосредственные или опосредованные слуги вершины другого сегмента;

- в зоне влияния одного из предикатов, сочиненных внутри простого или придаточного предложения, не могут находиться слуги других предикатов.

Из сказанного следует и для анализа существенно, что мы еще до построения сегментов знаем, что между каждыми двумя вершинами сегментов обязательно должен найтись хотя бы один оператор.

Выделим предикативные вершины сегментов:

(1) *Хлестаков* ***порхает*** *по пьесе, не* ***желая*** *толком понять, какой он* ***поднял*** *переполох, и жадно* ***стараясь*** *урвать все, что* ***подкидывает*** *ему счастливый случай. (В. Набоков, далее (Н))*

(2) *Иван,* ***зная*** *все это, заблаговременно* ***запасся*** *двумя вязками бубликов и колбасою и,* ***спросивши*** *рюмку водки, в которой не* ***бывает*** *недостатка ни в одном постоялом дворе,* ***начал*** *свой ужин,* ***усевшись*** *на лавке перед дубовым столом, вкопанным в глиняный пол. (Н. Гоголь)*

На этапах, когда сегменты еще не построены и функции операторов еще не определены, мы можем использовать тот факт организации линейной структуры, что между каждыми двумя вершинами обязательно должен быть Fk.

Безусловными вершинами сегментов являются глаголы в личной форме, деепричастия, краткие причастия и краткие прилагательные. Последние — с учетом возможности их вхождения в сложные сказуемые. Для полного причастия и прилагательного необходимо убедиться, что оно является вершиной сегмента — обособленного определительного оборота.

4. Свойство синтаксической несовместимости

Рассмотренные выше особенности задают простое, но важное свойство организации линейной структуры русского S, которое мы назовем принципом **синтаксической несовместимости**: внутри сегмента, а при наличии сочинения вершин — внутри зоны влияния одной из вершин сегмента — не могут одновременно находится слова, относящиеся к ситуациям разных предикативных вершин.

5. Использование принципа синтаксической несовместимости при синтаксическом анализе

Принцип синтаксической несовместимости может быть использован на разных этапах анализа. Рассмотрим его использование в двух модулях: 1) модуле разрешения морфологической неоднозначности и 2) в модуле сегментации.

1) Использование свойства синтаксической несовместимости при разрешении омонимии частей речи

Этап морфологического анализа является необходимым для синтаксической интерпретации S. На этом этапе порождается, в частности, неоднозначность морфологической интерпретации слов из-за случайных совпадений отдельных словоформ, принадлежащих лексемам разных частей речи.

В русском языке существует около 60 типов омонимии частей речи [2,3]. Важными являются типы омонимии, где одна из омонимичных частей речи — потенциальная вершина сегмента. К таким типам относятся, например:

- личная форма глагола vs. существительное (*белил, берегу, бури, вил…*)
- краткое прилагательное \ краткое причастие vs. наречие (*совершенно, дико, двусмысленно, забавно…*)
- краткое прилагательное \ краткое причастие (vs. существительное (*весел, весом, гол, долги…*)
- деепричастие vs. предлог (*для, благодаря, включая…*)
- деепричастие vs. полнозначное или местоименное полное прилагательное, \полное причастие (*скупая, строгая, заезжая, моя…*)

и др.

Один из способов разрешения омонимии — анализ грамматического контекста: проверка того, какие части речи и ЗП и в каком порядке окружают омоним. Именно такой способ разрешения частичной омонимии принят в описываемой системе. Анализ грамматического контекста хорош тем, что, задавая не очень большой, но достаточно представительный набор конфигураций, разрешающих омонимию каждого типа, мы создаем для русского языка фрагмент грамматики линейной структуры, который не требует лексической конкретизации и который можно далее пополнять и унифицировать.

Из свойства синтаксической несовместимости вытекает одно из условий такого фрагмента грамматики, действительное для всех перечисленных типов омонимии. Простота его проверки и довольно высокая частота линейных ситуаций, когда оно позволяет разрешить омонимию частей речи является веским аргументом для использования его перед более сложным анализом контекста.

Условие синтаксической несовместимости для разрешения омонимии Li= потенциальная предикативная вершина vs. Ln= не предикативная вершина: если у нас есть частичный омоним=Lo, одно из значений которого потенциально имеет синтаксическую функцию предикативной вершины, а в исследуемом контексте есть неомонимичный морфологический предикат=Pr и между Lo и Pr нет оператора, Lo не может иметь значение предикативной вершины.

Таким образом, если мы встречаем частеречные омонимы во фрагментах типа (омонимы подчеркнуты, а вершины выделены) *странно на меня* ***смотрит****; мог с тех пор стать совершенно другим; существенно* ***изменив****; решение, значительно с этого времени* ***пересмотренное****; на краю села к тому времени уже* ***построили****…; немедленно* ***взяв*** *в руки жгут;,* ***нет*** *у меня клея****; поднялась*** *буря* и т.д., где между омонимом указанных типов — а группы таких типов в словаре велики и в текстах часты — и неомонимичной предикативной вершиной нет F, алгоритм на основании условия синтаксической несовместимости легко разрешает омонимию.

Ниже в примерах (3) и (4) подчеркнуты интересующие нас омонимы и выделены полужирным неомонимичные Pr. В скобках справа от омонима приводится тип омонимии в результате морфанализа, в данном случае — «краткое прилагательное (Abr) vs. наречие (D)», а справа от стрелки — результат разрешения этой омонимии в этих контекстах в силу выполнения условия синтаксической несовместимости.

(3) *Рядом стоял воспитатель, и, когда серый резиновый мяч, которым играли в футбол,* ***подкатился*** *случайно (Abr\D → D) к его ногам, учитель словесности, инстинктивно продолжая очаровательное предание, сделал вид, что хочет его пнуть, неловко (Abr\D → D)* ***потоптался****, чуть не потерял голову и рассмеялся с большим добродушием.(H)*

(4) *Наплакавшись вдоволь, он поиграл с жуком, нервно (Abr\D → D)* ***поводившим*** *усами, и потом давил его камнем, стараясь повторить первоначальный сдобный хруст (H)*

2) Использование свойства синтаксической несовместимости при построении сегментов — при разрешении функциональной омонимии ЗП.

Синтаксическая несовместимость позволяет в некоторых случаях чрезвычайно упростить процедуру анализа и на этапе сегментации.

Для того, чтобы иметь возможность находить сегменты в S любой сегментной структуры, удобно использовать, как уже было показано в [2,9] рекурсивную процедуру, строящую сегменты в S справа налево. Для того, чтобы было понятно использование при этом принципа несовместимости, кратко опишем эту процедуру.

Используется следующая модель: русское S состоит из цепочки **β-сегментов** — простых (сочиненных и главных) предложений, каждое из которых может быть разорвано вложением в него некоторого числа **α-сегментов:** обособленных согласованных определений, деепричастных, предложных, вводных и сравнительных оборотов и придаточных S. Каждый **α-сегмент** в свою очередь может быть разорван следующего уровня вставлениями **α-сегментов,** причем количество вставлений, как и количество уровней вставлений, теоретически не ограничено.

Построение сегментов включает в себя 3 этапа:

(1) определение левых границ **α-сегментов: поиск α-отрезков —** безусловных минимальных левых компонент **α-сегментов**;

(2) построение **α-сегментов — поиск** правых границ **α-сегментов** с одновременным восстановлением целостности **α-сегментов,** разорванных вложениями;

(3) определение границ **β-сегментов** с элиминированием разрывов.

Правые границы определяет рекурсивный алгоритм, который, рассматривая поочередно справа налево в S минимальные отрезки **α-сегментов —** отрезки, определенные как безусловные фрагменты **α-сегментов** при установлении левых границ, и двигаясь слева направо от обрабатываемого **α-отрезка,** присоединяет еще не идентифицированные β-отрезки, если это грамматически допустимо, к **α-отрезку,** построенному к текущему моменту анализа.

Процедура повторяется соответственно правилам присоединения с учетом проективности **α-сегментов,** уже построенных правее анализируемого отрезка, пока не находим его правую границу.

Эта процедура начинает построение сегментных матрешек с самого глубокого вложения и позволяет анализировать на основе небольшого базисного синтаксического словаря возможных линейных конфигураций любые допустимые комбинации.

Рассмотрим построение сегментов на следующем примере. Определяем на первом шаге **α-** и **β**-отрезки:

(5) **β**=[*Едва уловимую особенность*], $\boldsymbol{\alpha}_7$=[***отличавшую*** *его сына от всех тех детей*], $\boldsymbol{\alpha}_6$=[*которые по его мнению должны были стать людьми*], $\boldsymbol{\alpha}_5$=[*ничем не* ***замечательными***], **β**=[*он* ***понимал*** *как тайное волнение таланта*], *и*], $\boldsymbol{\alpha}_4$=[*твердо* ***помня***], $\boldsymbol{\alpha}_3$=[*что* ***покойный тесть*** *был композитором*], **β**=[***он*** *в приятной мечте*], $\boldsymbol{\alpha}_2$=[*похожей на литографию, спускался ночью со свечой в гостиную*], $\boldsymbol{\alpha}_1$=[*где вундеркинд в белой рубашонке до пят играет на огромном черном рояле*]. (H)

На втором этапе строятся **α**-сегменты. Ниже (Рис.1) приведена условная схема движения по исходным отрезкам примера (5) при построении в нем **α**-сегментов.

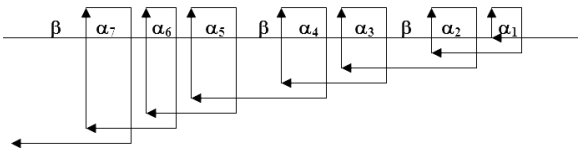

Рис. 1

β-отрезок — это часть S между двумя ЗП, в которой нет слов, маркирующих **α**-отрезки — минимальные правые составляющие **α**-сегментов. При анализе **β**-отрезок, ближайший к **α**-отрезку справа, может быть присоединен к нему тогда и только тогда, если хотя бы одно слово этого отрезка связано подчинительной или сочинительной связью с некоторым словом уже построенной части сегмента. Однако проверка того, имеется ли такая связь, сложна, громоздка и не всегда однозначна.

Использование принципа несовместимости позволяет во многих (но, естественно, не во всех) ситуациях этих проверок избежать. Если в очередном **β**-отрезке, который мы, удлиняя строимый сегмент, пытаемся присоединить, есть слово, которое по своим морфологическим характеристикам не может принадлежать строимому **α**-сегменту, это — в силу проективности сегментов [9] — означает, что построение очередного **α**-сегмента закончено.

Рассмотрим в (5) ситуации, где работает принцип несовместимости. При попытке присоединить к $\boldsymbol{\alpha}_4$=[*твердо* ***помня***] отрезок **β**=[***он*** *в приятной мечте*] мы видим, что в **β**-отрезке есть свободный неомонимичный Им.п. На этапе предсегментации мы уже построили именные группы с существительными в Им. п. в роли определений. Поэтому свободный Им.п. не может появиться внутри деепричастного оборота: деепричастие и свободное (по результатам анализа в модуле предсегментации) существительное в Им. п. не могут относиться к одному сегменту, т. е. синтаксически несовместимы.

Синтаксическая несовместимость деепричастия и существительного в Им.п. позволяет в некоторых случаях разрешить падежную омонимию Им.п.\ Вин п. В примере (2) в деепричастном обороте ***зная*** *все это* можно снять Им.п. у слов *все* и *это*, что в каких-то случаях может предупредить ошибки при анализе сочинения существительных.

При построении согласованного определения $\boldsymbol{\alpha}_5$=[*ничем не* ***замечательными***] ближайший справа **β**=[*он* ***понимал*** *как тайное волнение таланта*], свободный к моменту построения этого сегмента, не м.б. присоединен, так как предикативная вершина — глагол в личной форме — не может находиться в одном сегменте с ***замечательными*** — предикативной вершиной обособленного определения. По тем же причинам этот **β**-отрезок не может быть присоединен и к $\boldsymbol{\alpha}_7$.

Таким образом, если мы строим деепричастный или причастный оборот, **β**-отрезок справа от него при его удлинении не может быть к нему присоединен, если в этом **β**-отрезке есть глагол в личной форме или свободный Им.п.

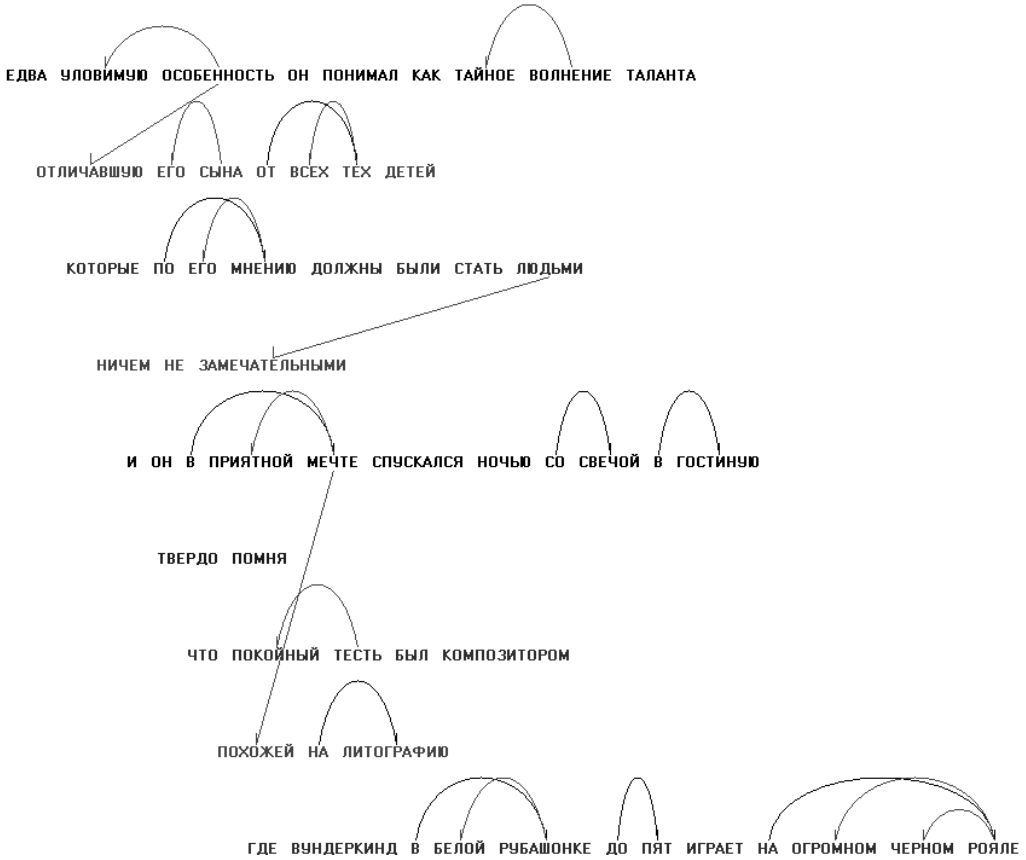

Рис. 2. Визуализация результата анализа примера (5), использующего в ходе рекурсивной процедуры построения сегментов принцип несовместимости (экспериментальная реализация модулей предсегментации и сегментации И. М. Ножова)

Таким образом, принцип несовместимости определят одно из простых и часто работающих в тексте условий присоединения. Для каждого типа **α**-сегментов можно задать список компонент, наличие которых в ближайшем справа **β**-отрезке означает, что этот отрезок не может быть присоединен и при этом — в силу проективности сегментов — построение рассматриваемого **α**-сегмента закончено.

6. Заключение

Рассмотрено свойство синтаксической несовместимости — важное свойство линейной организации русского S, которое можно использовать на разных уровнях синтаксического анализа для разрешения потенциальных неоднозначностей интерпретации S. Это свойство определяется семантикой линейной организации русского предложения и, в частности, является следствием исторически сложившейся в русском языке традиции чрезвычайно семантикализованного использования ЗП в письменном русском языке. Рассмотрение свойства синтаксической несовместимости еще раз показывает, насколько осмысление функциональной семантики русских ЗП, отражающей особенности смысловой организации линейной структуры русского S, может быть полезно для автоматического анализа.

Литература

1. *Теньер Люсьен,* Основы структурного синтаксиса. — М.: Прогресс, 1988.
2. *Кобзарева Т. Ю.* Иерархия задач поверхностно-синтаксического анализа русского предложения // НТИ, Сер.2, №1, 2007 — С 23–35.
3. *Кобзарева Т. Ю., Афанасьев Р. Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций // Труды международного семинара Диалог'2002 Протвино 2002. Т. 2. С 258–268.
4. *Зинькина Ю. В., Пяткин Н. В., Невзорова О. А.* Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005. — М.: Наука, 2005. С. 198–202.
5. *Кобзарева Т. Ю.* Омонимия и синонимия знаков препинания в русском тексте // Труды Международной конференции Диалог'2005. — М.: Наука, 2005. — С. 233–237.
6. *Иорданская Л. Н.* Синтаксическая омонимия в РЯ (с точки зрения автоматического анализа и синтеза). НТИ, сер. 2. 1967, №5 — С 9–17
7. *Дрейзин Ф. А.* Синтаксическая омонимия // Машинный перевод и прикладная лингвистика. М., 1988
8. *Мельчук И. А.* Автоматический синтаксический анализ. Т. 1. — Новосибирск.: Ред.-изд. отдел Сибирского отделения АН СССР, 1964.
9. *Кобзарева Т. Ю.* Принципы сегментационного анализа русского предложения. Московский лингвистический журнал // М., 2004. Т. 8. №1 — С. 31–80
10. *Иорданская Л. Н.* Автоматический синтаксический анализ. Т. 2. // Новосибирск: Наука, 1967
11. *Кобзарева Т. Ю.* Рекурсивность и проективность сочинительных связей в русском тексте // Компьютерная лингвистика и интеллектуальные технологии Труды Международной конференции Диалог 2006, Бекасово, 31 мая — 4 июня 2006 г. — М.: Наука, 2006. — С. 223–229.
12. *Плунгян В. А.* Общая морфология. Введение в проблематику. М., 2003.
13. *Кобзарева Т. Ю.* Морфанализ in vivo // Труды Международной конференции Диалог'2004, — М.: Наука, 2004 — С. 286–291.
14. *Кобзарева Т. Ю.* Некоторые свойства линейной структуры именных и предложных групп (Поверхностно-синтаксический анализ русского предложения)// Вестник РГГУ. № 8/07, Серия «Языкознание» (Московский лингвистический журнал № 9/2), Москва 2007. — С. 113–130.
15. *Кобзарева Т. Ю.* Построение графа связей сегментов (поверхностно-синтаксический анализ русского предложения) // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог, М.: Наука, 2008 — С. 192–198.