СИНТАКСИЧЕСКИЕ КОРРЕЛЯТЫ ПРОСОДИЧЕСКИ МАРКИРОВАННЫХ ЭЛЕМЕНТОВ ПРЕДЛОЖЕНИЯ И ИХ РОЛЬ В ЗАДАЧАХ СИНТЕЗА РЕЧИ ПО ТЕКСТУ[1]

SYNTACTIC CORRELATES OF PROSODICALLY MARKED ELEMENTS OF THE SENTENCE AND THEIR ROLE IN THE TASKS OF TEXT-TO-SPEECH SYNTHESIS

Иомдин Л.Л. **(*iomdin@iitp.ru*),** ***Институт проблем передачи информации РАН им А.А.Харкевича, Москва***

Лобанов Б.М. **(*lobanov@newman.bas-net.by*),** ***Объединенный институт проблем информатики НАН Беларуси, Минск***

Аннотация

Работа посвящена экспериментальному исследованию возможности использования синтаксического анализа письменного текста на начальном этапе алгоритма синтеза речи по тексту. Произведена попытка установить корреляции между элементами построенной автоматически синтаксической структуры предложения в виде дерева зависимостей, и просодически выделенными элементами этого предложения. Первые результаты эксперимента показывают, что данный подход имеет хорошие перспективы.

Введение

Синтез речи по тексту предполагает наличие автоматической процедуры формирования текущих контуров мелодии, силы звука, фонемной длительности и длительности пауз на основе анализа определенных свойств входного текста и его просодической разметки. Просодическая разметка текста заключается в его членении на синтагмы, разметке синтагм на акцентные единицы и маркировке интонационного типа синтагм в соответствии с определёнными правилами. В [1] и более подробно в [2] были описаны правила просодической разметки текста на основе его частичного синтаксического анализа (анализа словосочетаний) и указывалось, что в достаточно полной степени эта проблема может быть решена лишь с использованием глубокого синтаксического анализа. Приемлемой лингвистической теорией, на которой может строиться система такого анализа, представляется теория «Смысл Û Текст» И.А. Мельчука (см., например, [3]).

В данной работе исследуются связи между элементами синтаксической структуры предложения и экспертной просодической разметкой предложений на примере текстов новостных телевизионных передач на русском языке. Исследование таких связей стало возможным благодаря разработке системы «ЭТАП-3» [4,5] и созданию базы данных «Интонация русских информационных текстов» [6]. Опираясь на относительно небольшой фрагмент этой базы данных, мы попытались с помощью системы ЭТАП-3 получить предварительные ответы на следующие вопросы:

1. Существуют ли статистически значимые синтаксические корреляты **просодического**

выделенния слов в синтагмах?

2. Существуют ли статистически значимые синтаксические корреляты членения предложений на **предпаузальные и беспаузальные синтагмы**?
3. Существуют ли статистически значимые синтаксические корреляты особенностей членения предложений на предпаузальные и беспаузальные синтагмы, а также просодического выделения слов в синтагмах **для различных дикторов?**
4. Существуют ли статистически значимые синтаксические корреляты особенностей членения предложений на предпаузальные и беспаузальные синтагмы, а также просодического выделенния слов в синтагмах **для различных стилей речи?**
5. Если таковые закономерности существуют, то какова **числовая оценка их частотности**?

1. Синтаксический анализатор системы ЭТАП-3

Синтаксический анализатор, или парсер, многоцелевого лингвистического процессора ЭТАП-3 разработан в Лаборатории компьютерной лингвистики ИППИ РАН им А.А.Харкевича и используется в различных приложениях, в том числе в системе машинного перевода с русского языка на английский, в системе синонимического перифразирования, а также для построения синтаксически размеченного корпуса русского языка SynTagRus. [5. 7].

Этот парсер, опирающийся в значительной мере на уже упомянутую лингвистическую теорию «Смысл Û Текст», строит для каждого предложения письменного текста его синтаксическую структуру (СинтС) в виде дерева зависимостей, т.е. связного ориентированного графа без циклов. Каждый узел такого дерева соответствует некоторому слову предложения, а его дуги помечены именами синтаксических отношений (СинтО).

В СинтС каждого предложения имеется единственная вершина, которой непосредственно или опосредованно подчиняются все остальные узлы. Имена СинтО эксплицируют различные типы синтаксических связей между словами; в текущей версии парсера используется свыше 65 различных СинтО. Например, связь между глагольным сказуемым в качестве вершины и именным подлежащим при нем в качестве зависимого члена (*мальчик ←читает*) представляется **предикативным СинтО**; связь между предикатным словом и первым дополнением при нем (*читает → книгу, чтение → книги*) представляется **1-ым комплетивным СинтО**; связь между существительным и определяющим его прилагательным (*детская ← книга*) оформляется **определительным СинтО**, а связь между глаголом и наречным обстоятельством (*читает → вслух*) задается **обстоятельственным СинтО.**

СинтС предложения, генерируемая парсером ЭТАП-3, является упорядоченным деревом зависимостей – оно сохраняет информацию о порядке следования слов в предложении.

Алгоритм русского синтаксического анализа обращается к лингвистическим ресурсам двух основных типов: набору бинарных синтаксических правил, или синтагм[2], и так называемому комбинаторному словарю, содержащему богатую и разнообразную информацию о каждом входящем в него слове. Парсер работает пофразно и может функционировать в нескольких режимах, в частности, 1) в полностью автоматическом дежурном режиме, при котором для каждого предложения строится ровно одна СинтС; 2) в режиме множественного анализа, когда пользователь может потребовать от системы построить для неоднозначного предложения несколько СинтС или даже все возможные СинтС; 3) в интерактивном режиме, когда в определенных точках алгоритма парсер, встретив неоднозначную лексическую единицу или омонимичную синтаксическую конструкцию, предлагает пользователю выбрать ту или иную морфологическую, лексическую и/или синтаксическую интерпретацию элементов предложения и тем самым направить работу по некоторому конкретному пути.

Система ЭТАП-3 в целом и ее синтаксический анализатор рассчитаны в первую очередь

на тексты нейтрально-деловой прозы. Это, в частности, означает, что систему нецелесообразно применять к стилистически окрашенному материалу, к авторской художественной прозе, поэзии или же к разговорной речи.

2. Экспериментальный текстовый и аудиоматериал, отобранный для исследования

Ниже приводится экспериментальный текст, маркированный в соответствии с [3] знаками интонационной транскрипции (без указания тональных акцентов). Из общей базы данных [3] отобраны небольшие фрагменты, включающие записи (1–32) из новостных передач РТР и записи (33 – 37) из передач НТВ.

1. *[м1] Полтора часа *назад | из *Вены пришло *сенсационное *известие |, которое грозит *крупным международным *скандалом | и должно *повлиять | на судьбу арестованного в *Австрии | сотрудника || международного управления *РосКосмоса |||.*
2. *[м1] Австрийский *МИД | официально *признал |, что гражданин *России |, задержанный по подозрению в *шпионаже |, *имеет дипломатический *иммунитет |, а это *значит |, что по нормам международного *права | он не **мог быть *арестован |||.*
3. *[ж1] Москва требует немедленного *освобождения| нашего *гражданина.*
4. *[ж1] Австрийскому *послу вручена *нота |||.*
5. *[ж1] А в официальном *заявлении| сказано, что этот *шаг || властей *Австрии| расценивается как *недружественный,| наносящий *ущерб| двусторонним *отношениям || и что он не *укрепляет авторитета *Австрии,| как места *расположения *штаб-квартир || ряда международных *организаций |||.*
6. *[ж1] *Сейчас| на прямом эфире из *Вены| к нам присоединяется наш *специальный *корреспондент| Иван *Родионов |||.*
7. *[ж1] Иван, *здравствуйте |||!*
8. *[ж1] *Как из всей этой неловкой *ситуации |собираются *выходить австрийские *власти |||?*
9. *[ж1] И *почему такая *проволочка | с *подтверждением ||правового *статуса нашего *соотечественника |||?*
10. ***[м2] *Здравствуйте, *коллеги |||.***
11. ***[м2] Действительно, *сегодня| наступило || *решающее *развитие| в ситуации вокруг *арестовано российского || *сотрудника || *РосКосмоса ||.***
12. ***[м2] *Сегодня || пришло *подтверждение|| от|| правового управления || *ООН|| в *Нью-йорке |||.***
13. ***[м2] Я *позволю себе| коротко процитировать эту *бумагу || со слов| *официального представителя || *австрийского| *МинЮста| Томаса *Тайбленгера |||.***
14. ***[м2] А *австрийский *МИД || *сослался на то, что **теперь || *решение должно принимать| *юридическое *ведомство |||.***
15. ***[м2] *Позиция *нашего| дипломатического ведомства *известна:| оно, с *самого *начала ||, назвало *задержание || сотрудника *РосКосмоса| *«нарушением || международных *прав» |||.***
16. ***[м2] *Они будут *готовы *отпустить || нашего *гражданина |, как *только поступит *официальная || *реакция || российской *стороны |, и | *цитата |, как сказал Герхард *Ярыш |: «мы *готовы отпустить || *завтра ||, *самое *позднее | в *пятницу» |||.***

17. [м2] И вот *сейчас ||, уже перед *самым *эфиром |, поступила *информация |, что || российский гражданин *переведен все-таки | с *зальцбуржского следственного *изолятора |, где он был еще || сегодня *утром |||.

18. [м2] Я *говорил с представителями этого *изолятора ||, так *вот, он сегодня *вечером | уже оказался в *Вене ||, где || и | в течение *двух ближайших *дней |, если верить представителю прокуратуры ||, должен быть *освобожден |||.

19. [м2] *Сейчас || российский *посланник |, *представитель российского *посольства ||, *приглашен || в австрийский *МИД для *консультаций ||, по | поводу этой *новой | теперь уже *ситуации вокруг|| российского *гражданина |||.

*20. [ж2] В *России сегодня *зафиксированы | сразу *два *случая массовых *отравлений |||.*

*21. [ж2] Пятьдесят два *ребенка | попали в больницу | из подмосковного лагеря «*Смена» ||, и шестьдесят *четыре *человека | *госпитализированы в *Биробиджане |||.*

*22. [ж2] В *причинах разбирается *прокуратура |||.*

*23. [ж2] А в *Красноярске |, как раз *завершился суд | по похожему *делу |||.*

*24. [ж2] Оглашен *приговор в отношении *распорядителей и *поваров |, *обслуживавших в *марте губернаторский *бал ||, который больше *двухсот *гостей | покинули в *сплошном *расстройстве |||.*

*25. [ж2] В *кухне этих *происшествий | *разбирался Дмитрий *Кайстра |||.*

26. [м3] *Смена еще не *закончилась |, а в лагере с одноименным *названием | в *Рузском районе *Подмосковья | разыгралась настоящая *драма |||.

27. [м3] Почти *одновременно | *пятьдесят *детей | с диагнозом острая кишечная *инфекция | *были госпитализированы в *больницу | *поселка *Тучково |||.

28. [м3] *Решением *суда повар был *оштрафован |, а директор фирмы *поставщика | лишен *права || заниматься организацией общественного *питания | в течение *года и девяти *месяцев |||.

29. [м3] Теперь не *ясно |, *поставщики или *повара | были лучше осведомлены о качестве *сосисок ||, которыми как-то в *мае | *отобедали ученики одной из *школ | в *Нефтеюганске |||.

30. [м3] Следом за *трапезой, целый *класс | в полном *составе отправился на больничные *койки | инфекционного *отделения ||, где, к *слову сказать, и *встретил || *последний школьный *звонок |||.

31. [м3] *Врачи *говорят |, что *случаев отравления становится все *больше ||, и *причины *каждое *лето || - *одни и *те же |||.

32. [ж3] «*Чаще всего это бывает или *вода ||, опять же непригодная для *питья || *зараженная вода ||, *или | *нарушение работы | холодильного *оборудования ||, или нарушения *технологий *приготовления || пищевых *продуктов» |||.

*33. [м4] *Тысячи *пассажиров | "застряли" этой *ночью | в американском аэропорту *Лос-Анджелеса |||.*

*34. [м4] Из-за компьютерного *сбоя | в системе *идентификации || люди были вынуждены | находиться в *самолетах | более *семи *часов |||.*

*35. [м4] *Неполадки произошли | именно в *тех *компьютерах |, которые *отвечают || за предоставление *информации | о личных *данных ||, например, *сведений о нахождении в *розыске |||.*

*36. [м4] В *результате иностранные *пассажиры | не *могли пройти таможенный *контроль |||.*

*37. [м4] *Пока системные *администраторы | *устраняют *проблемы | в базе *данных ||, *прибывающие международные *рейсы | отправляют на *посадку || в*

*калифорнийский аэропорт *Онтарио | и *Лас-Вегас |||.*

В этих предложениях представлены стенограммы трех женских голосов (ж) и четырех мужских – (м), а также двух стилей речи: спонтанная речь (предложения 10-19 и 26-32, они выделены полужирным курсивом) и стиль чтения (остальные). Концы предложений отмечены знаками |||, концы предпаузальных синтагм – знаками ||, а концы интонационно выделенных беспаузальных синтагм – знаками |. Последние проставлены на основе аудитивного анализа соответствующих звуковых файлов. Просодически выделенные слова в синтагмах помечены звездочкой (*). В каждом предложении вручную проставлены знаки препинания.

3. Синтаксический эксперимент

Для целей настоящего исследования парсер ЭТАП-3 обрабатывал все предложения экспериментального корпуса в дежурном режиме: для каждого предложения строилась единственная синтаксическая структура (СинтС).

Все предложения подавались на вход парсера в практически непрепарированном письменном виде[3].

Почти для всех предложений парсер ЭТАП-3 построил адекватную СинтС. В одном случае (предложение 37) парсер неверно интерпретировал синтаксически неоднозначное высказывание и перепутал подлежащее глагола с прямым дополнением; кроме того, в ряде ситуаций были неточно установлены синтаксические хозяева предложно-именных групп. Эти погрешности, как впоследствии выяснилось, не оказали влияния на результаты эксперимента.

Ниже даются некоторые примеры полученных парсером СинтС и комментарии к ним, представляющие собой синтаксическую характеристику просодически выделенных слов в предложениях. На рис. 1 и 2 представлены СинтС предложений 1 и 5 из корпуса РТР, соответствующих записям дикторов М1 и Ж1 в стиле чтения текста, а на рис. 3 – СинтС предложения 32, соответствующего записи диктора Ж3 в стиле спонтанной речи. На рис.4 дается СинтС предложения 35 из корпуса НТВ, записанного диктором М4 в стиле чтения текста.

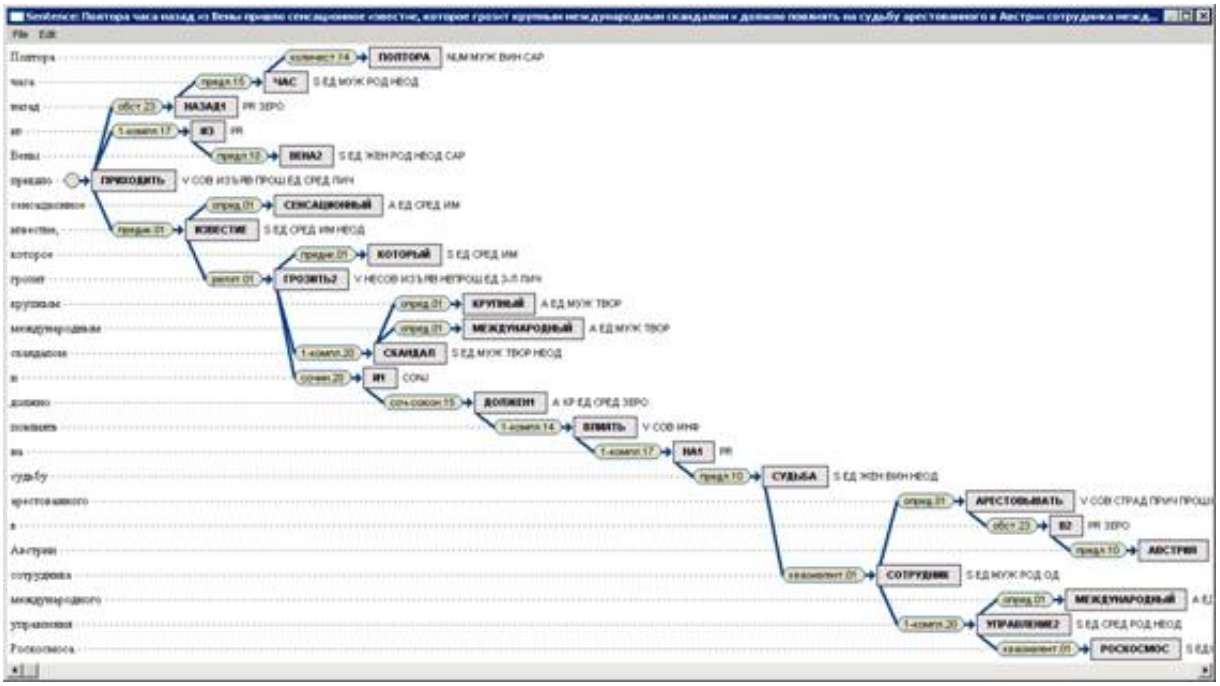

Рис. 1. СинтС предложения 1: *Полтора часа *назад из *Вены пришло *сенсационное *известие, которое грозит *крупным международным *скандалом и должно *повлиять на судьбу арестованного в *Австрии сотрудника международного управления *РосКосмоса.*

Просодически выделенные слова этого предложения синтаксически интерпретируются следующим образом: *Вены* – самый правый элемент группы первого дополнения; *назад* – самый правый элемент группы обстоятельства; *известие* – правый элемент группы подлежащего, от которой «отрезано» придаточное; *скандалом* – самый правый элемент группы первого дополнения; *Австрии* – самый правый элемент группы причастного оборота; *Роскосмоса* – самый правый элемент группы первого дополнения. Просодические выделения слов *сенсационный, крупный* и *повлиять* с точки зрения СинтС представляются случайными.

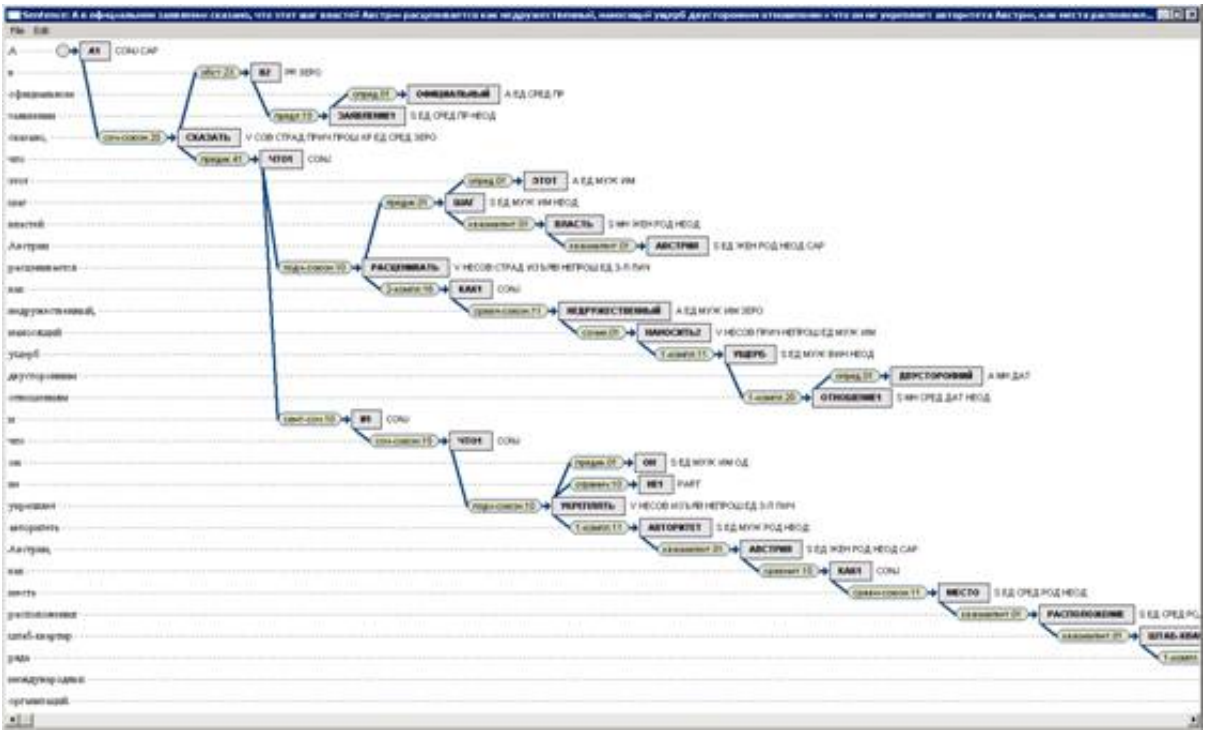

Рис. 2. СинтС предложения 5: *А в официальном *заявлении сказано, что этот *шаг властей *Австрии расценивается как *недружественный, наносящий *ущерб двусторонним *отношениям, и что он не *укрепляет авторитета *Австрии, как места *расположения *штаб-квартир ряда международных *организаций.*

В этом предложении просодически выделенные слова получают следующую синтаксическую интерпретацию: *заявлении* – самый правый элемент группы обстоятельства; *Австрии* – самый правый элемент группы дополнения; *недружественный* – правый элемент группы дополнения, от которого отрезана сочинительная группа; *ущерб* – правый элемент группы дополнения, от которой отрезана группа дополнения; *отношениям* – самый правый элемент группы дополнения; *укрепляет* – вершина второго однородного предложения; *Австрии* – правый элемент группы квазиагентивного дополнения, от которой отрезан сравнительный оборот; *штаб-квартир* – правый элемент группы квазиагентивного дополнения, от которой отрезана группа дополнения; *организаций* – самый правый элемент группы дополнения. Два акцентуированных слова и здесь представляются случайными с точки зрения СинтС: *шаг* и *расположения.*

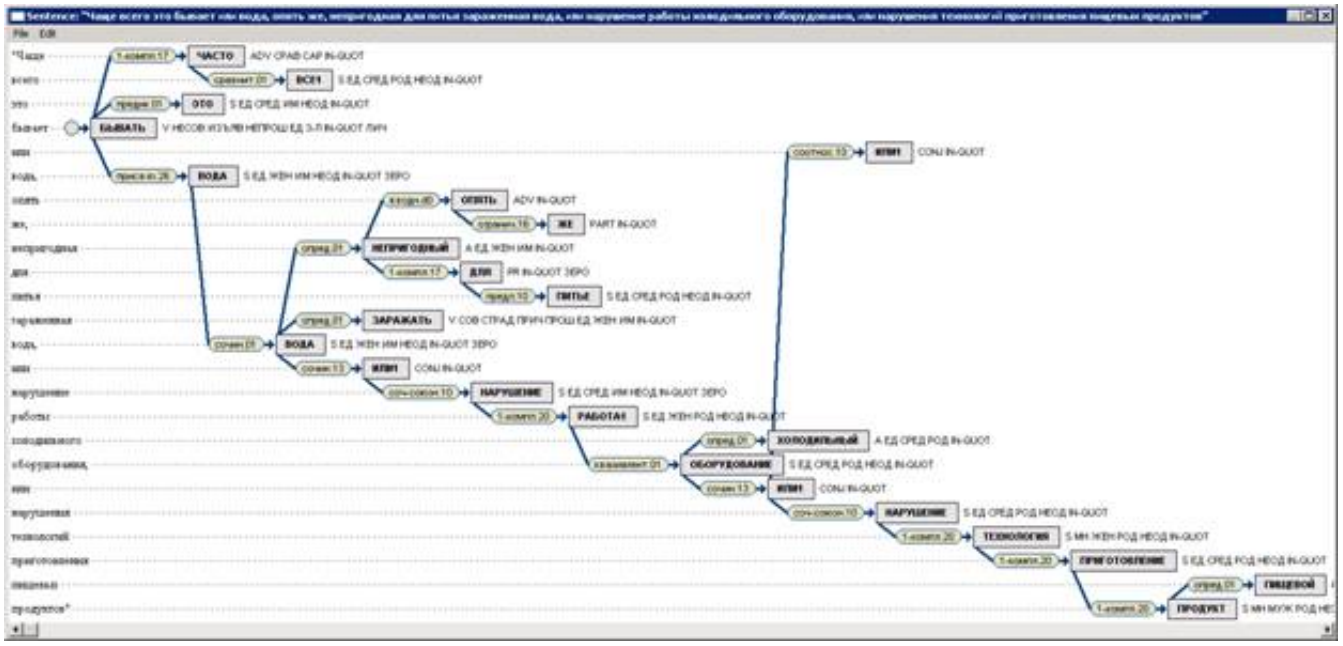

Рис. 3. Предложение 32: **Чаще всего это бывает или *вода, опять же непригодная для *питья, *зараженная вода, *или *нарушение работы холодильного *оборудования, или нарушения *технологий *приготовления пищевых *продуктов.*

В данном предложении просодически акцентуированные слова имеют такую интерпретацию: *вода* – правый элемент группы присвязочного дополнения, от которой отрезана вся сочинительная цепочка; *питья* – самый правый элемент группы дополнения; *оборудования* – правый элемент группы дополнения, от которой «отрезана» сочинительная цепочка; *технологий* – правый элемент группы дополнения, от которой «отрезана» группа внутреннего дополнения; *приготовления* – правый элемент группы дополнения, от которой «отрезана» группа внутреннего дополнения; *продуктов* – самый правый элемент группы дополнения. Акцентное выделение четырех слов - *чаще, зараженная, или, нарушение* и здесь представляется случайным. Характерно, что одно слово – *всего*, с точки зрения авторов, должно быть выделено, но фонетическая запись этого не подтверждает.

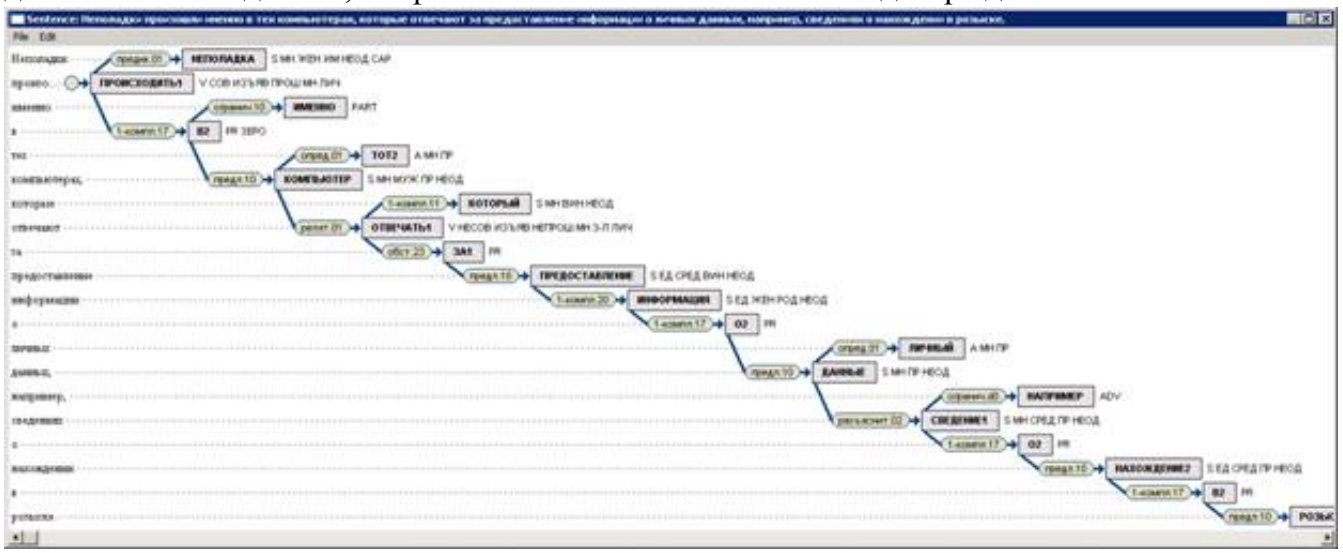

Рис. 4. Предложение 35. **Неполадки произошли именно в *тех *компьютерах , которые *отвечают за предоставление *информации о личных *данных, например, *сведений о нахождении в *розыске.*

В данном предложении просодически выделенные элементы интерпретируются следующим образом**:** *неполадки* – самый правый элемент группы подлежащего *компьютерах* –

правый элемент группы дополнения, от которой отрезано длинное придаточное; *отвечают* – вершина придаточного; *информации* – правый элемент группы дополнения, от которой отрезана внутренняя группа дополнения; *данных* – правый элемент группы дополнения, от которой отрезана разъяснительная группа; *розыске* – самый правый элемент группы дополнения; Два слова – *тех* и *сведений* – акцентуированы с точки зрения структуры случайно.

4. Обсуждение результатов

Внимательное исследование итогов эксперимента позволило авторам сформулировать, в первом приближении, несколько простых правил идентификации просодически маркированных элементов предложения [4].

А. Правила просодического выделения.

Просодически выделенными словами являются:

1) абсолютная вершина предложения;

2) вершины всех частей сложносочиненного предложения;

3) вершины всех придаточных предложений;

4) самые правые субстантивные элементы группы подлежащего, дополнения или обстоятельства при вершинах, перечисленных в пп. 1-3;

5) самый правый субстантивный элемент первой именной подгруппы в группах, перечисленных в п. 4;

6) отдельные классы лексических единиц и конкретные лексические единицы, стоящие в определенной позиции (наречия-детерминанты в начале слов, числительные и количественные существительные).

Б. Правила членения предложения на просодические синтагмы

1) Как следует из анализа изученных текстов, примерно в 90% случаев граница синтагмы выставляется непосредственно после конца просодически акцентированного слова. Остальные 10% приходятся на индивидуальную, синтаксически немотивированную установку границы синтагмы после неакцентированного слова.

2) Хорошим признаком конца синтагмы является наличие знака препинания (в более чем 90% случаев появление границы синтагмы коррелирует с присутствием такого знака).

3) Появление границы синтагмы подчиняется некоторым статистическим закономерностям и в основном зависит от стиля речи (см. табл. 1).

Таблица 1.

Статистические характеристики акцентуации и членения предложения на синтагмы

Стиль речи	Кол-во слов (всего)	Кол-во выделенных слов	Кол-во синтагм (всего)	Кол-во синтагм перед паузой	Кол-во знаков препинания	Среднее кол-во слов в синтагме
Стиль речи «чтение»	294	111 (38%)	70	31	76	4,2
«Спонтанная речь»	287	134 (47%)	105	60	55	2,7
Весь корпус	581	245 (42%)	175	91	131	3,3

5. Заключительные замечания

Из полученных результатов, по нашему мнению, логически вытекает следующий план

дальнейшей работы: пополнение синтаксического анализатора правилами маркировки просодически выделенных элементов предложения, а также правилами его синтагматического членения. Тем самым будет сделан важный шаг в сторону совершенствования системы синтеза речи за счет блока высокоуровневого синтаксического анализа.

Литература

1. Лобанов Б.М. Алгоритм сегментации текста на синтаксические синтагмы для синтеза речи // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2008). – М.: Наука, 2008. С. 323-329.
2. Лобанов Б. М. Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник // Минск: Белорусская Наука, 2008. 342 с.
3. Мельчук И.А. Опыт теории лингвистических моделей «Смысл Û Текст». Семантика, синтаксис. Отв. ред. А.А. Холодович. М. Наука. ГРВЛ. 1974г. 314 с.
4. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992. 256 с.
5. Богуславский И.М., Иомдин Л.Л., Валеев Д.Р., Сизов В.Г. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // Труды Международной конференции <Корпусная лингвистика -2008>. СПб.: Санкт-Петербургский государственный университет, 2008. С. 56-74.
6. Кодзасов С.В., Архипов А.В., Захаров Д.М., Кривнова О.Ф. База данных «интонация русских информационных текстов // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2008). – М.: Наука, 2008. С. 206-209.
7. Апресян Ю.Д., Богуславский И.М., Иомдин Б.Л., Санников А.В., Санников В.З., Сизов В.Г., Цинман Л.Л. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–2005 г. (результаты и перспективы). М: «Индрик», 2005. С. 193-214.

---

[1] Авторы благодарны Российскому фонду фундаментальных исследований (грант № 08-06-00373) и Белорусскому фонду фундаментальных исследований (грант № Ф08Р-016) за частичную финансовую поддержку настоящего исследования. Мы хотели бы также выразить свою признательность С.В. Кодзасову и Л.М. Захарову за предоставленную нам возможность использовать в процессе работы базу данных «Интонация русских информационных текстов».

[2] Тем самым термин «синтагма» используется здесь иначе, чем это принято в литературе, посвященной автоматической обработке устной речи (в том числе и в настоящей статье).

[3] Исключение составляло предложение (16) *Они будут готовы отпустить э… нашего гражданина…*, где заполняющий паузу неструктурный элемент э… был опущен, поскольку мешал бы построению приемлемой синтаксической структуры.

[4] Класс просодического выделения, скажем, противопоставление тематического и рематического интонационных контуров, на данном этапе эксперимента не учитывался.