ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
на тему:

Автоматическое выделение и классификация конструкций на основе синтаксически размеченного корпуса (в задаче снятия неоднозначности)

основная образовательная программа магистратуры по направлению подготовки 45.04.02 «Лингвистика»

Исполнитель:

Обучающийся 2 курса
Образовательной программы
«Прикладная и экспериментальная лингвистика»
очной формы обучения
Бамбурова Ольга Александровна

Научный руководитель:
к.ф.н., доц. Хохлова М.В.

Рецензент:
к.ф.н. Копотев М.В.

Санкт-Петербург
2018

Оглавление

Введение

Благодаря развитию информационных технологий, коммуникация в современном мире вышла на новый уровень развития. Такой стремительный прогресс, в свою очередь, задаёт ещё более высокие стандарты для разрабатываемых средств коммуникации. Как следствие, задачи, связанные с автоматической обработкой данных, становятся все более востребованными. Значительная часть этих данных представлена текстами, написанными на естественном языке, что определяет необходимость предоставить инструменты для автоматической обработки таких текстов. Синтаксический анализ является одним из основных этапов автоматической обработки текста, целью которого является распознавание синтаксической структуры предложения, а также выделение и классификация её отдельных элементов.

Целью данного работы является разработка инструмента для автоматического распознавания и извлечения синтаксической структуры предложения из размеченного корпуса текстов и классификации отдельных элементов структуры.

В соответствии с поставленной целью работы сформулированы следующие **задачи** исследования:

- анализ синтаксически размеченных корпусов текстов, доступных на платформе Natural Language Toolkit[1]
- создание программы для автоматического извлечения правил из синтаксически размеченного корпуса текстов
- анализ извлечённой грамматики и её оптимизация
- тестирование оптимизированной грамматики на контрольной выборке
- оценка работы разработанного инструмента и анализ полученных результатов

[1] https://www.nltk.org/

Объектами исследования являются синтаксическая структура предложения в английском языке, а также набор правил, позволяющих сформировать грамматику, которая покроет как можно больше шаблонов таких синтаксических структур.

Предметом данного исследования являются синтаксические отношения в английском языке и способы их классификации с точки зрения грамматики зависимостей.

Материалом для данного исследования послужил корпус синтаксически размеченных предложений Penn Discourse Treebank (PDTB)[2], предоставленный в открытом доступе на платформе Natural Language Toolkit.

Актуальность данного исследования определена востребованностью инструментов для синтаксического анализа текстов при решении ряда задач в прикладной лингвистике. Так, автоматический синтаксический анализ применяется в системах машинного перевода, автоматической атрибуции, информационно-поисковых системах.

Практическая значимость данного исследования заключается в разработке инструмента, позволяющего ускорить и упростить создание модуля синтаксического анализа теста, который, впоследствии, может быть встроен в любую систему автоматического анализа текста. Наш инструмент может быть использован в проекте, аналогичном Open Corpora[3], который полагается на вклад экспертов и волонтёров в создании разметки для корпуса текстов.

Работа включает в себя введение, теоретическую и практическую главы, заключение, список используемой литературы и приложение. В **теоретической главе** раскрывается специфика существующих подходов к моделированию естественного языка, проводится обзор существующих в исследуемой сфере проектов.

---

[2] https://catalog.ldc.upenn.edu/LDC99T42

[3] http://www.opencorpora.org/

В **практической главе** содержится описание алгоритма работы разработанного нами инструмента, рассказывается о проведённой нами оптимизации, а также приводятся результаты тестирования нашего инструмента на контрольной выборке.

Приложение содержит конечный вариант грамматики после оптимизации.

Глава 1. Синтаксический анализ естественного языка

Введение

Синтаксический анализ является одним из основных этапов автоматической обработки текста, целью которого является распознавание синтаксической структуры предложения, а также выделение и классификация её отдельных элементов. Объектом анализа является цепочка словоформ, образующих предложение (или его часть), для которой устанавливается факт соответствия или несоответствия ряду условий, заданных правилами грамматики конкретного языка. В случае соответствия заданным правилам, анализируемое предложение и его структура могут быть описаны и представлены в форме, принятой в рамках данной грамматики.

Знание синтаксической структуры предложения или отдельных его частей требуется при решении ряда задач в прикладной лингвистике. Так, автоматический синтаксический анализ применяется в **системах машинного перевода** с одного естественного языка на другой. Результаты анализа могут использоваться как для семантической интерпретации, так и непосредственно для преобразования синтаксической структуры переводимого предложения – в структуру выходного предложения переводящего языка. При этом устанавливается набор соответствий между грамматическими средствами переводимого и переводящего языков.

Точное и формальное представление синтаксической структуры предложения делает возможными статистические исследования синтаксиса естественных языков, которые в свою очередь служат базой для разработки статистических моделей языка. Классическим примером могут послужить **системы автоматической атрибуции** (установления авторства) текстов, которые в частности опираются на данные о распределении частот грамматических конструкций, свойственном автору.

Однако стоит отметить, что иногда применяется не полный, а приближенный синтаксический анализ – например, в **информационно-**

поисковых системах, где в основном распознается структура именных, или даже сугубо субстантивных (с существительным в роли главного слова) словосочетаний. Такой подход обусловлен тем, что целью обработки текстов в таких системах является индексирование – распознавание наименований, понятий, или терминологических словосочетаний в тексте документов или запросов.

Широкая сфера применения синтаксического анализа обуславливает существование разных подходов к моделированию структуры предложения в прикладной лингвистике. Каждый подход, в свою очередь оперирует оптимальными методами анализа, ориентированными на то или иное представление синтаксической структуры. В настоящее время широко используются **три формальных модели**: грамматика составляющих (дерево непосредственно составляющих), грамматика зависимостей (дерево зависимостей) и гибридные грамматики, совмещающие свойства двух упомянутых выше (ориентированные структуры составляющих).

1.1. Грамматика непосредственно составляющих

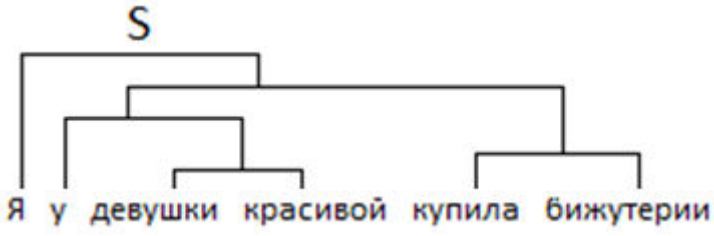
В грамматике непосредственно составляющих предложение *S* рассматривается, как **линейно упорядоченная** цепочка единиц. В качестве таких единиц могут выступать знаки пунктуации и другие символы, словоформы или, в некоторых случаях, единицы, больше, чем одна словоформа (например, сложные союзы). Множество единиц (точек) цепочки складываются во множество отрезков, наибольшим из которых будет являться само предложение S. Чтобы множество отрезков цепочки можно было назвать системой непосредственно составляющих, они должны удовлетворять следующему правилу:

«Любые два отрезка множества R либо не пересекаются, либо один из них содержится в другом» (1)

Однако встречаются случаи, анализ которых труден или даже невозможен в рамках данного правила. В частности, это касается художественных текстов, в которых привычный «правильный» порядок словоформ и отношений между ними нарушен в угоду ритму, как например, фраза из произведения А.С. Пушкина «Евгений Онегин»: *«Он из Германии туманной привёз учёности плоды».*

Ещё чаще нарушение порядка слов случается в разговорной речи в повседневной жизни. На первый взгляд, предложение *«Я у девушки красивой купила бижутерии»* можно представить следующим образом:

а)



В таком случае предложенный анализ соотносится с принципом **проективности**, сформулированным выше (1), однако, слово «красивая» будет определением по отношению к слову «девушка», а не «бижутерия», что может не соответствовать смыслу предложения. Если же мы попытаемся

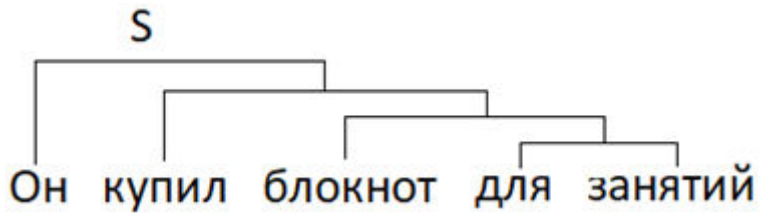
проанализировать второй возможный вариант предложения, то столкнёмся с тем, что полученная структура не является отрезком в линейно упорядоченной цепочке единиц, а значит, противоречит сути модели непосредственно составляющих. Такие структуры называют **разрывными структурами составляющих**:

b)

В противовес неприменимости модели непосредственно составляющих к предложениям с «нетрадиционным» порядком слов, семантически неоднозначные (омонимичные) предложения – не являются проблемными для анализа, так как каждой из возможных интерпретаций будет соответствовать отдельная структура.

Так, например, кажущееся однозначным, на первый взгляд, предложение *«Он купил блокнот для занятий»* может быть разобрано двумя способами, представленными ниже:

(c)



(d)

В первом случае словосочетание «для занятий» будет определением по отношению к слову «блокнот» (блокнот какой? – для занятий), а во втором будет описывать цель покупки (купил блокнот для чего? – для занятий). Способность различать омонимичные предложения на синтаксическом уровне является неоспоримым достоинством грамматики непосредственно составляющих.

Несложно заметить, что представленные выше синтаксические модели предложений (*a, b, c, d*) не несут в себе дополнительной информации о составляющих и их типах, что значительно снижает их ценность с точки зрения лингвистики. Чтобы получить **размеченную** модель, грамматика составляющих должна быть снабжена метками или тэгами, позволяющих различать и классифицировать словоформы и, как следствие, конструкции. Такие грамматики, в свою очередь, соотносят порождаемые предложения с размеченными деревьями составляющих – и именно поэтому называются **порождающими грамматиками**. Помимо разметки такие грамматики содержат правила «разложения» составных синтаксических структур на их элементы, и строго говоря, являются правилами подстановки единиц низшего уровня вместо структур уровнем выше, и при необходимости – наоборот.

На *Рисунке 1* приведён образец синтаксической модели-дерева предложения, сгенерированного размеченной порождающей грамматикой:

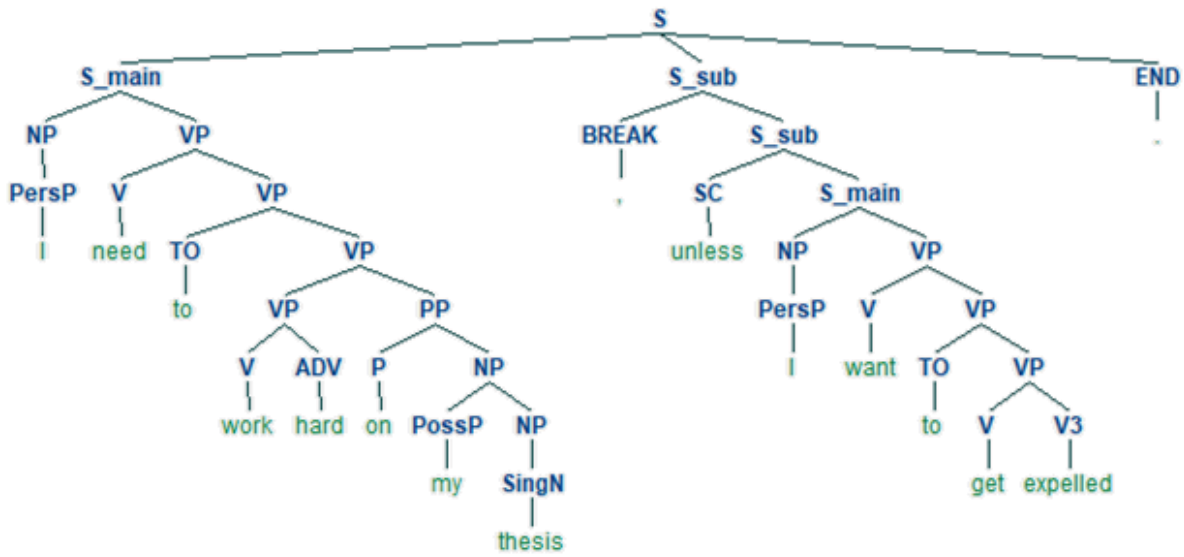


Рис. 1

Можно заметить, что у данного предложения *S* есть две *основные* непосредственно составляющие (далее НС):

- S_main: «*I need to work hard on my thesis*»
- S_sub: «,*unless I want to get expelled*»

Структура S_main, в свою очередь, имеет следующие НС:

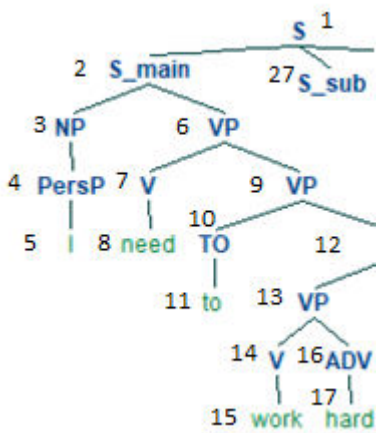
- NP: «*I*»
- VP: «*need to work hard on my thesis*»

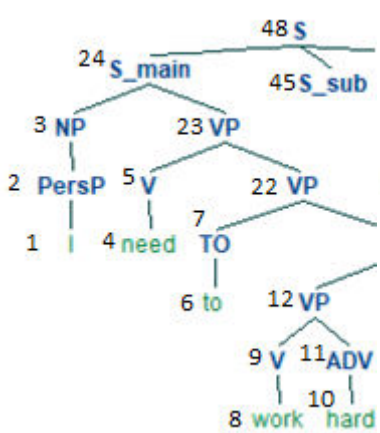
– которые, в последствии тоже раскладываются на составляющие: NP – на терминальную (конечную) единицу, и VP – на нетерминальную.

Таким образом, эти составляющие не входят в предложение *S* непосредственно, а являются компонентами его составляющих. Можно сформулировать следующее правило:

«Ни одна единица цепочки не является НС предложения S, но является НС по отношению к одной из его составляющих» (2)

Подобный подход к анализу синтаксической структуры предложения позволяет задать чёткие **правила преобразования** одних структур в другие – до тех пор, пока они будут не состоять из терминальных единиц. Как итог, анализ структуры предложения по уже заданным правилам, можно свести к поиску всех возможных дериваций (вариантов ветвления дерева предложения), который может осуществляться различными способами. К числу таких способов относятся **нисходящие**, **восходящие** и **комбинированные** алгоритмы.





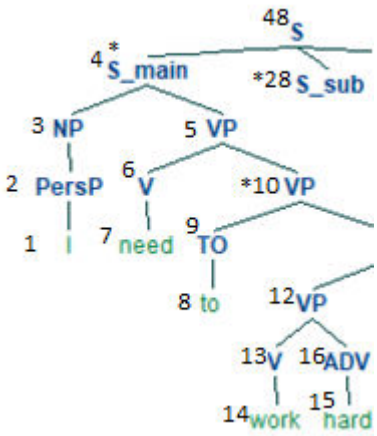


Рис. 2 **Рис. 3** **Рис. 4**

Нисходящие алгоритмы строят синтаксическое дерево сверху вниз, начиная с самой верхней непосредственно составляющей, которой является

само предложение *S*. Порядок построения изображён на *Рисунке 2*. По сути, данные алгоритмы поочерёдно воспроизводят заданные правила, но применяют их сначала к первым компонентам в имеющейся последовательности. В результате, такие алгоритмы сразу выстраивают предполагаемую структуру предложения, однако для того, чтобы привести её в соответствие с входными данными, им необходимо осуществить перебор всех дериваций, порождаемых грамматикой. Процедура повторяется для каждой «ветви» до тех пор, пока **начальные** компоненты порождённого и фактического высказываний не совпадут, что ведёт худшей производительности и большей ресурсозатратности.

Восходящие алгоритмы строят синтаксическое дерево снизу вверх. В процессе нижестоящие составляющие заменяются вышестоящими до того момента, пока для замены не потребуется соединить несколько «ветвей» в одну – в этот момент движение вверх прекращается до тех пор, пока все позиции уровнем ниже не будут заполнены. Порядок построения изображён на *Рисунке 3*. Специфика работы таких алгоритмов обуславливает их высокую производительность, так как не ведёт к порождению чрезмерного количества дериваций, что характерно для нисходящих алгоритмов. Однако восходящие алгоритмы неприменимы для работы с рекурсивными грамматиками, допускающими пропуски составляющих (эллипсис), что ведёт либо к невозможности анализа предложений с эллипсисом, либо к необходимости внедрять гораздо больше правил, чем потребовалось бы с рекурсивной грамматикой.

Комбинированные (гибридные) алгоритмы, как можно понять из названия, совмещают в себе вышеописанные стратегии. Порядок построения изображён на *Рисунке 4*. Как следствие, такие алгоритмы обладают более высокой производительностью, чем нисходящие, и вместе с тем применимы к рекурсивным грамматикам с эллиптическими составляющими. Их преимуществом является то, что в отличие от восходящих алгоритмов движение вверх прекращается не в тот момент, когда требуется объединение

нескольких «ветвей», а на шаг позже, что позволяет сравнить порождаемое дерево с фактическим и отбросить порождение тех деривации уровнем ниже, что точно не впишутся в уже сформированную часть.

1.2. Грамматика зависимостей

В грамматике зависимостей предложение *S* также рассматривается, как линейно упорядоченная цепочка единиц, на которой, в отличие от грамматики непосредственно составляющих, задаются **бинарные отношения зависимости** между единицами. Как следует из названия, бинарное отношение отражает фактическую зависимость между двумя словоформами, которые образуют словосочетание в предложении. Такой подход можно рассматривать как обобщение традиционно выделяемых в предложении отношений – управления, согласования и примыкания – в которых участники обычно «неравноправны» как в грамматическом, так и в смысловом плане. Набор таких зависимостей между словоформами предложения может быть представлен в виде дерева зависимостей.

Деревом зависимостей предложения *S* называется **конечный** граф на множестве словоформ (узлов) предложения *S* такой, что:

1) существует единственная словоформа, не зависящая ни от какой другой (эта словоформа называется вершиной дерева *V*),

2) всякая словоформа, отличная от стоящей в вершине (*V*), зависит ровно от одной другой словоформы,

3) в графе отсутствуют замкнутые пути (в котором начало и конец совпадают). **Путь** в дереве зависимостей – это последовательность словоформ, в которой каждая следующая зависит от предыдущей.

Учитывая правила, перечисленные выше, можно заключить, что *до каждого узла дерева ведёт один единственный путь от вершины дерева*. Стоит отметить, что помимо прямой зависимости между словоформами (узлами), расположенными одна за другой на пути от вершины – существует **косвенная зависимость**. На *Рисунке 5* приведён пример: если между узлом N_1 и узлом N_3 данного пути лежит ещё один узел N_2, то N_1 косвенно подчиняет N_3 (или наоборот: словоформа N_3 косвенно зависит от словоформы N_1).

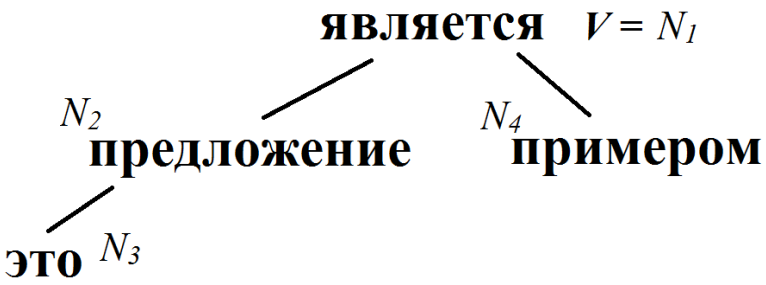


Рис.5

Множество всех словоформ, косвенно и прямо зависящих от словоформы N_1, называется **группой зависимости** этой словоформы. Так на *Рисунке 5* явственно видно, что в группу зависимости словоформы N_1 «является» входят словоформы «это», «предложение» и «примером», а в группу зависимости словоформы N_2 «предложение» – только словоформа «это»; остальные словоформы не имеют группы зависимости в конкретном случае. Количество и состав возможных зависимых слов определяется способностью главного слова вступать с ними во взаимодействие, или **валентностью**. Например, валентность глагола определяет, может ли он иметь подлежащее, прямое дополнение, косвенное дополнение.

Нельзя не обратить внимания на то, что, как и грамматика составляющих, грамматика зависимостей основана на правиле **проективности** (или ограничена им). *Проективной структурой* называется такая структура, в которой группа зависимости каждой словоформы является неразрывным отрезком в линейном порядке слов предложения. Формальным выражением нарушения свойства проективности в дереве зависимостей будет пересечение проекций связей между словоформами, как изображено на *Рисунке 6*:

Рис.6

При этом появляются так называемые «**разрывные**» группы зависимости отдельных словоформ (т. е. группы, не являющиеся отрезками в линейном порядке слов предложения).

Как и в грамматике составляющих, в грамматике зависимостей различают **размеченные** и неразмеченные модели. Модели, представленные на *Рисунках 5 и 6*, не являются размеченными, так как не содержат информации о типах зависимостей в предложении. Размеченный вариант модели может выглядеть следующим образом:

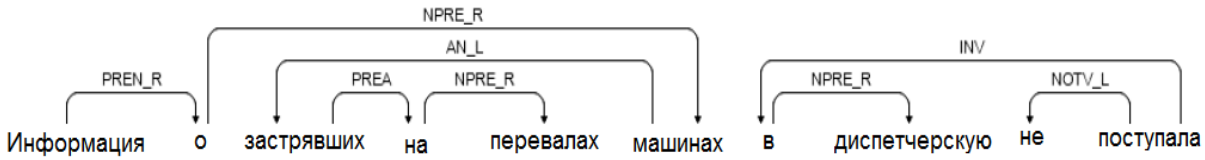


Рис.7

На *Рисунке 7* обозначены следующие виды зависимостей (L – левая связь, R – правая связь):

- PREN_R – от именной группы к предложной
- NPRE_R – от предлога к существительному
- AN_L – от существительного у адъективному модификатору
- PREA – от отглагольного адъективного модификатора к предлогу
- INV – от глагола к предлогу при инверсированном порядке слов
- NOTV_L – связь от глагола к отрицанию

Помимо дополнительной информации, размеченные деревья зависимостей предоставляют возможность различать омонимичные предложения (или конструкции в их составе) в том случае, когда структуры графов (или отдельные пути в их составе) совпадают. *Рисунок 8* иллюстрирует подобную ситуацию в предложении «*Преследование тигра затянулось*»:

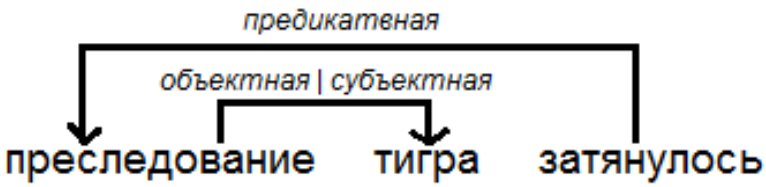


Рис.8

Два возможных типа связи, указанные для словоформ «преследование» и «тигра», соотносятся с двумя возможными толкованиями словосочетания: наличие **субъектной** связи будет означать, что «тигр преследует кого-то», а наличие **объектной** связи – что «кто-то преследует тигра». Мы видим, что разметка будет зависеть от трактовки предложения, при этом само дерево останется неизменным. Набор типов грамматических отношений и, как следствие, подробность и глубина разметки – определяются конкретной грамматикой языка и задачами, для решения которых создаётся модель.

Тот факт, что грамматика зависимостей берёт за основу связи между словоформами, определяет особенности алгоритмов построения деревьев зависимостей: обычно они основаны на **правилах продукций** – условных переходах вида «если…, то …». Такое правило может выглядеть следующим образом:

$$«Vt\ S4 \Longrightarrow Vt \rightarrow [Acc]\ S4»$$

– что можно прочитать как: если за *Vt* (переходным глаголом) следует *S4* (имя существительное в винительном падеже), то между ними устанавливается подчинительная связь $\rightarrow$ *[Acc]*. Совокупность правил продукций представляет собой продукционную модель, которая теоретически «производит» все возможные корректные выводы о заданном наборе объектов. При этом правила обычно нумеруются, что позволяет исключить работу одних правил, если ещё не сработали другие. Тем не менее, если правила в какой-то момент всё же начинают противоречить друг другу, то проблему можно решить с помощью дополнительных управляющих структур, искусственно разрешающих возникающие противоречия.

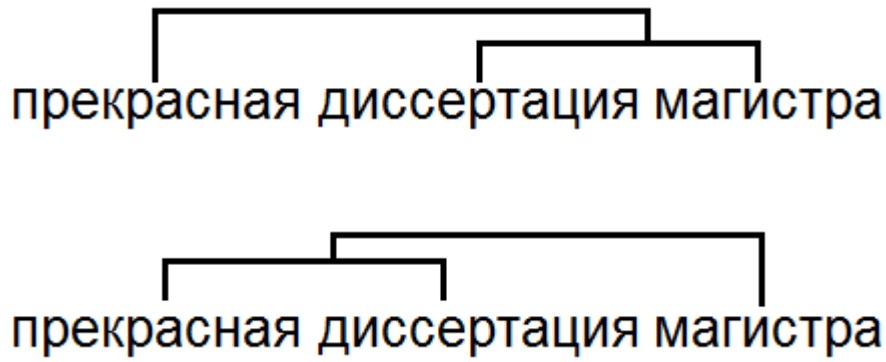
1.3. Сравнительный анализ грамматики составляющих и грамматики зависимостей. Гибридные грамматики

Грамматика составляющих (ГС) и грамматика зависимостей (ГЗ) берут за основу принципиально различную информацию о структуре предложения и его компонентов. Основные различия этих двух представлений синтаксической структуры предложения состоят в следующем:

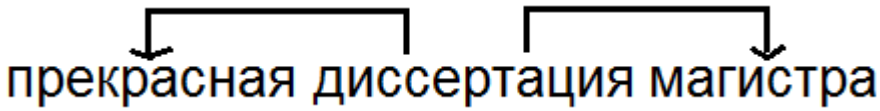
1) Элементарными единицами в ГЗ являются вхождения словоформ, а в ГС – словосочетания (включая в качестве частного случая отдельные словоформы и полное предложение).

2) ГЗ базируется на «неравноправном» отношении между двумя словоформами N_1 и N_2 – когда одна из них зависит от другой. Как следствие, отношение синтаксической связи является ориентированным: $N_1 \rightarrow N_2$ или $N_2 \rightarrow N_1$. В ГС отношения непосредственно составляющих некоторой составляющей не иерархичны, т. е. среди них не выделяется ни одна непосредственно составляющая в качестве главной (основной), от которой зависели бы другие непосредственно составляющие.

3) ГС допускает возможность объединять в одну составляющую «семантически близкие» словосочетания, когда ГЗ не может различить подобные смысловые оттенки. Примером здесь может служить анализ словосочетания «*прекрасная диссертация магистра*», для которого возможны две структуры непосредственно составляющих:

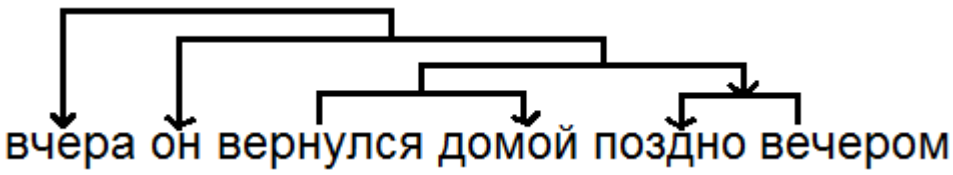


– и лишь одна структура зависимостей:



Образно говоря, ГЗ «воспринимает» предложение как мозаику, а ГС – как конструктор. По аналогии с **мозаикой** – первый кусочек будет являться вершиной предложения, а последующие элементы будут подбираться в соответствии с формой (валентностью) и рисунком (морфологическими признаками) этого кусочка (по правилам продукции «если…, то…»). Получается, что в ГЗ *важную роль играет контекст* – характеристики смежных единиц, которые позволяют определить, какие правила продукции необходимо применить в конкретном случае. ГС, в свою очередь, *лишена привязки к контексту*, так как не проводит принципиальных различий между «блоками», из которых будет собрано предложение. Целью в таком случае будет являться не соотнесение «блоков» друг с другом, а лишь их расположение в необходимом порядке в соответствии с их «размером» – количеством элементарных единиц в составе.

Несмотря на такие кардинальные различия в «восприятии» предложения и его элементов, существует возможность скомбинировать эти две грамматики. Например, можно ввести в ГС дополнительную информацию об ориентации синтаксических связей, а именно – в каждой составляющей выделить главную непосредственно составляющую. Такая гибридная грамматика называется **ориентированной грамматикой составляющих** (ОГС). Дерево, построенное в соответствии с ОГС, может выглядеть следующим образом (структуры, к которым ведут стрелки, технически не являются зависимыми – но «неглавными»):



В некоторых прикладных задачах комбинированный подход реализуется с помощью **расширенных сетей переходов** и основанных на них РСП-грамматиках. В частности, именно на такой грамматике основан синтаксический анализ предложений в системе ПРОМТ. РСП-грамматики – это комбинация ГС и Марковских цепей (сетей переходов), в которых переходы производятся не только от словоформы к словоформе, но и между словоформами и непосредственно составляющими (НС) или непосредственно между НС. При этом для каждого «блока» НС строится своя независимая сеть, позволяющая анализировать их независимо друг от друга. Фрагмент варианта такой сети для предложения «*Старушка вынула из рабочего ящика нательный золотой крестик Наташи*» представлен на рисунке ниже:

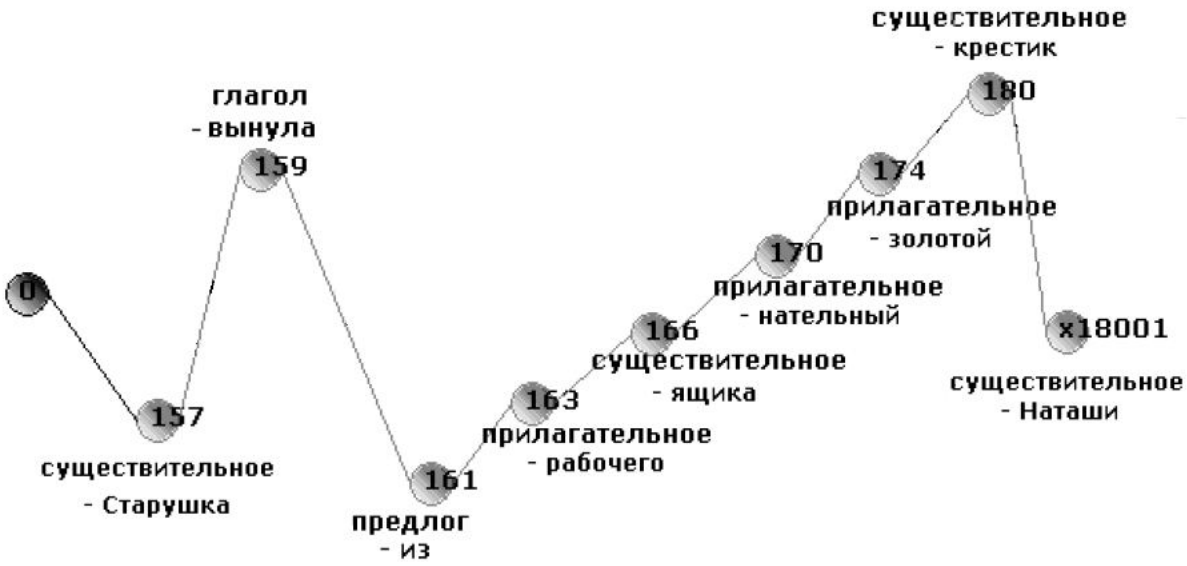


Рис.9

Можно заметить, что представленный фрагмент сети также является графом, а каждая словоформа – узлом, что согласуется с методами ГЗ. Однако принцип, по которому меняется направление пути от начальной до конечной единицы, указывает на разделение структуры предложения ещё и на именную и глагольную группы (и при необходимости более мелкие конструкции) – что свидетельствует о влиянии теории ГС.

В нашей работе мы остановили выбор на корпусах текстов, размеченных в соответствии с принципами грамматики непосредственно составляющих,

так как она является оптимальной по ряду критериев. Во-первых, работа с ГС, в отличие от ГЗ, не требует досконального знания грамматической сочетаемости словоформ, на которых основаны правила продукции. Также, нет необходимости различать типы зависимостей меду словоформами, что при анализе на основе ГЗ вызывает ряд проблем. Среди таких проблем – адекватное представление сочинительной связи, как в случае с предложением «*Лена, Оля и Лиза успешно защитили диссертации*». Если союз «и» является зависимым, то управлять им будет «Лена», «Оля» или «Лиза», однако, словоформы используются в единственном числе, что противоречит множественному числу глагольной формы «защитили» в предложении. Если же главным является союз, то словосочетание «Лена, Оля и Лиза» не будет обладать свойствами существительного, а значит, его будет невозможно встроить в правила продукции.

Придаточная связь также является проблемой в системе ГЗ, что явственно видно в предложении «***Магистрант**, диссертацией **которого** гордился весь ВУЗ*». Так, словоформа «которого» может быть подчинена словоформе «диссертацией» по падежному признаку (родительный падеж), либо словоформе «магистрант» – по родовому признаку (мужской род). И в том и в другом случае, одно из свойств зависимого слова будет противоречить свойствам главного слова. Попытки объяснить ситуацию через анафорическую связь не кажутся убедительными, так как связь анафора и антецедента не всегда «грамматична». В предложении «*Декан сказал **студенчеству**, что **они** должны идти в магистратуру*» анафорическая связь выделенных словоформ очевидна, однако, они согласованы ни в роде, ни в числе, ни в падеже. Следовательно, связь между словом «который» и определяемым существительным является не только анафорической, но и грамматической, но если это так, то нарушается постулат ГЗ о том, что каждое слово может быть грамматически зависимым не более чем от одного слова.

Список проблем, которые невозможно решить средствами ГЗ, можно продолжить, и он свидетельствует о недостаточности данной грамматики при работе с естественным языком. Что касается корпусов, размеченных в рамках гибридной грамматики, то таких, что удовлетворяли бы задачам и средствам исследования, нет в открытом доступе. В связи с перечисленными выше причинами, наше исследование сосредоточено на корпусах текстов, размеченных согласно грамматике непосредственно составляющих.

1.4. Корпусы синтаксически размеченных текстов

Исследования в сфере автоматической обработки языка свидетельствуют о том, что значительного и быстрого прогресса в извлечении данных, как из текста, так и разговорной речи, можно достичь, используя крупные корпусы текстов, материал для которых подбирался по какому либо признаку или ряду признаков, как например жанр, автор, период создания стиль и т.п. Такие корпуса служат важным инструментом для исследователей в области обработки естественного языка, распознавания речи и интегрированных систем разговорного языка, а также в теоретической лингвистике. Аннотированные корпуса представляют ценность и для таких проектов, как автоматическое построение статистических моделей для грамматики письменного и разговорного языка, разработка явных формальных теорий разных грамматик письменности и речи, исследование просодических явлений в речи и оценки и сопоставления адекватности моделей синтаксического анализа.

В рамках нашей работы мы провели анализ корпусов текстов, размеченных в соответствии с грамматикой непосредственно составляющих и доступных на платформе Natural Language Toolkit. Помимо определённых выше критериев, в работе не рассматривались корпусы текстов на языке, которым автор исследования не владеет на уровне, достаточном для анализа синтаксических связей в предложении. Анализ выявил, что наиболее репрезентативным корпусом текстов, соответствующим критериям, указанным выше, является **Penn Discourse Treebank** (PDTB).

Penn Discourse Treebank (PDTB) является проектом, реализованным в Университете Пенсильвании. Цель проекта состояла в том, чтобы аннотировать Wall Street Journal corpus, состоящий из более миллиона слов. В течение первых трёх лет проекта (1989-1992 годы) корпус был аннотирован информацией о частях речи (POS) [4] . Следуя лексически

[4] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

обоснованному подходу к аннотации, PDTB маркирует отношения, реализуемые явным образом, взятыми из синтаксически четко определенных классов, а также отношения между смежными предложениями, когда никакой видимой связи не наблюдается. Аргументы отношений аннотируются в каждом случае.

Для явных связок аргументы не ограничены в терминах их расстояния от соединительной линии и могут быть найдены в любом месте текста. Между соседними предложениями, в которых отсутствует явная связка, используются четыре сценария[5]:

1) предложения могут быть связаны отношением дискурса, которое не реализует во втором предложении, и в этом случае создается соединительная (называемая неявной связью) для выражения;

2) предложения могут быть связаны отношением дискурса, которое реализуется посредством некоторого альтернативного несвязанного выражения, и в этом случае эти альтернативные лексикализации аннотируются как носители отношения (помеченные как AltLex);

3) предложения могут быть связаны не отношением дискурса, а просто отношением связности на основе сущности, и в этом случае наличие такого отношения помечено (как EntRel);

4) предложения не могут быть связаны вообще, и в этом случае они обозначаются как таковые (NoRel).

В дополнение к структуре аргументов отношений PDTB обеспечивает:

- смысловые аннотации для каждого отношения дискурса, а также фиксирует многозначность связок;

- аннотации атрибутов отношений и каждый их аргумент, причем каждый экземпляр атрибуции дает соответствующий интервал текста и четыре функции для захвата семантического вклада атрибуции.

[5] https://catalog.ldc.upenn.edu/LDC2008T05

Для смысловых, неявных и AltLex-отношений предусмотрены как смысловые, так и атрибутивные аннотации, но не для EntRel и NoRel.

Лексически обоснованный подход в PDTB предоставляет четко определенный уровень структуры дискурса, который будет поддерживать извлечение ряда выводов, связанных с дискурсивными связями.

На сегодняшний день группа PDTB провела различные эксперименты на корпусе, в частности, рассмотрев следующие вопросы:

- выравнивание между синтаксисом и дискурсом, особенно в отношении атрибуции;

- смысловое смещение дискурсных связей;

- сложность зависимостей в дискурсе.

Группа проекта PDTB будет продолжать изучать эти проблемы и сосредоточиться на более масштабных проектах, таких как разбор речи, автоматическое обобщение и генерация естественного языка. Дальнейшая работа также исследует фундаментальные проблемы в дискурсе.

Выводы

В теоретической главе была раскрыта специфика существующих подходов к моделированию естественного языка. Также, был проведён обзор существующих в исследуемой сфере проектов. Было установлено, что синтаксический анализ является одним из основных этапов автоматической обработки текста, целью которого является распознавание синтаксической структуры предложения, а также выделение и классификация её отдельных элементов.

Глава 2. Разработка инструмента для автоматического извлечения правил из корпуса Pennsylvania Treebank

Введение

Для начала работы над инструментом для синтаксического анализа, мы должны обладать набором определенных данных о тех единицах, которые собираемся анализировать: границы предложений, границы словоформ, их возможные морфологические интерпретации. Платформа NLTK (**Natural Language Toolkit**) была выбрана как наиболее гибкая, содержащая основные алгоритмы для автоматической обработки текста, не требующая существенной перестройки для работы с русским языком, обладающая хорошей совместимостью и широким кругом инструментов для исследования и работы с естественным языком.

Так же внутри среды существует набор готовых библиотек для решения задач классификации, токенизации, кластеризации, машинного обучения. Инструменты NLTK позволяют осуществлять полный цикл автоматической обработки текста: от графематического анализа и токенизации до синтаксического анализа и логической семантики. В NLTK включены не только готовые инструменты для анализа текстов, но и корпусы и алгоритмы, позволяющие решать задачи автоматической обработки текста на основе машинного обучения.

NLTK позволяет проводить синтаксический анализ на основе формальных грамматик. Исследователям предоставляется возможность самим создавать такие грамматики и, пользуясь встроенными инструментами, получать синтаксический разбор. Среди ресурсов, предоставляемых NLTK, присутствуют не только инструменты для обработки текста, но и лингвистические данные в виде корпусов, которые можно использовать, в том числе, для машинного обучения.

Как уже было описано выше, с помощью конечного набора правил (грамматики) можно описать многочисленные и потенциально бесконечные

конструкции некоторого естественного языка. Теория генеративных грамматик рассматривает язык как очень большой набор всех грамматически правильных предложений, а грамматику как формальное описание того, каким образом эти предложения могут быть построены. Грамматики используют рекурсивные правила (или продукции). Модуль синтаксического анализа в NLTK позволяет работать с несколькими типами грамматик: контекствно-свободными (CFG), **вероятностно контекстно-свободными** (PCFG), лексикализованными и контекстно-зависимыми. Вероятностно контекстно-свободной грамматикой называют такую грамматику непостредственно составляющих, где каждому терминальному и нетерминальному элементу привисан «вес», или вероятность. Как было описано выше, мы решили работать именно с такой грамматикой.

1.1. Алгоритм извлечения набора правил из разметки синтаксического корпуса

В ходе кропотливой работы нами был разработан алгоритм, который позволяет извлечь синтаксическую разметку из корпуса Penn Discourse Treebank. Ниже представлен код с пошаговыми комментариями и размеченный построчно:

«import nltk; *# загрузка библиотеки Natural Language Toolkit*
trans = str.maketrans("",#$:`'",''1234567'',''1234567890-=|''); *# определяем, какие символы необходимо заменить, так как среда разработки воспринимает их как служебные*
Minimum_Repetitions_Required = 1; *# задаём минимальный порог, определяющий количество вхождений правила в корпус, необходимый для того, чтобы правило попало в файл вывода (грамматику)*

def AddTreeToDictionary(Tree, Dictionary): *# определяем функцию пополнения словаря, которая принимает 2 аргумента «Tree» и «Dictionary»*
TreeLabel = Tree.label().translate(trans); *# вызываем функцию «trans» для названий объектов «Tree»*
if TreeLabel not in Dictionary: *# определяем условие «если название дерева ещё не в словаре…*
Dictionary[TreeLabel] = {}; *# …то добавляем его в качестве ключа» и задаём значение по умолчанию = пусто*

TreeBranches=(); *# создаём пустой кортеж…*
for TreeBranch in Tree: *# …для дочерних объектов объекта «Tree»*
if type(TreeBranch) is nltk.tree.Tree: *# проверяем тип данных дочерних объектов («TreeBranch»). Если это nltk.tree.Tree то...*

TreeBranches = TreeBranches + (TreeBranch.label().translate(trans),); # *...добавляем название в кортеж, при этом заменяя служебные знаки (функция trans)*

AddTreeToDictionary(TreeBranch, Dictionary); # *функция рекурсивно вызывает себя, чтобы проверить каждый объект, спускаясь вниз по структуре дерева, до тех пор, пока TreeBranch = nltk.tree.Tree*

else: # *в момент, когда объект имеет любой тип данных, кроме nltk.tree.Tree (а мы знаем, что это string), то...*

TreeBranches = TreeBranch; # *...мы закрываем кортеж для данного объекта*

if TreeBranches not in Dictionary[TreeLabel]: # *если полученного кортежа нет во вложенном словаре...*

Dictionary[TreeLabel][TreeBranches] = 1; # *...то добавляем его*

else: # *иначе...*

Dictionary[TreeLabel][TreeBranches] += 1; # *...прибавляем к вхождениям существующего ключа 1*

def Rare(Tree, Dictionary): # *определяем функцию «Rare» для объектов «Tree», которая отсеивает правила, число вхождений которых меньше, чем порог, заданный в параметром «Minimum_Repetitions_Required»*

TreeBranches=(); # *создаём пустой кортеж…*

for TreeBranch in Tree: # *...для дочерних объектов объекта «Tree»*

if type(TreeBranch) is nltk.tree.Tree: # *проверяем тип данных дочерних объектов («TreeBranch»). Если это nltk.tree.Tree то...*

TreeBranches = TreeBranches + (TreeBranch.label().translate(trans),); # *...добавляем название в кортеж, при этом заменяя служебные знаки (функция trans)*

if Rare(TreeBranch, Dictionary): # *проверяем, входит ли объект «TreeBranch» в существующий словарь функции «Rare»*

return True; *# да, закончить функцию для объекта*

else: *# иначе…*

return False; *# нет, добавить объект в словарь*

if Dictionary[Tree.label().translate(trans)][TreeBranches] < Minimum_Repetitions_Required: *# сверяем объекты в словаре с минимальным порогом вхождений, заданным в Строке 3*

return True; *# да, ниже заданного порога – не записывать в конечный файл грамматики*

else: *# иначе…*

return False; *# нет, выше заданного порога – записать в конечный фай грамматики*

def main(): *# задаём главную функцию «main»*

Dictionary = {}; *# задаём наполнение словаря по умолчанию = пусто*

Parsed_Sents = list(nltk.corpus.treebank.parsed_sents()); *# задаём переменную «Parsed_Sents», которая является списком всех предложений в корпусе*

for Parsed_Sent in Parsed_Sents: *# для каждого отдельного предложения в списке…*

AddTreeToDictionary(Parsed_Sent, Dictionary); *# … вызываем функцию «AddTreeToDictionary»*

if Minimum_Repetitions_Required > 1: *# если минимальный порог вхождений > 1, то начинаем проверку предложений на соответствие условию (если = 1, то шаг пропускается)*

s = 0; *# переменная для подсчёта всех предложений в списке*

b = 0; *# переменная для подсчёта предложений, не подходящих под условие*

for Parsed_Sent in list(Parsed_Sents): *# с каждым предложением в списке…*

s+=1; *# переменная «*s*» увеличивается на 1*

if Rare(Parsed_Sent,Dictionary): *# вызываем функцию «Rare» для проверки словаря на редкие предложения*

b+=1; *# переменная «*b*» увеличивается на 1*

Parsed_Sents.remove(Parsed_Sent); *# и редкое предложение удаляется из словаря*

print("before: %i" % s); *# вывести на печать*

print(" after: %i" % (s-b)); *# вывести на печать*

Dictionary = {}; *# снова задаём наполнение словаря по умолчанию = пусто…*

for Parsed_Sent in Parsed_Sents: *# чтобы собрать словарь заново…*

AddTreeToDictionary(Parsed_Sent, Dictionary); *# без учёта правил, что входили в редкие исключённые предложения*

 # блок подсчёта вероятностей

```
for DictionaryKey in Dictionary.keys():
    Sum = 0;
    for SubDictionaryKey in Dictionary[DictionaryKey].keys():
        Sum += Dictionary[DictionaryKey][SubDictionaryKey];
    for SubDictionaryKey in Dictionary[DictionaryKey].keys():
        Dictionary[DictionaryKey][SubDictionaryKey] = float(Dictionary[DictionaryKey][SubDictionaryKey]) / Sum;
```

 # блок создания «сырого» файла вывода (без служебных символов, необходимых для работы грамматики)

```
f = open('PythonRawOutput.txt', 'w')
f.write(str(Dictionary));
f.close();
```

 # блок создания конечного файла вывода (со служебными символами, необходимыми для работы грамматики)

```
   f = open('PythonOutput.txt', 'w')
   s = '';
   for DictionaryKey in Dictionary.keys():
      f.write(s);
      s = DictionaryKey + ' ->';
      for SubDictionaryKey in Dictionary[DictionaryKey].keys():
         if type(SubDictionaryKey) is str:
            s += ' ' + repr(SubDictionaryKey);
            #s += ' [' + repr(Dictionary[DictionaryKey][SubDictionaryKey]) +']
|';
            #s += '' [%.4f] |'' % Dictionary[DictionaryKey][SubDictionaryKey];
         else:
            for Element in SubDictionaryKey:
               s += ' ' + Element;
         s += '' [%.5f] |'' % Dictionary[DictionaryKey][SubDictionaryKey];
      s = s[:-2] + '\n';
   s = s[:-1];
   f.write(s);
   f.close();
```

main() *# вызываем основную функцию»*

После завершения работы программы, мы получаем файл вывода с грамматикой. Именно для неё и будет проводиться дальнейшая оптимизация.

1.2. Анализ и оптимизация извлечённой грамматики

Анализ полученной грамматики показал, что она является небинарной и хаотично рекурсивной, что не является оптимальным, так как это ресурсозатратно, и делает грамматику неуниверсальной (правила буквально копируют структуру предложения, без разделения на дополнительные вложенные структуры).

В результате оптимизации удалось добиться сокращения количества правил за счёт упорядочивания рекурсии, бинаризации и удаления дублирующихся правил.

Примером оптимизации может являться следующее правило:

Состояние «до»: CONJP -> ADJP CC ADJP [0.00726] | RB RB JJ [0.00435] | 66 RB 77 JJR [0.00145] | QP JJR [0.00145] | JJS JJ [0.00435] | VBN PP [0.00290] | RB RB JJ PP [0.00145] | JJ RB SBAR [0.00145] | CD NN [0.08273] | RB JJR IN [0.00145] | ADVP JJ [0.00581] | RB 66 JJ CC JJ 77 [0.00145] | RB [0.00435] | ADJP 2 CC ADJP [0.00145] | JJ S [0.00726] | ADJP PP [0.01451] | JJR [0.01451] | RBR JJ [0.03193] | NP JJ [0.01451] | JJ SBAR [0.00290] | NN NN [0.00145] | CD JJ [0.00145] | JJ 2 JJ 2 JJ CC NN [0.00145] | VB JJR [0.00145] | CD NNS [0.00145] | NONE [0.00290] | NN CC NN [0.00145] | RB VBG [0.00290] | NPADV JJR [0.00435] | JJ NP [0.00145] | JJ 2 JJ CC JJ [0.00435] | JJ CC RB [0.00145] | DT [0.00145] | JJ VBN [0.00145] | JJ CC JJ [0.02612] | NPADV JJ [0.00145] | NNS PRN VBN [0.00145] | RB RB [0.00581] | NN JJ [0.00145] | JJS JJ S [0.00145] | JJ JJR [0.00145] | JJ RB [0.00145] | RBR [0.00145] | NNP 2 JJ [0.01161] | JJR PP [0.00145] | IN NN [0.00145] | JJR CC JJR [0.00290] | ADJP 2 ADJP 2 ADJP [0.00145] | RB JJ PPLOC [0.00145] | NNP NNP [0.00435] | JJ CC VBG [0.00145] | NNP JJ [0.00871] | JJ JJS [0.00145] | ADVPTMP VBN [0.00145] | JJ PRN [0.00145] | JJ 2 CC JJ [0.00145] | RB JJ CC JJ [0.00290] | ADJP 2 CC ADJP 2 [0.00145] | JJ PP [0.06241] | JJ NPTMP [0.03048] | 66 RB VBN [0.00145] | NNS CC NNS [0.00145] | JJR VBN [0.00145] | JJR JJ [0.00581] | NNP 2 NNP JJ [0.00145] | JJ [0.10160] | QP NN [0.01161] | RB DT [0.00290] | VBN CC JJ [0.00145] | NN RB SBAR [0.00145] | VBN

[0.00581] | NNP NNP 2 JJ [0.00145] | JJ PPTMP [0.00290] | NN [0.00290] | RB VBN PRT PP [0.00145] | ADVPTMP JJ [0.00290] | DT ADJP CC ADJP [0.00145] | 4 CD JJ [0.00145] | RB RB S [0.00145] | JJ JJ [0.01161] | QP NONE [0.08418] | 4 CD NONE [0.04064] | RBS VBN [0.00145] | RB JJ [0.11176] | RB JJ PP [0.01016] | VBN JJ [0.00145] | VBN PPCLR [0.00145] | 4 JJ NONE [0.01016] | RB RBR JJ [0.00290] | JJ CD NN [0.00145] | NP JJ PP [0.00145] | JJ PPLOC [0.00290] | RB JJR [0.03193] | ADVP VBN [0.00290] | RBS RB JJ [0.00145] | RB RB PP S [0.00145] | VBG CC VBG [0.00145] | JJ CC NNP [0.00145] | CD CD NN [0.01161] | 4 JJ [0.00145] | ADVP JJ PP [0.00145] | RBR JJ PP [0.00145] | ADVPTMP RB JJ [0.00145] | ADVP RB JJR [0.00145] | QP RB JJ [0.00145] | RB VBN [0.06096] | RBS JJ [0.02612] | 66 JJ 77 CC 66 JJ [0.00145]

Состояние «после»: CONJP -> QP JJR [0.00145] | JJS JJ [0.00435] | VBN PP [0.00290] | CD NN [0.08273] | ADVP JJ [0.00581] | JJ S [0.00726] | ADJP PP [0.01451] | RBR JJ [0.03193] | NP JJ [0.01451] | JJ SBAR [0.00290] | NN NN [0.00145] | CD JJ [0.00145] | VB JJR [0.00145] | CD NNS [0.00145] | RB VBG [0.00290] | NPADV JJR [0.00435] | JJ NP [0.00145] | JJ VBN [0.00145] | NPADV JJ [0.00145] | RB RB [0.00581] | NN JJ [0.00145] | JJ JJR [0.00145] | JJ RB [0.00145] | NNP 2 JJ [0.01161] | JJR PP [0.00145] | IN NN [0.00145] | NNP NNP [0.00435] | NNP JJ [0.00871] | JJ JJS [0.00145] | ADVPTMP VBN [0.00145] | JJ PRN [0.00145] | JJ PP [0.06241] | JJ NPTMP [0.03048] | 66 RB VBN [0.00145] | JJR VBN [0.00145] | JJR JJ [0.00581] | QP NN [0.01161] | RB DT [0.00290] | JJ PPTMP [0.00290] | ADVPTMP JJ [0.00290] | 4 CD JJ [0.00145] | JJ JJ [0.01161] | RBS VBN [0.00145] | RB JJ [0.11176] | VBN JJ [0.00145] | VBN PPCLR [0.00145] | JJ PPLOC [0.00290] | RB JJR [0.03193] | ADVP VBN [0.00290] | 4 JJ [0.00145] | RB VBN [0.06096] | RBS JJ [0.02612]

Подробный лог введённых изменений приведён ниже:

1) (=) -> (1-0)
2) (PRP$) -> (PRP1)

3) (:) -> (1-1) *** and (: ->)

4) (-NONE-) -> (NONE) *** and -NONE- ->

5) (,) -> (1-2) *** and (, ->)

6) (`` ) -> ( 1-3 ) *** and (`` ->)

7) ('') -> (1-4) *** 2 single brackets *** and ('' ->) *** and 1 more

8) (-LRB-) -> (LRB) *** and (-LRB- ->)

9) (-RRB-) -> (RRB) *** and -RRB- ->

10) (.) -> (1-5) *** and (. ->)

11) (WP$) -> (WP1) *** and (WP$ ->)

12) 6.891798759476223e-05 -> 0.00006891798759476223

13) 4.2337002540220155e-05 -> 0.000042337002540220155

14) 8.467400508044031e-05 -> 0.00008467400508044031

15) 7.595321282090232e-05 -> 0.00007595321282090232

16) ($) -> (1-6) *** and $ ->

17) (#) -> (1-7) *** and # ->

18) (ADVP|PRT) -> (ADVP-PRT) *** and (ADVP|PRT ->)

19) [0.\[0-9]+\] - decided to see if works without probabilities

20) ***[0.[0-9]+\] and \[0.[0-9]+\] work equally?

21) '([^']+)' - for all strings

22) String data corrections:

23) copied from P3:

24) JJ -> 'and', 'a' - deleted

25) IN -> 'a' - deleted

26) NNP -> 'British'

27) Implemented by mistake -> discarded:

28) PP-TMP-2 -> IN NP [1.0] - duplicate string *** deleted the latter

29) NP-SBJ-3 -> DT NNP CC NNP NNS [1.0] - partial duplicate (duplicates only this value) - deleted

30) NP-SBJ-2 -> DT JJ NN [1.0] - partial duplicate (duplicates only this value) - deleted

31) PP-TMP-3 -> IN NP [1.0] - duplicate string - deleted

32) ADJP-PRD-1 -> JJR [0.6666666666666666] | RB VBN [0.3333333333333333]

33) ADJP-PRD-1 -> VBN [1.0]

34) *** deleted the latter (checked with the context, there's a rule)

35) VBD ADJP-PRD-1 - both are 2-3 verb forms -> seems unlkely that they stand together in a sentence WITHOUT an intermediate)

36) SBAR-ADV-3 -> IN S [1.0] - partial duplicate - deleted

37) ADVP-DIR-4 -> RB NP PP [1.0] - duplicate string - deleted

38) PP-LOC-2 -> IN NP [1.0] - duplicate string - deleted

39) PP-TMP-1 -> IN NP [1.0] - DS - deleted

40) ADVP-3 -> RB NP [1.0] - DS - deleted

41) VP-2 -> VBN NP PP-CLR [1.0]

42) VP-2 -> VBN NP [0.3333333333333333] | VP 1-1 VP 1-1 VP [0.3333333333333333] | VBG NP [0.3333333333333333]

43) *** 1st rule is possible, added to the original string

44) Result: VP-2 -> VBN NP PP-CLR [0.25] | VBN NP [0.25] | VP 1-1 VP 1-1 VP [0.25] | VBG NP [0.25]

45) NP-2 -> 1-6 CD NONE [0.5] | QP NONE [0.5] - DS - deleted

46) NP-3 -> NNP [0.3333333333333333] | NNP NNPS [0.6666666666666666] - DS - deleted

47) NP-SBJ-1 -> PRP1 CD CD NN NN [0.3333333333333333] | NP PP [0.3333333333333333] | DT NN [0.3333333333333333] - has to be added, 1st is missing, but possible, bgger string has 94 com's + 1 = 95 -> probability is 1\95

48) *** NP-SBJ-1 -> PRP1 CD CD NN NN is missing from the 2nd longer rule

49) *** probabilities in the 2nd rule does not sum to 1

50) *** added NP-SBJ-1 -> PRP1 CD CD NN NN to the 2nd rule, corrected probabilities

51) RESULT: [0.0105263157894736842] AND 1st entry is [0.0009737098344695282] to add up to 1

52) NP-TMP-2 -> JJ JJ NN [1.0] - DS - deleted

53) PP-3 -> IN NP-LGS [1.0] - DS - deleted

54) PP-CLR-2 -> TO NP [0.5] | IN NP [0.5] - left this one, as TO= participle 'to', and IN = prepositions, both combine well enough to have 0.5 value

55) PP-CLR-2 -> TO NP [0.3333333333333333] | IN NP [0.6666666666666666]

56) *** just because think it's right :)

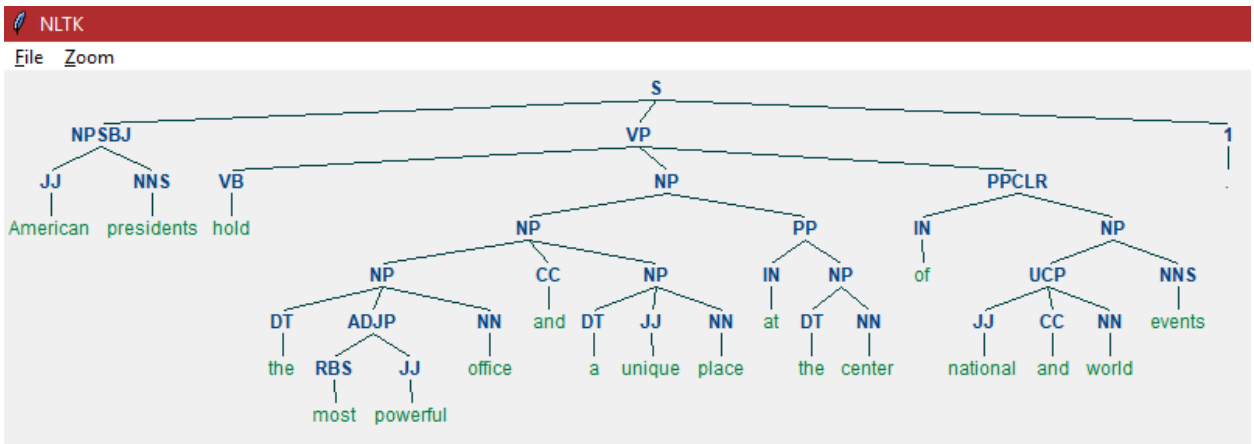
57) NO S -> NP-SBJ-1 VP 1-5 RULE!!!! - discovered a defect in the original script

1.3. Тестирование оптимизированной грамматики на контрольной выборке

Тестирование оптимизированного варианта грамматики проводилось на выборке в сто предложений, правила из которых не извлекались вместе с остальными. В тестировании мы не акцентировали внимание на правильности разбора с экспертной точки зрения, а только исключительно на способности инструмента предложить хоть какой-то вариант разбора.

Грамматике удалось успешно разобрать 87 предложений, что свидетельствует о том, что рекурсивные грамматики являются более универсальными. Среднее время обработки составило ~ 13 секунд, так что теорию о том, что бинарные грамматики являются менее ресурсоёмкимим, подтвердить не удалось.

Результат разбора предложения может выглядеть следующим образом:



Можно заметить, что у данного предложения *S* есть две *основные* непосредственно составляющие (далее НС):

- NPSBJ: «*American presidents*»
- VP: «*hold the most powerful office and a unique place at the center of national and world events*»

Структура NPSBJ, в свою очередь, имеет НС JJ и NNS, которые, впоследствии раскладываются на терминальные (конечные) составляющие: «*American*» и «*presidents*».

1.4. Оценка работы разработанного инструмента и анализ полученных результатов

В грамматике непосредственно составляющих предложение *S* рассматривается, как линейно упорядоченная цепочка единиц. В качестве таких единиц могут выступать знаки пунктуации и другие символы, словоформы или, в некоторых случаях, единицы, больше, чем одна словоформа (например, сложные союзы). Множество единиц (точек) цепочки складываются во множество отрезков, наибольшим из которых будет являться само предложение S.

В нашей работе мы использовали *комбинированные (гибридные) алгоритмы*, как можно понять из названия, совмещают в себе нисходящие и восходящие алгоритмы. Порядок построения изображён на *Рисунке 4*. Как следствие, такие алгоритмы обладают более высокой производительностью, чем нисходящие, и вместе с тем применимы к рекурсивным грамматикам с эллиптическими составляющими. Их преимуществом является то, что в отличие от восходящих алгоритмов движение вверх прекращается не в тот момент, когда требуется объединение нескольких «ветвей», а на шаг позже, что позволяет сравнить порождаемое дерево с фактическим и отбросить порождение тех деривации уровнем ниже, что точно не впишутся в уже сформированную часть.

Нисходящие алгоритмы строят синтаксическое дерево сверху вниз, начиная с самой верхней непосредственно составляющей, которой является само предложение *S*. Порядок построения изображён на *Рисунке 2*. По сути, данные алгоритмы поочерёдно воспроизводят заданные правила, но применяют их сначала к первым компонентам в имеющейся последовательности. В результате, такие алгоритмы сразу выстраивают предполагаемую структуру предложения, однако для того, чтобы привести её в соответствие с входными данными, им необходимо осуществить перебор всех дериваций, порождаемых грамматикой. Процедура повторяется для каждой «ветви» до тех пор, пока **начальные** компоненты порождённого и

фактического высказываний не совпадут, что ведёт худшей производительности и большей ресурсозатратности.

Восходящие алгоритмы строят синтаксическое дерево снизу вверх. В процессе нижестоящие составляющие заменяются вышестоящими до того момента, пока для замены не потребуется соединить несколько «ветвей» в одну – в этот момент движение вверх прекращается до тех пор, пока все позиции уровнем ниже не будут заполнены. Порядок построения изображён на *Рисунке 3*. Специфика работы таких алгоритмов обуславливает их высокую производительность, так как не ведёт к порождению чрезмерного количества дериваций, что характерно для нисходящих алгоритмов. Однако восходящие алгоритмы неприменимы для работы с рекурсивными грамматиками, допускающими пропуски составляющих (эллипсис), что ведёт либо к невозможности анализа предложений с эллипсисом, либо к необходимости внедрять гораздо больше правил, чем потребовалось бы с рекурсивной грамматикой.

Выводы

В практической главе была раскрыта специфика разработки алгоритмы, оптимизации полученной грамматики и приведены примеры правил «до» и «после оптимизации». Также были приведены результаты тестирования полученного инструмента на выборке. Результаты позволяют говорить об успехе проделанной работы.

Заключение

В нашей выпускной квалификационной работе мы описали основные явления синтаксиса английского языка при помощи грамматики структур составляющих.

В ходе работы были выполнены все поставленные задачи:

- анализ синтаксически размеченных корпусов текстов, доступных на платформе Natural Language Toolkit
- создание программы для автоматического извлечения правил из синтаксически размеченного корпуса текстов
- анализ извлечённой грамматики и её оптимизация
- тестирование оптимизированной грамматики на контрольной выборке
- оценка работы разработанного инструмента и анализ полученных результатов

Возможные направления для развития исследования включают в себя ряд шагов для повышения точности синтаксического анализа: расширение и усовершенствование грамматики, подключение дополнительных инструментов для получения более подробной информации о словоформах на уроне токенизации, а также применение парсера для разнообразных задач автоматической обработки текста и его интеграция с другими инструментами NLTK.

Список использованной литературы

1. Беляева Л. Н. Автоматический (машинный) перевод // Прикладное языкознание: Учебник. СПб., 1996.
2. Буторов, В. Д. Моделирование синтаксиса естественного языка / В. Д. Буторов; В. В. Богданов; Г. Я. Мартыненко; А. С. Штерн; И. В. Азарова. Прикладное языкознание / отв. ред. А. С. Герд - СПб. : Изд-во СПбГУ, 1996.
3. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения. М.: Наука, 1985.
4. Иомдин Л. Л., Петроченков В. В., Сизов В. Г., Цинман Л. Л. Синтаксический анализатор системы ЭТАП: современное состояние. / ред. А.е. Кибрик. // В кн.: Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог» (2012). — Вып. 11. — Т.2. — М.: Изд-во рГГУ, 2012.
5. Иорданская, Л.Н. Автоматический синтаксический анализ. Том 2. Межсегментный синтаксический анализ. / Л.Н. Иорданская; ред. А.А. Ляпунова, О.С. Кулагина. Новосибирск: Наука, 1967.
6. Каневский, Е.А., Семантико-синтаксический анализатор SEMSIN. / Е.А. Каневский, К.К. Боярский; ред. А.Е. Кибрик. // Научно-технический вестник информационных технологий, механики и оптики. — СПб: Университет ИТМО, 2015.
7. Леонтьева Н. Н. Автоматическое понимание текстов. Системы, модели, ресурсы М.: Academia, 2006.
8. Мельчук И.А. Автоматический синтаксический анализ. Т. 1. Общие принципы. Внутрисегментный синтаксический анализ. / И.А. Мельчук; ред. А.А. Ляпунова, О.С. Кулагина. Новосибирск: Наука, 1967.
9. Старостин, А.С. Алгоритм синтаксического анализа, используемый в системе морфо-синтаксического анализа «TREETON», 2008.
10. Старостин, М.Г. Мальковский; ред. Л.Л. Иомдин, Н.И. Лауфер, А.С. Нариньяни и др. / В кн.: Компьютерная лингвистика и

интеллектуальные технологии: Труды международной конференции «Диалог 2007». — М.: Изд-во рГГУ, 2007.

11. Теньер Л. Основы структурного синтаксиса. М.: Прогресс, 1988.

12. Тестелец Я. Г. Введение в общий синтаксис. М.: РГГУ, 2001.

13. Bies, Ann; Ferguson, Mark; Katz, Karen; MacIntyre, Robert. Bracketing Guidelines for Treebank II Style. Department of Computer and Information Science, University of Pennsylvania, 1995.

14. Bird, S. Natural Language Processing with Python: analyzing Text with the Natural Language Toolkit. / S. Bird, E. Klein, E. Loper. Beijing, 2009.

15. Brill, Eric. Discovering the lexical features of a language. In Proceedings, 29th Annual Meeting of the Association for Computational Linguistics. Berkeley CA, 1991.

16. Brill, Eric. A Corpus-based Approach to Language Learning. PhD Dissertation, University of Pennsylvania, 1993.

17. Brill, Eric; Magerman, David; Marcus, Mitchell P.; and Santorini, Beatrice. Deducing linguistic structure from the statistics of large corpora. In Proceedings, DARPA Speech and Natural Language Workshop, 1990.

18. Church, Kenneth W. Memory Limitations in Natural Language Processing, MIT LCS Technical Report 245. Master's thesis, Massachusetts Institute of Technology, 1980.

19. Church, Kenneth W. A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the Second Conference on Applied Natural Language Processing. 26th Annual Meeting of the Association for Computational Linguistics, 1988.

20. Cohen S. B., Satta G., Collins M. Approximate PCFG Parsing Using Tensor Decomposition // Proc. of NAACL 2013.

21. Francis, W. Nelson and Henry Kucera. Frequency Analysis of English Usage. Lexicon and Grammar. Houghton Mifflin, Boston, 1982.

22. Garside, Roger, Geoffrey Leech, and Geoffrey Sampson. The Computational Analysis of English. A Corpus-based Approach. Longman, London, 1987.

23. Gildea, D. Synchronous Context-Free Grammars and optimal Parsing Strategies. / D. Gildea, G. Satta // Computational Linguistics, 2016.

24. Hindle, Donald. Acquiring disambiguation rules from text. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989.

25. Jurafsky D., Martin J. H. Speech and Language Processing, 2nd Edition. Prentice Hall, 2008.

26. Kroch, Anthony S. and Ann Taylor. The Penn-Helsinki Parsed Corpus of Middle English, Second Edition. Department of Linguistics, University of Pennsylvania, 2000.

27. Levine, R.D. Head-Driven Phrase Structure Grammar Linguistic approach, Formal Foundations, and Computational Realization. / R.D. Levine, W. D. Meurers. The Ohio State University, 2004.

28. Magerman, David, and Marcus, Mitchell P. Parsing a natural language using mutual information statistics. In Proceedings of AAAI-90, 1990.

29. Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(2), 1993.

30. Marcus, Mitchell P., Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn Treebank: Annotating predicate-argument structure. In ARPA Human Language Technology Workshop, 1994.

31. Meteer, Marie; Schwartz, Richard; and Weischedel, Ralph. Studies in part of speech labelling. In Proceedings, Fourth DARPA Speech and Natural Language Workshop. February 1991.

32. Marneffe, de M.-C. Generating Typed Dependency Parses from Phrase Structure Parses. / M.-C. de Marneffe, B. MacCartney, C. D. Manning, 2006.

33. Niv, Michael. Syntactic disambiguation. In The Penn Review of Linguistics, 14, 1991.

34. Pollard, C. Head-Driven Phrase Structure Grammar. / C. Pollard, I. a. Sag. Chicago: University of Chicago Press and Stanford, 1994.

35. Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. A Comprehensive Grammar of the English Language, Longman, London, 1985.

36. Santorini, Beatrice. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990.

37. Santorini, Beatrice and Mary Ann Marcinkiewicz. Bracketing Guidelines for the Penn Treebank Project. Department of Computer and Information Science, University of Pennsylvania, 1991.

38. Santorini, Beatrice. Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing), 1990.

39. Sleator, D. Parsing English with a Link Grammar. / D. Sleator, D. Temperley. // Carnegie Mellon University Computer Science technical report CMU-CS-91-196, 1991.

40. Taylor, A. The Penn treebank: an overview. / A. Taylor, M. Marcus, B. Santorini; ed. A. Abeille. // Building and Using Parsed Corpora, volume 20 of Text, Speech and Language Technology, 2003.

41. Veilleux, N. M., and Ostendorf, Mari. Probabilistic parse scoring based on prosodic features. In Proceedings, Fifth DARPA Speech and Natural Language Workshop, 1992.

42. Weischedel, Ralph; Ayuso, Damaris; Bobrow, R.; Boisen, Sean; Ingria, Robert; and Palmucci, Jeff. Partial parsing: a report of work in progress. In Proceedings, Fourth DARPA Speech and Natural Language Workshop, 1991.

Приложение

Приложение содержит конечный вариант грамматики после оптимизации. Оптимизация позволила сократить количество правил с 3914, до 1604.

«S -> NPSBJ PPTMP VP 1 [0.00097] | CC 2 PP 2 NPSBJ VP 1 77 [0.00032] | NPSBJ 2 PP 2 VP [0.00032] | NPSBJ VP 5 [0.00064] | SBARTMP 2 NPSBJ VP 1 [0.00225] | CC SBARTMP 2 NPSBJ VP 1 [0.00064] | S 2 S 1 [0.00032] | NPSBJ PPTMP VP [0.00032] | ADVPTMP NPSBJ VP 1 [0.00129] | PP NPSBJ VP 1 [0.00064] | ADVP 2 NPSBJ VP [0.00064] | NPSBJ NPTMP VP [0.00064] | NPSBJ PRN VP 1 77 [0.00064] | PPTMP 2 ADVP 2 NPSBJ VP 1 [0.00032] | CC PPTMP 2 NPSBJ VP 1 [0.00032] | S 2 NPSBJ VP 1 [0.00032] | NPSBJ 66 NPPRD 77 [0.00032] | 66 S 2 CC S 1 77 [0.00032] | S 5 S 1 77 [0.00032] | S 2 CC S 1 77 [0.00032] | STPC 2 NPSBJ VP 1 [0.01320] | SNOMSBJ VP [0.00161] | NPSBJ PP VP [0.00032] | ADVP 2 NPSBJ VP 1 77 [0.00032] | CC NPSBJ VP 1 [0.01675] | PPLOC 2 NPSBJ VP 1 77 [0.00064] | NPSBJ NPTMP VP 1 [0.00032] | ADVP PRN NPSBJ VP 1 [0.00064] | SBARTMP 2 NPSBJ ADVP VP 1 [0.00032] | NPSBJ VP 1 [0.26216] | ADVP NPSBJ VP [0.00032] | 66 NPSBJ 77 VP [0.00032] | NPSBJ 2 VP 1 [0.00032] | PP 2 NPSBJ VP 1 [0.01385] | PPTMP 2 NPSBJ VP [0.00097] | NPSBJ 2 PPTMP 2 VP [0.00064] | PP 2 SBARADV 2 NPSBJ VP 1 [0.00032] | SPRP 2 NPSBJ VP 1 [0.00064] | SBARTMP 2 NPSBJ VP [0.00032] | SBARADV 2 NPSBJ VP [0.00097] | PPLOC 2 ADVP 2 NPSBJ VP 1 [0.00032] | PPTMP NPSBJ VP [0.00129] | NPSBJ ADVPTMP VP [0.00483] | PPLOC 2 PP 2 NPSBJ VP 1 [0.00032] | PPTMP 2 NPSBJ ADVP VP 1 [0.00032] | NPSBJ ADVP VP [0.00451] | ADJPTMP 2 NPSBJ VP 1 [0.00032] | S 5 S 1 [0.00515] | S CC S 1 77 [0.00097] | PPLOC 2 NPSBJ VP [0.00032] | NPSBJ 2 SBARADV 2 VP 1 [0.00032] | NPTMP NPSBJ VP 1 [0.00032] | PP 2 NPSBJ VP 1 77 [0.00032] | CC PP 2 NPSBJ VP 1 [0.00097] | CC NPSBJ ADVP VP 1 [0.00161] | ADVPTMP NPSBJ VP [0.00064] | PP NPSBJ VP [0.00161] | SADV 2 NPSBJ VP 1 [0.00322] | NPSBJ 2 ADVP 2 VP 1 [0.00161] | NPSBJ VP 2 [0.00032] | PPLOC NPTMP 2

NPSBJ VP 1 [0.00322] | SBARPRP 2 NPSBJ VP [0.00032] | 66 NPSBJ VP 1 77 [0.00580] | CC SBARADV 2 NPSBJ VP 1 [0.00064] | NPTMP NPSBJ VP [0.00064] | NPSBJ 2 PP 2 VP 1 [0.00161] | SNOMSBJ VP 1 [0.00064] | LST NPSBJ VP 1 [0.00161] | PPPRP NPSBJ VP [0.00064] | NPSBJ VP 2 CC S 1 [0.00032] | NPSBJ PPPRD [0.00064] | NPSBJ NPPRD [0.00515] | ADVPTMP 2 NPSBJ VP 1 [0.00451] | PPTMP PRN NPSBJ VP 1 [0.00064] | 66 S CC S 1 [0.00032] | NPSBJ VP [0.43800] | PP 2 NPSBJ ADVP VP 1 [0.00032] | NPTMP 2 NPSBJ VP 1 [0.00354] | NPSBJ ADVP VP 1 77 [0.00032] | LRB NPSBJ VP 1 RRB [0.00032] | PPLOC 2 NPSBJ VP 1 [0.01063] | PPTMP NPSBJ VP 1 [0.00064] | NPSBJ ADVPTMP VP 1 [0.00483] | S 5 CC S 1 [0.00032] | CC PRN NPSBJ VP 1 [0.00064] | S 2 CC S [0.00161] | 66 NPTTLSBJ 77 VP [0.00032] | 66 S CC S 1 77 [0.00064] | NPSBJ VP 1 77 [0.00805] | S 2 CC S 1 [0.00934] | CC PPLOC 2 NPSBJ VP 1 [0.00064] | S 2 77 CC S 1 [0.00032] | NPSBJ ADJPPRD [0.00386] | PPTMP 2 NPSBJ VP 1 [0.01514] | SADV 2 NPSBJ VP [0.00032] | STPC 2 77 NPSBJ VP 1 [0.00161] | 66 SBARADV 2 NPSBJ VP 1 77 [0.00097] | PPTMP 2 NPSBJ ADVPTMP VP 1 [0.00032] | 66 NPSBJ VP 5 [0.00064] | NPSBJ ADVP VP 1 [0.01256] | SADV 2 NPSBJ VP 1 77 [0.00032] | CC NPSBJ VP 1 77 [0.00064] | 66 STPC 2 77 NPSBJ VP 1 [0.01063] | S 2 CC FRAG 1 [0.00032] | NPSBJ 2 VP [0.00032] | PPTMP 2 NPSBJ PRN VP 1 [0.00032] | CC NPSBJ ADVPTMP VP 1 [0.00064] | PPPRP 2 NPSBJ VP 1 [0.00161] | NPSBJ 66 VP [0.00193] | SBARADV 2 NPSBJ VP 1 77 [0.00032] | PPLOC NPSBJ VP 1 [0.00097] | NPSBJ 2 SADV 2 VP 1 [0.00097] | NONE [0.05024] | ADVP NPSBJ VP 1 [0.00097] | S CC S 1 [0.00419] | ADVP 2 NPSBJ VP 1 [0.00934] | ADVP PPLOC 2 NPSBJ VP 1 [0.00032] | ADVP 2 PP 2 NPSBJ VP 1 [0.00032] | CC ADVPTMP NPSBJ VP 1 [0.00064] | NPSBJ 2 PPLOC 2 VP 1 [0.00032] | PP 2 ADVP 2 NPSBJ VP 1 [0.00032] | SBARADV 2 NPSBJ VP 1 [0.00580] | NPSBJ PRN VP 1 [0.00258] | 66 NPSBJ VP 1 [0.00258] | NPSBJ ADVPMNR VP [0.00097] | NPSBJ RB VP [0.00064] | S CC S [0.00258] | NPSBJ PRN VP [0.00032]

ADJP -> JJR JJ [0.00472] | QP NONE [0.08019] | JJ JJ [0.01887] | NPADV JJR [0.00943] | ADJP PP [0.01887] | RB RBR JJ [0.00472] | NN [0.00472] | RB RB JJ [0.00472] | JJ S [0.00472] | NNP JJ [0.01415] | ADJP CC ADJP [0.01887] | CD NN [0.10377] | RB VBN [0.05189] | RB JJ PP [0.01415] | JJR [0.01415] | JJ CC JJ [0.03302] | RB JJ [0.12736] | RB JJR [0.04717] | 4 JJ NONE [0.00943] | RB DT [0.00472] | VBN [0.00472] | JJR CC JJR [0.00472] | RB RB [0.00472] | NP JJ [0.02830] | NNP 2 JJ [0.01887] | CD CD NN [0.01415] | 4 CD NONE [0.02830] | RBS JJ [0.00943] | JJ PPTMP [0.00472] | JJ PPLOC [0.00943] | JJ PP [0.06604] | JJ NPTMP [0.02830] | JJ [0.11792] | RB JJ CC JJ [0.00472] | ADVP VBN [0.00472] | RBR JJ [0.05189] | NNP NNP [0.00472] | ADVP JJ [0.00472]

ADJPPRD -> RB [0.00510] | ADJP 2 ADJP [0.01020] | RB VBN S [0.00510] | VBN PP [0.01531] | RB RBR JJ [0.00510] | RBR JJ PP [0.01020] | JJ S [0.06122] | JJ RB S [0.01020] | VBN [0.02551] | ADJP SBAR [0.05102] | RB JJ PP [0.02551] | JJR [0.02041] | RB JJ S [0.01531] | RB JJ [0.04592] | RB JJR [0.01020] | RB JJ SBAR [0.00510] | VBN SBAR [0.01531] | ADJP CC ADJP [0.02551] | NONE [0.01531] | JJ IN [0.00510] | RB VBN [0.01531] | JJ CC JJ [0.01531] | RB VBN PP [0.01020] | RBS JJ [0.00510] | JJ PPLOC [0.00510] | JJ PP [0.12755] | ADJP PP [0.05102] | RBR JJ S [0.01020] | RBR JJ [0.01531] | JJ SBAR [0.01531] | ADVP JJ [0.00510] | JJ [0.34184]

ADJPTMP -> JJ PP [1.00000]

ADVP -> IN PP [0.00699] | RB [0.77972] | RB RB [0.02797] | IN [0.00350] | RBR IN [0.00699] | RB NP [0.02098] | RBR RB [0.00699] | ADVP PP [0.00350] | JJS [0.00350] | RB PP [0.03497] | RB JJ [0.00350] | RB JJR [0.00350] | DT [0.00350] | RB RBR RB [0.00350] | NONE [0.03846] | RB NP PP [0.01049] | IN JJS [0.00699] | RBR [0.00699] | IN NN [0.00350] | RB RBR [0.00350] | JJ [0.00699] | ADVP SBAR [0.00350] | IN NP [0.00350] | JJR [0.00350] | IN DT [0.00350]

ADVPCLR -> RB [0.14286] | JJR [0.07143] | NP JJR [0.07143] | RB JJR [0.14286] | RB NP PP [0.14286] | RB PP [0.07143] | RB NP [0.28571] | NONE [0.07143]

ADVPDIR -> IN [0.18182] | RB PP [0.09091] | RB [0.54545] | RBR [0.09091] | RB NP [0.09091]

ADVPLOC -> IN [0.08333] | RB [0.66667] | NONE [0.25000]

ADVPLOCPRD -> NONE [1.00000]

ADVPLOCPRDTPC -> RB [1.00000]

ADVPMNR -> ADVP PP [0.03390] | RB [0.61017] | RB RB [0.08475] | NONE [0.20339] | RBR [0.01695] | RBS RB [0.01695] | RBR RB [0.03390]

ADVPPRD -> RB [0.85714] | JJ NP PP [0.14286]

ADVPTMP -> RB [0.61745] | JJ [0.00671] | RB RB [0.03356] | NP IN [0.04027] | IN [0.00671] | NP RBR [0.00671] | RB RBR [0.01342] | ADVP PP [0.00671] | IN RB [0.00671] | NP RB [0.03356] | NONE [0.17450] | RB CC RB [0.00671] | NP JJR [0.04698]

CONJP -> RB RB IN [1.00000]

LST -> LS 1 [1.00000]

NAC -> NNP PP [1.00000]

NACLOC -> NNP 2 NNP 2 [0.42857] | NNP NNP 2 NNP 2 [0.57143]

NP -> DT NP CC NP [0.00012] | DT NN POS [0.00950] | QP NNS [0.00407] | NP 5 NP 1 [0.00185] | DT NN CC NNS [0.00012] | NNP CC NNP POS [0.00037] | PDT DT NN [0.00049] | NNP NNP NNP NNP [0.00296] | NP PP [0.09996] | NP CC NP PP [0.00012] | JJ 2 JJ NN [0.00012] | DT ADJP NNP NN [0.00037] | NP JJ NN [0.00272] | JJ NNP NNP NNP [0.00025] | DT JJ NN [0.03061] | CD NNP NNS [0.00037] | DT JJ NN S [0.00025] | NP 2 NP 1 [0.00037] | JJ CD [0.00111] | PRP4 NN [0.00740] | PRP4 JJ NNS [0.00136] | 66 JJ 77 NNS [0.00012] | NNP CD NNS [0.00025] | NN CC JJ NNS [0.00012] | NN NNS NNS [0.00025] | NP 2 PP 2 SBAR [0.00012] | DT QP NNS [0.00012] | DT JJ NNP NNP [0.00012] | QP NONE [0.01752] | DT ADJP NNS [0.00049] | DT ADJP NN NNS [0.00012] | NNS POS [0.00062] | PRP4 JJ NN NNS [0.00012] | JJ NX [0.00049] | DT NN CC NN [0.00111] | NNP NNP 1 [0.00037] | DT JJ NN QP [0.00025] | DT NX [0.00012] | DT NNP NNP NNP NN [0.00012] | DT VBN NN [0.00123] | DT 66 NN 77 [0.00012] | NN NN NN [0.00086] | JJ JJ NNS [0.00518] |

NNP JJ NN [0.00012] | JJR NNS [0.00197] | NP NX [0.00025] | DT NNS S [0.00025] | NP VP [0.00987] | NP PPTMP [0.00370] | NP NNP NNP NNP NN [0.00012] | NP [0.00037] | CD NP NNS [0.00025] | NNS NNS [0.00123] | DT CD NNS [0.00160] | NNP NNPS NNP NNP [0.00012] | CD NN JJ NNS [0.00012] | JJ NNP NNP [0.00025] | NP 2 VP 2 [0.00012] | DT JJR NN NN [0.00025] | NPTMP 5 NP [0.00012] | DT 66 NN [0.00025] | NP NNP NNP NNP [0.00012] | CD NN NNS [0.00049] | ADJP NN NNS [0.00037] | NP 5 NP 5 [0.00012] | DT NN NNS POS [0.00025] | NP VP 1 [0.00012] | NP 5 NP [0.00123] | DT JJ NNP NN [0.00037] | DT JJ NNS NN [0.00012] | DT NNP POS [0.00049] | NNP NNPS POS [0.00037] | NNP 2 NNP CC NNP [0.00049] | DT UCP NN [0.00012] | DT NNP NNP POS [0.00136] | NNP POS [0.01000] | NP PP SBAR [0.00210] | NNP 1 [0.00062] | VBG NN [0.00062] | DT CD [0.00062] | DT 66 NN 77 NN [0.00025] | DT NNS NN [0.00099] | DT JJ VBG NN [0.00025] | NN NNP NNP [0.00049] | NP PP PP PP [0.00025] | CD JJ NNS [0.00160] | NN CD [0.00012] | DT NN NN NN [0.00160] | NNS NN [0.00025] | JJR NN [0.00111] | NP ADVPTMP [0.00062] | NP NN SBAR [0.00025] | PRP4 NN NNS NN [0.00012] | RB JJR [0.00025] | PRP4 VBN NN [0.00025] | NP 77 PPLOC [0.00025] | CD NN [0.01333] | DT VBN NN NNS [0.00025] | NONE [0.04875] | DT NNPS CC NNP NNP [0.00012] | NNP NNP POS [0.00629] | DT VBG JJ NN [0.00012] | PRP4 NNP NNP NNP [0.00012] | NNP CC NNP NNP [0.00012] | NP CONJP NP [0.00025] | DT JJ JJ NN [0.00210] | CD NN NN [0.00012] | PRP4 NNS POS [0.00012] | JJ NNP NN NNS [0.00012] | DT VBG NNS [0.00037] | ADJP JJ NN [0.00037] | NP PP 2 NP [0.00025] | DT NN NNS [0.00160] | CD JJ NN [0.00037] | 5 NP PP 2 SBAR 1 [0.00037] | DT 66 NN NN 77 [0.00012] | NNP NNP NNPS [0.00037] | JJ JJ NN NN [0.00025] | JJR JJ NNS [0.00049] | JJR NN NN NN [0.00012] | DT JJ NN NNS [0.00074] | RB JJ DT NN [0.00025] | DT VBN NNS [0.00012] | DT CD JJ NNS [0.00025] | NP QP NONE [0.00037] | DT NN SBAR [0.00136] | DT ADJP JJ NN [0.00111] | DT NN NN POS [0.00062] | NP 5 S 1 [0.00012] | NP CC NP [0.01172] | NP 2 SBAR 2 [0.00099] | JJ DT NN [0.00012] | NP PP PPTMP [0.00136] | NNP NNPS [0.00123] | JJ VBG [0.00012] | NN CC NN [0.00086] | DT

NNS NNS [0.00025] | CD NNS [0.01283] | JJS NNS [0.00037] | NNP CD [0.00284] | CD [0.01246] | NP ADVP [0.00037] | NP 2 NP 2 NP [0.00049] | NP NNS S [0.00025] | JJ NNP NNP NN [0.00012] | VBN JJ NN [0.00012] | NN 2 NN CC NN [0.00012] | DT JJ CC JJ NNS [0.00012] | RB NN NNS [0.00012] | PRP4 NNP NNP NN [0.00025] | QP NN [0.00346] | QP [0.00111] | PRP4 NN POS [0.00012] | NP PPLOC [0.01518] | NP PP 2 VP [0.00012] | NNP NNP NNP 1 [0.00025] | PRP4 NACLOC NN [0.00025] | NN CC NNS [0.00037] | DT NAC NN [0.00012] | DT NNP NNP [0.00395] | NP SBARLOC [0.00037] | DT JJ NNP NN NN [0.00037] | NNS SBAR [0.00074] | NP 2 NP 2 SBAR [0.00012] | NAC POS [0.00012] | PRP4 NN NNS [0.00111] | JJ NNP NNS [0.00074] | NP PP VP [0.00123] | NP PP 2 PP [0.00012] | DT VBN ADJP NN [0.00012] | NP 2 VP [0.00074] | NP PPDIR PPDIR [0.00012] | NP NPLOC 1 [0.00025] | DT NNP NNP CC NNP NNP [0.00025] | DT JJ NN POS [0.00037] | NNP NNS NNS [0.00012] | NP NN CC NN NN NN [0.00012] | NN VBG NN [0.00012] | DT JJ [0.00111] | VBG JJ NNS [0.00012] | DT NNP NNP NNP NNP [0.00136] | DT NNP NNP NNS [0.00025] | DT NNP CD [0.00049] | DT 66 JJ 77 NN [0.00012] | NP 2 ADVPTMP [0.00012] | JJ NN S [0.00049] | CD VBN NNS [0.00037] | DT JJS NNS [0.00074] | NP 5 NP 5 NP 5 NP 5 NP 5 NP [0.00012] | VBN NNS [0.00099] | NP CD NN NNS [0.00025] | RB NN [0.00049] | PRP4 JJ NN NN [0.00025] | NP 2 PP [0.00160] | NP PP PPLOC [0.00173] | DT ADJP NNS NN [0.00012] | DT NNPS [0.00111] | JJR [0.00136] | PRP4 NN VBG NN [0.00025] | PRP4 NN S [0.00049] | NP JJ NN NNS [0.00025] | NNP NNP NNP [0.01382] | NNS CC NNS [0.00086] | ADJP NNP NNS [0.00012] | NP 5 NP 5 NP 5 NP 5 NP [0.00012] | NNP CD 2 CD [0.00012] | DT VBN NN NN [0.00012] | NNP NN NN [0.00012] | NNP 2 NNP [0.00012] | NP SBARTMP [0.00012] | RB CD [0.00012] | NNP NNP NNP NNP POS [0.00012] | JJ NNS NNS [0.00037] | DT JJ CD NNS [0.00160] | NN NN CC NN [0.00012] | JJ NNP [0.00025] | NP 2 CONJP NP [0.00012] | NP JJ JJ NN [0.00025] | NNS CC NN [0.00025] | NP RRC [0.00012] | PRP4 NNS [0.00629] | DT NNP NNP NN [0.00074] | VBN NN NN [0.00012] | NNP NNP NNS [0.00012] | DT ADJP NN [0.00506] | ADJP NN NN [0.00025] | NP ADJP PP

[0.00012] | JJS [0.00086] | DT NNP NNP NNP NNP NNP [0.00012] | NP SBAR [0.01518] | RB DT JJ NNS [0.00025] | DT 66 NN S [0.00012] | DT JJR NN [0.00086] | NNP NNPS NNP [0.00074] | DT 66 JJ NN 77 [0.00012] | DT JJ NNS [0.00592] | DT JJ CD [0.00025] | JJ CD POS [0.00012] | ADJP NNS [0.00086] | CD CD [0.00123] | NN POS [0.00099] | JJ NNS POS [0.00025] | DT NNP CD NN [0.00049] | 66 NPTTL 2 77 NP [0.00025] | RB DT JJ NN [0.00037] | PRP4 NX [0.00012] | NP NNP NN [0.00049] | RB [0.00086] | JJ VBG NNS [0.00037] | NNP NNP CC NNP [0.00099] | DT JJ NN VBG NN [0.00012] | NNP NNP CC NNP NNP [0.00012] | RB JJ NNS [0.00012] | NNP NNP NN NN [0.00012] | NP VP PP [0.00025] | DT JJ JJ NNS [0.00025] | NNS S [0.00025] | DT JJR NNS [0.00012] | DT NNP NNS [0.00049] | RB CD NN [0.00025] | VBG NNS [0.00123] | PRP4 NN NN [0.00173] | JJ VBN NNS [0.00012] | NP 2 NP 2 CC NP [0.00025] | DT JJ QP NONE [0.00037] | 66 NPTTL 77 PRN [0.00012] | DT ADJP QP NONE [0.00012] | ADJP DT NN [0.00012] | 4 CD NONE [0.01135] | NP CD NN [0.00012] | NNP JJ NNS [0.00012] | NP 2 NP 2 NP CC NP [0.00049] | NP 2 NPLOC 2 [0.00086] | NNP NN NNP [0.00012] | NP 5 S [0.00012] | NN SBAR [0.00049] | NNP NNP [0.03135] | NP 2 SBAR [0.00420] | PRP4 NN NN NN [0.00012] | JJR NN NNS [0.00012] | DT NN NNS NN [0.00037] | PRP4 NN CC NN NN [0.00025] | NP VBN NN [0.00025] | NP 2 PP 2 [0.00025] | ADJP NN [0.00062] | PRP4 NNP NN NN [0.00012] | QP NN NNS [0.00012] | JJ NN CC NN NN NNS [0.00012] | DT [0.00531] | CD CD NN [0.00062] | NP 5 NP 5 NP [0.00025] | DT CD CC CD [0.00012] | DT JJ CD NN [0.00037] | NP 2 NP CC NP [0.00185] | DT NNS CC NNS [0.00025] | PRP4 ADJP NN [0.00012] | NP PRN [0.00111] | NP PPLOC PPTMP [0.00025] | RB NNS [0.00025] | NP 2 NP CC NP 2 [0.00025] | JJ JJ NN NNS [0.00049] | NP PP SBARTMP [0.00012] | NP PRN SBAR [0.00025] | JJ CC JJ NNS [0.00012] | DT NNP NN [0.00568] | PRP4 CD NN NNS [0.00012] | DT JJ NNP NNP NN [0.00037] | DT VBG NN [0.00111] | NP NP [0.00222] | DT NNP CC NNP NNP [0.00012] | ADJP JJ NNS [0.00025] | DT JJ JJ NN NN [0.00099] | NN JJ NNS [0.00012] | DT JJS JJ NNS [0.00037] | NP ADJP JJ NN [0.00012] | DT JJ NNS POS [0.00012] | PRP [0.01074] | NP NNP

NNP [0.00136] | DT NACLOC NN [0.00025] | DT NN [0.08614] | NN NN NNS [0.00197] | NNPS [0.00037] | DT NNS POS [0.00037] | NP PPLOC SBAR [0.00099] | NP 2 ADVP [0.00123] | PRP4 NNP NNP NNS [0.00012] | NP NN NN NNS [0.00012] | NP 2 NPLOC [0.00074] | DT NN CC NN NN [0.00025] | NP NN [0.00728] | VBG [0.00049] | NNP NNP NNP POS [0.00037] | NNP NNP NNP NNP JJ NN [0.00049] | NN CC NN NNS [0.00074] | DT NNP JJ NN [0.00037] | NP PP PP PPTMP [0.00025] | DT CD NN NN [0.00012] | DT NACLOC NN NN [0.00012] | NNP NNP NNP NNP NNP NNP [0.00012] | JJ NNS CC NN [0.00012] | DT JJS NN NN [0.00025] | PRP4 CD NNS [0.00025] | DT JJ QP NNS [0.00037] | NN [0.04517] | DT JJ 2 JJ NN [0.00049] | PRP4 UCP NNS [0.00012] | NP 2 CC NP 2 [0.00160] | NN NNS [0.01370] | NNP [0.03727] | JJ NN NN NNS [0.00037] | DT NNP [0.00432] | RB JJ [0.00012] | DT NNP NNP NNP NN NN [0.00025] | NNP DT NNP [0.00012] | DT JJS JJ NN [0.00012] | NP PP ADJP 2 PP PP 1 [0.00012] | NP NPTMP [0.00074] | CD CC CD [0.00012] | DT JJ NN 66 NN 77 NN [0.00012] | NP ADJP [0.00346] | DT NN NNP [0.00012] | VB NNS [0.00012] | JJ NNS [0.02962] | JJ 2 JJ NN NNS [0.00012] | PRP4 NNP [0.00012] | JJ NN NN NN [0.00049] | NP QP NNS [0.00012] | NP PP PPDIR [0.00025] | ADJP NNP NNP [0.00012] | VB [0.00025] | NNP NNP NNP CC NNP [0.00012] | DT JJ VBN NN [0.00025] | DT ADJP NN NN [0.00086] | NP 2 NP 2 NP 2 NP 2 NP CC NP [0.00049] | NP JJS NNS [0.00025] | NP PP ADVP [0.00012] | NP NNS [0.00272] | DT VBG NN NN [0.00025] | NP NPADV [0.00617] | DT JJ 2 JJ NNS [0.00025] | DT JJS NN [0.00099] | JJ NNS S [0.00012] | DT 4 CD NONE [0.00012] | DT NNP NN POS [0.00025] | NP PPDIR [0.00148] | JJ NNS CC NNS [0.00049] | RB DT [0.00037] | DT NNS [0.01493] | QP JJ NNS [0.00049] | NP PP 2 SBAR [0.00037] | VBN NN NNS [0.00049] | NN JJ NN [0.00037] | NP NPLOC [0.00049] | NP PP 2 NP 2 [0.00025] | NNP NN [0.00086] | JJ NN NNS [0.00444] | ADJP NNP [0.00037] | NP PP PP [0.00346] | DT VBG NNP NN [0.00012] | NP VBG NN [0.00025] | NNS [0.03999] | NP JJ NN NN [0.00062] | NP ADVPLOC [0.00012] | NP ADVP PP [0.00012] | NP 2 NP 2 [0.00259] | DT JJ NNP NNS [0.00012] | 66 NP 2 77 SBAR [0.00025] | DT JJ VBN NNS [0.00012] | DT JJ NN CC NN

[0.00012] | DT ADJP JJ NN NN [0.00012] | NP 2 CC NP [0.00296] | DT NN QP NONE [0.00012] | JJ NNS NN [0.00012] | NP JJ NNS [0.00160] | NP NN NNS [0.00062] | JJ JJ NN [0.00259] | DT JJ NN NN [0.00432] | NP JJ JJ NNP NN [0.00025] | PRP4 ADJP JJ NN [0.00012] | NP PRN PP [0.00012] | JJ NN CC NN NNS [0.00049] | JJR NN NN [0.00037] | NP SBARPRP [0.00062] | DT JJ JJ JJ NN [0.00025] | NP 2 RB NP [0.00025] | NP 2 NP [0.00642] | NNP 2 NNP CC NNP NNP [0.00012] | DT JJ NNP [0.00025] | JJ NN [0.01752] | NP 2 NP 2 PP 1 [0.00025] | DT VBG CD NNS [0.00037] | NN NN [0.00901] | NP 2 ADJP [0.00025] | PDT PRP4 NNS [0.00012] | NP PP NPTMP [0.00062] | NP NN NN [0.00197] | NP 2 SBARLOC [0.00012] | VBN NN [0.00099] | NP NN S [0.00123] | JJ NN CC NN [0.00062] | NP QP 1 [0.00062] | NP NNP [0.00025] | PRP4 NN NN NNS [0.00012] | JJ CD NNS [0.00012] | NP NN JJ [0.00025] | JJ NN NN [0.00321] | PRP4 NNS S [0.00012] | DT NNP NNP NNP [0.00197] | NNP NNS [0.00197] | NP 2 NP 2 NP 2 NP [0.00012] | QP NNP NNS [0.00012] | PRP4 JJS NNS [0.00012] | NP PP 1 [0.00037] | NP ADJP NN [0.00012] | JJS JJ NNS [0.00012] | NP PP ADVPTMP [0.00012] | NP PP S [0.00049] | DT CD CD [0.00025] | PRP4 JJ NN [0.00309] | DT NN JJ NN [0.00049] | DT JJS [0.00049] | DT JJ 66 NN NN 77 [0.00025] | JJ NN POS [0.00111] | PRP4 JJ ADJP NN [0.00012] | NNP CC NNP [0.00148] | CD JJ NN NNS [0.00037] | DT NN S [0.00210] | DT JJ NN NN NN [0.00025] | NN S [0.00012] | DT NNP NNP NN NN [0.00037] | JJ [0.00160] | DT NNP NNP NNP NNP POS [0.00012] | NP 5 NP 5 NP 5 CC NP [0.00012] | DT NN NN [0.01469] | RB DT NN [0.00025] | NNP NNP NNPS NNP [0.00025] | DT CD NN [0.00123] | DT NNP NN NN [0.00099]

NPADV -> QP NONE [0.03390] | QP NN [0.01695] | DT [0.01695] | NP SBAR [0.01695] | DT NN [0.89831] | CD NN [0.01695]

NPCLR -> NN [1.00000]

NPEXT -> NP NPADV [0.07273] | QP NNS [0.01818] | QP NN [0.01818] | QP [0.03636] | CD NNS [0.10909] | 4 CD NONE [0.03636] | CD CD [0.03636] | CD [0.29091] | CD NN [0.38182]

NPHLN -> NNP NNPS [0.22222] | NNP NNP 5 [0.16667] | NP PRN 5 [0.33333] | NP 5 [0.11111] | NNP 5 [0.11111] | NNP NNP [0.05556]

NPLGS -> DT NP CC NP [0.01695] | NN [0.01695] | NNS [0.03390] | NNP NNP NNP NNP [0.01695] | NN NNS [0.03390] | NP 2 NPLOC [0.01695] | NNP [0.05085] | NP PP [0.20339] | DT NN NN [0.01695] | DT NNP [0.01695] | NP CC NP [0.11864] | DT JJ NN [0.01695] | DT NN [0.05085] | NP 2 NP SBAR [0.03390] | JJ NNS [0.03390] | NP VP [0.01695] | NNP NNP NNP NNPS NNP [0.01695] | DT NNP NNP NNP [0.03390] | NN NNS CC NNS [0.03390] | DT NNP NN [0.05085] | NNP NNP [0.10169] | NP 2 SBAR [0.05085] | NNP CC NNP [0.01695]

NPLOC -> NNP [0.35714] | NNP 1 [0.01786] | NP 2 NP [0.44643] | NNP NNP [0.14286] | NNP 2 NNP [0.03571]

NPPRD -> QP NONE [0.02073] | NP 2 NP [0.02073] | NP PPLOC SBAR [0.01036] | NN [0.01554] | NNS [0.00518] | NP PPLOC [0.02591] | JJ NN [0.00518] | CD NN [0.01036] | NNP [0.01554] | DT [0.00518] | NP ADJP PP [0.00518] | NP PP [0.30570] | DT NN NN [0.03627] | DT JJ NN NN [0.00518] | DT 66 JJ NN [0.00518] | NP SBAR [0.08290] | DT NN SBAR [0.01036] | NP PP PPLOC [0.00518] | DT NN [0.07254] | DT JJ NN [0.11917] | NP 2 NP CC NP [0.01036] | JJ NNS [0.00518] | NP ADJP [0.01036] | NP PP 2 SBAR [0.00518] | DT NN S [0.00518] | 4 CD NONE [0.01036] | PRP4 NN [0.01036] | NN NN [0.02073] | CD [0.01036] | NP VP [0.02073] | ADJP JJ NNS [0.00518] | NP 5 NP [0.00518] | NP PRN PP [0.00518] | NP CC NP [0.01036] | NP PP SBAR [0.02591] | JJ NN NNS [0.00518] | DT JJ [0.00518] | NONE [0.02591] | NP PP PP [0.01036] | NP 2 SBAR [0.01036]

NPSBJ -> DT NP CC NP [0.00214] | NP ADVPLOC [0.00061] | DT NNP NNP NNP NNP [0.00061] | QP NNS [0.00031] | NNP NNP CC NNP [0.00031] | NP NNS [0.00520] | DT JJ NN NNS [0.00061] | NNP NNP NNP NNP [0.00183] | DT JJS NN [0.00031] | PRP4 [0.00031] | NP PP [0.06145] | NP PPDIR [0.00061] | PRP4 JJ NNS [0.00031] | NP PP 2 NPLOC 2 [0.00061] | DT NN VBG NN [0.00092] | DT NN SBAR [0.00061] | NP JJ NN [0.00092] | JJ NNP NNP NNP

[0.00031] | NP CC NP [0.00734] | NP 2 SBAR 2 [0.00887] | DT JJ NN [0.01162] | NP PP PPTMP [0.00092] | NNP NNPS [0.00092] | NP 2 ADJP 2 [0.00153] | CD NNS [0.00031] | NP [0.00031] | PRP4 NN [0.00520] | DT VBN NN [0.00061] | NP S [0.00153] | CD [0.00092] | NP JJS NN [0.00031] | NP PPLOC 2 VP 2 [0.00031] | NP 2 NPLOC 2 [0.00611] | JJ NNPS [0.00031] | NNP NN [0.00214] | JJ NN NNS [0.00183] | JJR NN NNS [0.00061] | NNP NNP NN [0.00061] | NP PP PP [0.00122] | NNP NNP [0.05595] | DT JJ NNP NNP [0.00061] | DT NNP NNP NNP [0.00183] | DT NNPS [0.00031] | DT NNP JJ NN [0.00061] | NNS [0.03332] | NP PPLOC [0.01192] | NNP NNPS NNP [0.00031] | NP 2 NP 2 [0.02048] | DT NN NN NN [0.00061] | NN NN NNS [0.00183] | DT NN NN [0.00887] | DT [0.00978] | NP 2 CC NP 2 [0.00153] | PRP4 NN NNS [0.00061] | VBG NNS [0.00061] | NP 2 CC NP [0.00031] | NP 2 NP CC NP [0.00061] | NN NN NN [0.00092] | JJ JJ NNS [0.00092] | NP PRN [0.00183] | DT NNP NNP NN NNS [0.00031] | NP 2 VP [0.00061] | JJR NNS [0.00214] | NP NN NNS [0.00122] | DT JJ JJ NN NN [0.00031] | DT JJ NN NN [0.00122] | NP VP [0.00397] | NP PPTMP [0.00183] | NP 2 PP 2 [0.00061] | DT JJ [0.00061] | NP 2 SBAR [0.00031] | NNS NNS [0.00183] | DT NNP NNP NNS [0.00031] | DT CD NNS [0.00183] | NN NN [0.00581] | DT NNP NN [0.00764] | JJ NNP NNP [0.00031] | DT VBG NN [0.00031] | NP 2 NP [0.01131] | NP NP [0.00092] | DT NNP CC NNP NNP [0.00061] | DT NNP NNS [0.00092] | DT JJS NNS [0.00061] | NP 5 NP 5 [0.00031] | NP PP PPLOC [0.00031] | NP NN NN [0.00214] | PRP [0.14888] | EX [0.01009] | DT VBN NNS [0.00031] | NP NNP NNP [0.00245] | DT NACLOC NN [0.00061] | DT JJS NN NN [0.00031] | NNP NNP NNP [0.00734] | DT NN [0.06848] | DT NNP CC NNP NN [0.00031] | NNPS [0.00031] | JJ NNS [0.01315] | NP NN S [0.00122] | NNP POS [0.00153] | NP PP S [0.00031] | NNP NN NN [0.00061] | DT NNS NN [0.00153] | NP PP SBAR [0.00031] | NP NNP NNP NNS [0.00061] | NP 2 NPLOC [0.00122] | DT NN NNS [0.00275] | JJ NN NN [0.00153] | NNP NNP NN NNP NNP [0.00061] | JJ [0.00092] | NP PP VP [0.00031] | DT NNS [0.01773] | NNP NNS [0.00764] | NP NN [0.00459] | DT JJ ADJP NN [0.00061] | DT CD NN NN [0.00092] | NP 2 UCP 2 [0.00031] | DT

VBG NNS [0.00061] | NP NNP NNP NNP [0.00061] | JJ NNP [0.00031] | NP PP PP PP [0.00061] | NN [0.01162] | PRP4 JJ NN [0.00061] | NP 2 VP 2 [0.00245] | PRP4 NNS [0.00183] | DT NNP NNP NN [0.00245] | JJ NN [0.00428] | NNP NNP NNS [0.00092] | NNP [0.04158] | JJ NN NN NNS [0.00061] | CD JJ NNS [0.00061] | JJS [0.00031] | NN NNS [0.01406] | VBG NN [0.00031] | JJS NNS [0.00031] | NP SBAR [0.01039] | DT NNP NNP NNP NNP NNP NNP [0.00031] | NP 2 ADVP NP 2 [0.00031] | CD NN [0.00245] | NP NPTMP [0.00031] | NONE [0.23754] | NP ADJP [0.00275] | DT ADJP NN [0.00061] | NP JJ NN NN [0.00061] | DT JJ NNS [0.00459] | DT DT NNS [0.00031] | NP 2 NP 2 NP 2 [0.00061] | NN NNS CC NNS [0.00031] | ADJP NNS [0.00031] | PDT DT JJ NN [0.00031] | DT JJ JJ NN [0.00061] | CD NN NN [0.00031] | NP NNP [0.00061] | DT NNP CD NN [0.00031] | JJ NNS NNS [0.00031] | NNP CC NNP [0.00122] | JJ JJ JJ NNS [0.00031] | NNP NN NNS [0.00122] | NNS CC NNS [0.00092] | DT NNP NNP [0.00489] | NNP NNP NNP NNP NNP [0.00122] | NP PP 2 SBAR 2 [0.00092] | DT NNP CC NNP NNS [0.00031] | DT CD NN [0.00031] | DT NNP [0.00978]

NPTMP -> RBR DT NN [0.00917] | NNP CD 2 CD [0.01835] | NP 2 NP [0.03670] | JJ NNP [0.04587] | NNP CD [0.06422] | CD NNS [0.00917] | JJR DT NN [0.00917] | JJ NNP CD [0.00917] | IN NN [0.01835] | JJ NN [0.13761] | NN CD [0.00917] | NNP [0.13761] | NN [0.23853] | NP PP [0.01835] | CD [0.04587] | NP SBAR [0.02752] | RB DT NN [0.00917] | NP ADVP [0.00917] | DT NN [0.14679]

NPTMPCLR -> NNP CD 2 CD [0.11111] | NNP CD [0.88889]

NPTTL -> DT NNP NNP [0.25000] | NNP NNP [0.75000]

NPTTLSBJ -> NNP [1.00000]

NX -> NX 2 NX CC NX [0.07407] | JJ NN NN [0.03704] | JJ NNS [0.07407] | NX PP [0.03704] | JJ NN [0.07407] | NN [0.18519] | NNS [0.07407] | NX CC NX [0.22222] | NN NNS [0.07407] | NN NN [0.14815]

PP -> IN ADVP [0.00058] | JJ IN NP [0.00351] | IN 66 NPTTL 77 [0.00058] | IN [0.00117] | IN NP [0.82174] | VBN PP [0.00643] | IN NPLOC [0.00058] |

NPADV IN NP [0.00058] | RB RB IN NP [0.00058] | ADVP IN NPLGS [0.00058] | IN PP [0.00117] | VBG PP [0.00409] | TO NP [0.05845] | TO SNOM [0.00117] | NONE [0.00468] | IN NP PPTMP [0.00058] | IN SBARNOM [0.00292] | IN SNOM [0.03098] | ADVP IN NP [0.00292] | IN ADVPTMP [0.00058] | VBG NP [0.00643] | IN ADJP [0.00058] | IN NP ADVPTMP [0.00058] | IN PPTMP [0.00175] | IN 66 NP 77 [0.00058] | CC NP [0.00175] | PP CC PP [0.00292] | IN NP NPTMP [0.00234] | ADVP TO NP [0.00058] | IN NP 2 [0.00058] | RB IN SNOM [0.00117] | IN SBAR [0.00058] | PP PP [0.00117] | RB IN NP [0.00117] | IN NPLGS [0.03390]

PPCLR -> IN PP [0.00477] | IN SBARNOM [0.00955] | IN SNOM [0.02625] | IN ADVPTMP [0.00239] | ADVP IN NP [0.00716] | IN ADJP [0.00477] | RB ADJP [0.00716] | PP 2 CC PP [0.00239] | IN 66 SNOM [0.00239] | IN 66 NP [0.00239] | PP CC PP [0.00239] | IN NP [0.68258] | ADVP TO NP [0.00239] | TO NP [0.22673] | IN SBAR [0.00716] | PP PP [0.00716] | TO SNOM [0.00239]

PPDIR -> IN NP PPTMP [0.00690] | IN NP [0.40690] | IN ADVPTMP [0.00690] | TO NP [0.57931]

PPDIRCLR -> TO NP [1.00000]

PPDTV -> TO NP [1.00000]

PPEXT -> IN NP [1.00000]

PPLOC -> IN NP [0.97175] | IN SNOM [0.00565] | PP CC PP [0.00565] | ADVP IN NP [0.00847] | NONE [0.00847]

PPLOCCLR -> IN NP [1.00000]

PPLOCPRD -> IN NP [0.88889] | ADVP IN NP [0.11111]

PPMNR -> IN NP [0.58974] | IN SNOM [0.35897] | ADVP IN NP [0.05128]

PPPRD -> IN NP [0.78261] | IN SNOM [0.08696] | ADVP IN NP [0.13043]

PPPRDLOC -> IN NP [1.00000]

PPPRP -> IN IN NP [0.45455] | IN NP [0.45455] | IN SNOM [0.04545] | JJ PP [0.04545]

PPPUT -> IN NP [1.00000]

PPTMP -> IN ADJP [0.00373] | IN SNOM [0.01119] | ADVP IN NP [0.00746] | RB SNOM [0.00373] | VBG NP [0.00373] | IN NP [0.95522] | PPDIR PPDIR [0.00373] | PP 2 PP [0.00373] | NONE [0.00373] | TO NP [0.00373]

PPTMPCLR -> IN NP [1.00000]

PPTMPPRD -> IN NP [1.00000]

PPTPC -> IN NP [1.00000]

PRN -> 2 S 2 [0.19565] | 2 SINV 2 [0.02174] | LRB NP RRB [0.13043] | 5 CC NP 5 [0.02174] | 5 NP [0.04348] | LRB NP 2 NPLOC RRB [0.02174] | 2 PPTMP 2 [0.02174] | 2 NPSBJ VP 2 [0.04348] | 5 SBAR 5 [0.04348] | LRB NPLOC RRB [0.15217] | 2 PP 2 [0.13043] | 5 NP 5 [0.04348] | 2 ADVP 2 [0.04348] | 2 NP 2 [0.02174] | LRB S RRB [0.02174] | 5 S 5 [0.02174] | 2 ADVPTMP 2 [0.02174]

PRT -> IN [0.10417] | RB [0.04167] | RP [0.85417]

QP -> 4 CD CD CC JJR [0.00375] | RB 4 CD [0.01124] | CC RB [0.00375] | IN CD CD [0.00749] | JJR IN CD CD [0.00375] | RB JJ IN CD [0.00749] | RBR IN 4 CD CD [0.00749] | JJR IN 4 CD [0.00749] | IN CD [0.05243] | RB 4 CD CD [0.01498] | CD CC JJR [0.00375] | RBR IN CD [0.01124] | IN 4 CD CD CC 4 CD CD [0.00749] | IN 4 CD CD [0.04494] | RB CD CD [0.01124] | JJR IN 4 CD CD [0.01873] | CD CC CD [0.00749] | 4 CD CD [0.49064] | CC CD [0.00375] | RB IN CD [0.00375] | JJ NNS [0.00375] | RB CD [0.02247] | IN TO CD [0.00375] | CD NN TO CD NN [0.01873] | IN JJS CD [0.00375] | 3 CD CD [0.01498] | CD CD [0.17228] | IN 4 CD [0.00749] | JJR IN CD [0.02996]

RRC -> VP [1.00000]

SADV -> NPSBJ ADVP VP [0.07500] | NPSBJ ADJPPRD [0.07500] | NPSBJ VP [0.85000]

SBAR -> SBAR 2 CC SBAR [0.00725] | SBAR CC SBAR [0.00362] | NONE S [0.38889] | DT S [0.00362] | IN 66 S [0.00121] | WHADVP IN S [0.00121] | NONE [0.01691] | IN 2 S [0.00121] | IN S [0.20773] | WHNP 66 S [0.00121] | WHADVP S [0.04106] | WHPP S [0.00242] | WHNP S [0.32246] | SBAR 2 SBAR [0.00121]

SBARADV -> IN S [0.94118] | RB IN S [0.01471] | ADVP IN S [0.02941] | IN FRAG [0.01471]

SBARCLR -> IN S [1.00000]

SBARLOC -> WHADVP S [0.25000] | WHPP S [0.75000]

SBARMNR -> IN S [1.00000]

SBARNOM -> IN S [0.07143] | WHADVP S [0.14286] | WHNP S [0.64286] | SBAR CC SBAR [0.14286]

SBARNOMPRD -> WHNP S [1.00000]

SBARPRD -> IN S [1.00000]

SBARPRP -> IN S [0.52174] | IN NN S [0.08696] | IN 66 S [0.08696] | WHNP S [0.08696] | ADVP IN S [0.08696] | WHADVP S [0.13043]

SBARQ -> WHNP SQ 1 [1.00000]

SBARTMP -> IN S [0.66667] | WHADVP S [0.33333]

SCLR -> NPSBJ VP [1.00000]

SHLN -> NPSBJ VP 1 [0.63636] | NPSBJ VP 5 [0.18182] | NPSBJ VP [0.18182]

SINV -> PPTPC VP NPSBJ 1 [0.01754] | 66 STPC 2 VP NPSBJ 1 [0.01754] | 66 STPC 2 77 VP NPSBJ 1 66 [0.01754] | STPC 2 VP NPSBJ 1 [0.22807] | VP NPSBJ 5 66 S 1 [0.01754] | STPC 2 77 VP NPSBJ 1 [0.07018] | 66 STPC 2 77 VP NPSBJ 1 [0.52632] | VP NPSBJ [0.01754] | VP NPSBJ 2 66 S 1 77 [0.03509] | ADVPLOCPRDTPC VP NPSBJ 1 [0.01754] | 66 S 2 77 VP NPSBJ 1 [0.01754] | VP NPSBJ 5 66 S 1 77 [0.01754]

SNOM -> NPSBJ RB VP [0.01075] | NPSBJ ADVPTMP VP [0.01075] | NPSBJ ADVP VP [0.01075] | NPSBJ VP [0.96774]

SNOMSBJ -> NPSBJ VP [1.00000]

SPRD -> NPSBJ VP [1.00000]

SPRP -> NPSBJ VP [1.00000]

SPRPCLR -> NPSBJ VP [1.00000]

SQ -> VBZ NPSBJ VP [0.33333] | NPSBJ VP [0.66667]

STPC -> CC NPSBJ VP [0.01562] | SBARTMP 2 NPSBJ VP [0.00781] | SBARADV 2 NPSBJ VP [0.03125] | SADV 2 NPSBJ VP [0.00781] | PP 2 NPSBJ VP [0.01562] | NPSBJ ADVPTMP VP [0.02344] | NPSBJ VP [0.71875] | SNOMSBJ VP [0.01562] | NPSBJ ADVP VP [0.05469] | ADVP 2 NPSBJ VP [0.02344] | NPSBJ 66 VP [0.01562] | PPTMP 2 NPSBJ VP [0.02344] | S CC S [0.01562] | S 2 CC S [0.03125]

UCP -> NN CC JJ [0.33333] | JJ CC NN [0.33333] | ADJP CC NP [0.33333]

VP -> VBD NPEXT NPTMP SBARTMP [0.00020] | VB NP PPDTV [0.00040] | VB NP PPTMP SBARPRP [0.00020] | VBZ NPPRD PPLOC [0.00119] | VBD NPEXT PPDIR [0.00416] | VBG NP SPRP [0.00079] | VBN NP PPCLR [0.01385] | VB ADVP [0.00040] | VBZ NPPRD ADVPLOC [0.00020] | VBP PPMNR [0.00040] | VBD PPDIR [0.00099] | VBP RB ADVPTMP VP [0.00020] | PRN VBN NP [0.00020] | VBN NPEXT [0.00040] | VB PRT NP [0.00158] | VBZ RB VP [0.00435] | VBG PPCLR PPCLR [0.00020] | VBD ADJPPRD PP [0.00079] | VBD S PPLOC [0.00059] | VBP NP SBARPRP [0.00040] | ADVPTMP VBN NP PPCLR [0.00040] | VB NP PRT [0.00040] | VBD ADVPTMP NPPRD [0.00040] | MD RB VP [0.00574] | VBD NPEXT SCLR [0.00059] | VBD PP PP [0.00020] | VBD NP PPTMP [0.00237] | VBP PPCLR [0.00336] | VBG PRT NP [0.00079] | VB PPCLR [0.00890] | VB NP PPTMP [0.00435] | VBD PPCLR 2 ADVPCLR [0.00079] | VBD ADVPCLR PPLOC [0.00040] | VBN NP PPLOCCLR [0.00198] | VBG NP PPCLR PPCLR [0.00020] | VBP ADJPPRD SBARADV [0.00020] | VBD NPEXT PPTMP [0.00020] | VBZ SBARNOMPRD [0.00020] | VBD ADJPPRD ADVPTMP [0.00040] | VBN S PPTMP [0.00020] | VBN PRT NP PP [0.00020] | VBZ CC VBZ NP [0.00059] | VBD NPCLR [0.00020] | VB VP [0.01801] | ADVP VBD S [0.00020] | VBN PPLOC [0.00020] | VBD PPTMPPRD [0.00040] | VBD S [0.01622] | VBN NP PPCLR SCLR [0.00059] | VB NP PPCLR NPTMP [0.00020] | VBN PPPRP [0.00040] | VBD NPEXT [0.00020] | VB ADVPMNR SBARADV [0.00020] | VBN NP PP PPCLR [0.00079] | JJ PPCLR [0.00020] | VBN PPCLR PPTMP [0.00020] | VBP PPLOC [0.00020] | VB SBAR [0.00475] | VBN NP

SBARADV [0.00040] | VB NP ADVPMNR [0.00257] | VBZ ADVP PPCLR [0.00040] | VB PPDIR PPPRP [0.00020] | VBD SPRP [0.00020] | MD ADVP VP [0.00336] | VBP NPPRD [0.00198] | VBD NP ADVP [0.00059] | VBD PPDIR PPDIR PPTMP [0.00020] | VBZ PPLOCPRD [0.00040] | VBD ADVPTMP VP [0.00079] | VBP ADVPMNR VP [0.00020] | VBD PPDIR ADVPTMP [0.00020] | VBN NP SBARCLR [0.00020] | ADVPMNR VBG NP [0.00040] | VBG NP PPTMP [0.00079] | VBD NPPRD 2 PP [0.00040] | NNS NP [0.00020] | VBN SBAR [0.00237] | VBZ RB ADJPPRD SBAR [0.00059] | VBN NP PPTMP 2 PP [0.00020] | VBD ADVPCLR PPCLR [0.00020] | VBN NP PPMNR PP [0.00020] | ADVPMNR VBG [0.00020] | VBD NP 2 PPTMP [0.00020] | VBD NPEXT ADVPTMP [0.00020] | VBN PPCLR PP [0.00020] | VB PPDIR PPDIR [0.00099] | VBP ADJPPRD [0.00653] | NN [0.00020] | VBN NP SBARTMP [0.00099] | VBP NP PPPRP [0.00020] | VBD PPDIR PPDIR [0.00198] | VBG ADVPDIR PP [0.00020] | VBN NP PPDIR [0.00040] | VBD PPCLR PP [0.00059] | VB PPDIR [0.00198] | VBG S [0.00673] | VBD NPEXT PPCLR [0.00178] | VBG NP ADVPMNR [0.00040] | VBG PPLOC [0.00059] | VBP S [0.00495] | ADVP VBN NP [0.00040] | VBD NP SBARPRP [0.00059] | ADVPMNR VBD NP [0.00040] | VBN NP PP PPPRP [0.00020] | VB NP ADVPTMP [0.00178] | VBN NP PPLOC 2 SADV [0.00020] | VBN NP NPTMP [0.00237] | VBD NP S [0.00099] | VBD NP ADVPTMP [0.00198] | NONE [0.00099] | VBN NP PPCLR PPTMP [0.00040] | VBP PP [0.00020] | VBD NP NP [0.00059] | VBD NP PPMNR [0.00020] | VBN PPCLR PPMNR [0.00040] | VBG ADVPDIR PPDIR [0.00040] | VBN NPPRD [0.00119] | VP 2 VP CC VP [0.00059] | VB PPCLR PPMNR [0.00040] | VB PPCLR PPLOC [0.00020] | VBN PRT PPCLR [0.00020] | VBZ NPPRD SBAR [0.00020] | VBZ RB ADVP VP [0.00119] | VBN NP [0.01801] | VBN CC VBN NP [0.00020] | VBZ PPDIR [0.00059] | VBN PPMNR [0.00020] | VBG PPCLR [0.00356] | VB NP 2 SBARADV [0.00059] | VBZ NPPRD ADVPTMP [0.00040] | VBD ADVPPRD PP [0.00059] | VBP PPPRDLOC [0.00020] | VBD NP PPDIR PPDIR [0.00020] | VBD NP [0.03324] | VBD NPTMP PPLOC [0.00040] | VB SPRD [0.00020] | VBZ ADVPMNR VP [0.00059] | VBD 66 ADJPPRD [0.00040]

| VBP RB NPPRD [0.00040] | VBN NP PPEXT [0.00020] | VBZ NP 2 SBARADV [0.00040] | VBN NP PPTMP PPCLR [0.00040] | VBN NP PRT [0.00059] | VB S NPTMP [0.00020] | VBD PPLOC SBAR [0.00059] | VBG PPTMP [0.00040] | VBP ADVPLOCPRD [0.00020] | VBG ADVP [0.00040] | VBD SPRD [0.00079] | VP 2 NPADV [0.00020] | VBZ RB ADVPTMP VP [0.00040] | VBZ NP [0.02256] | VBN NP PPMNR [0.00119] | VBD 2 66 S [0.00079] | VBZ ADJPPRD NPTMP [0.00040] | ADJPPRD PPLOC [0.00020] | VB SBARNOM [0.00040] | VBZ ADVPTMP VP [0.00079] | VBP NP S [0.00020] | VBZ SBAR [0.01306] | VBP ADJPPRD SBARPRP [0.00020] | VB PPLOCPRD SBARTMP [0.00040] | VBZ ADVPTMP NPPRD [0.00020] | VBP NP SPRP [0.00040] | VBZ PRN NP [0.00020] | VBZ ADVP NPPRD [0.00079] | VBZ 66 VP [0.00059] | VBD NP PPCLR [0.00633] | VBN NP PP 2 SADV [0.00020] | VBP NPCLR PPCLR [0.00020] | VBN NP ADVP [0.00020] | VBZ PPLOCCLR [0.00020] | VBD CC VBD NP [0.00020] | VBN [0.00119] | VBN ADJPPRD [0.00079] | VBP NP SCLR [0.00040] | VBD NP SBARTMP [0.00040] | VBG NP 5 SADV [0.00020] | VBP PPCLR ADVPTMP [0.00020] | VB PPCLR NPTMP [0.00020] | VBZ NP PPDIR [0.00040] | TO 66 VP [0.00040] | VBD ADVP VP [0.00119] | VB ADVPDIR [0.00040] | NN SBAR [0.00020] | VBD PPCLR SPRP [0.00020] | VBP RB VP [0.00554] | VBN ADVPMNR [0.00020] | VBN NP PPLOC PP [0.00020] | VBZ NP SBARTMP [0.00040] | VBP ADVP NPPRD [0.00020] | VBN NP PP [0.01108] | VBZ PPCLR [0.00277] | VBD RB NPPRD [0.00040] | VBD PPPRD [0.00040] | VB S [0.01207] | VBZ RB ADJPPRD 2 ADVP 2 SBAR [0.00040] | VBP NP PPMNR [0.00059] | VBG SPRP [0.00040] | VP 2 VP 2 CC VP [0.00020] | VBN NP ADVPTMP [0.00178] | VBZ ADJPPRD SBAR [0.00059] | VBZ PPMNR [0.00020] | VBD NP NPTMP [0.00158] | VBN NP 2 SADV [0.00059] | VBD ADJPPRD PPTMP 2 PP [0.00020] | VBP NP ADVPLOC [0.00020] | VB NPPRD PP [0.00040] | VBZ S 2 SBARADV [0.00040] | VB NP SPRP [0.00059] | JJ NP [0.00020] | VBG SADV [0.00020] | VBN NP PP PP [0.00020] | VBN NP 2 PP [0.00040] | VBN ADVPMNR VP [0.00020] | VBD ADVPMNR VP [0.00040] | VBN S [0.01108] | VBD ADJPPRD

SBAR [0.00020] | VBN PPTMP [0.00059] | VBD [0.00218] | VBP [0.00059] | VBZ SBARTMP [0.00040] | VBP PPCLR PPCLR [0.00040] | VBD NP 2 66 S [0.00020] | VBD RB VP [0.00693] | VB ADVPMNR [0.00020] | VP [0.00020] | VBP NP PPTMP [0.00020] | VBD PRT NP PPTMP [0.00020] | VB NP NPTMP [0.00059] | VBN PP [0.00059] | VBP ADVP VP [0.00040] | VB NP PPEXT [0.00040] | VB NP PPCLR PPMNR [0.00020] | VBZ PPTMP VP [0.00040] | VBG PPLOCCLR [0.00020] | VBZ RB ADJPPRD [0.00099] | VP 2 VP [0.00020] | VB 66 NPPRD [0.00020] | VP CC VP PPTMP [0.00040] | VB NP 2 SADV [0.00040] | VBZ NPPRD PPTMP [0.00059] | VB PPTMP [0.00158] | VBP NPTMP [0.00040] | VBP SBAR [0.00871] | VB NP SPRPCLR [0.00020] | MD VP [0.05639] | VBG NP NP [0.00040] | VB ADJPPRD SBARTMP [0.00040] | VBN ADVP [0.00040] | VBG NP PPLOC [0.00198] | VBZ SBAR 2 SADV [0.00020] | VBZ NP NPTMP [0.00040] | VBN SBARNOM [0.00020] | VB PPPRD PPTMP [0.00020] | VBZ 66 NPPRD [0.00020] | VB NP ADVP [0.00178] | VBD NP PPLOC [0.00119] | VBZ ADVP VP [0.00178] | VBZ NP PPMNR [0.00040] | VBN S 2 SADV [0.00040] | VBG NP NPTMP [0.00059] | VBN NP S [0.00040] | VBD SBAR [0.05065] | VBP VP [0.02216] | ADVPMNR VP CC VP PP [0.00059] | VBN NP PPDIR PPTMP [0.00040] | VBD NP PPPRP [0.00020] | VBN NP PPCLR PPPRP [0.00059] | VB PRT [0.00099] | MD ADVPTMP VP [0.00099] | VB ADVPMNR VP [0.00040] | VBZ NP PPLOC [0.00059] | VBD NPTMP SBAR [0.00099] | VB ADJPPRD PP [0.00040] | VBD RB ADJPPRD [0.00020] | RB VBG NP [0.00020] | VB NP SCLR [0.00020] | VBP NP PPLOC [0.00020] | VB ADJPPRD [0.00356] | NONE NP [0.00020] | VBZ 66 NP [0.00020] | VP 5 NP [0.00020] | NN PPCLR [0.00020] | VB NP PPDIRCLR [0.00020] | VB NP PP [0.00277] | ADVPMNR VB NP [0.00020] | VBN 66 ADJPPRD [0.00020] | VB NP PPLOC PPTMP [0.00040] | ADVP VBD PPDIR PPDIR [0.00020] | VB NP NPADV [0.00020] | VBP ADVPTMP [0.00040] | VB ADVPTMP [0.00020] | VBD PRT PP [0.00020] | NN NP [0.00040] | VBG ADJPPRD [0.00099] | VB ADVPPRD [0.00020] | VB PPLOCPRD [0.00020] | VB ADVPCLR [0.00020] | VB NP PPPUT [0.00040] | VBN NP PPCLR SBARTMP [0.00059] | VBD PPCLR 2 ADVP [0.00059] | VBD

NP PPDIR [0.00059] | VB ADVP ADVPTMP [0.00020] | VBD NP PPPUT [0.00040] | VBD NPPRD 2 ADVPCLR [0.00020] | VBP PPLOCPRD [0.00040] | VBD NP SPRP [0.00040] | VBZ PPPRD [0.00099] | VB NP PPPRP [0.00059] | VBN NP PP PPTMP [0.00020] | VBZ ADJPPRD S [0.00059] | VBZ ADJPPRD ADVPTMP [0.00040] | VBN NP PPCLR ADVPTMP [0.00020] | VBG NP 2 SADV [0.00059] | ADVP VBG NP [0.00040] | VBG NP PPCLR [0.00455] | VBD PPEXT PPTMP [0.00040] | VBZ NPPRD 2 SBARADV [0.00020] | VBN NP SBARMNR [0.00020] | VBG NP PP [0.00040] | VB NP PPLOC [0.00237] | VBZ NP ADVPTMP [0.00040] | VP CC VP [0.01642] | ADVPMNR VBN NP [0.00040] | VBD NP PPDTV [0.00040] | VBZ ADVPMNR PPCLR [0.00040] | VBZ NP NP [0.00040] | TO ADVP VP [0.00020] | VBN NP PPLOC SBAR [0.00020] | VBD RB ADVP VP [0.00079] | VBZ S SPRP [0.00020] | VBZ NP PP [0.00059] | VBD NP 2 SBARTMP [0.00020] | VBZ NP PPTMP [0.00079] | VBZ ADJPPRD PP [0.00040] | VBD SBAR 2 SADV [0.00020] | VBG NP ADVPTMP [0.00020] | VBD ADVP ADJPPRD [0.00020] | VBG ADVPMNR PP [0.00020] | VBZ VP [0.03463] | VB NP PPCLR ADVPMNR [0.00020] | VBD NPCLR PPCLR [0.00020] | VP 2 CC VP [0.00257] | VBP NP PP [0.00020] | VBZ RB NPPRD [0.00119] | VBZ RB PPPRD [0.00040] | VBP PPTMPCLR [0.00020] | ADVPMNR VBZ NP [0.00020] | VBG [0.00198] | VBN NP PPCLR PPLOC [0.00040] | VBN PPPRD [0.00040] | VB NP ADVPLOC [0.00059] | VBD PP [0.00178] | VBD PPTMP [0.00099] | VBD ADVP [0.00020] | VBG NPPRD [0.00059] | VB ADVPDIR NP [0.00020] | VB PRT NP SPRP [0.00040] | VBN PPDIR [0.00059] | VBG ADVPMNR [0.00040] | VBZ RB ADJPPRD S [0.00040] | VP CC RB VP [0.00020] | VBP ADJPPRD PP [0.00020] | VBD PPCLR PPTMP [0.00079] | VB S PPTMP [0.00020] | VB S SBARTMP [0.00020] | VBN NP SPRP [0.00040] | VB NP NP [0.00099] | VB PPCLR PPTMP [0.00059] | VBG NP ADVPCLR [0.00020] | VBP RB ADVP VP [0.00079] | VBD SBAR 2 PP [0.00059] | VB ADJPPRD SBARADV [0.00020] | VBZ NP PPCLR [0.00277] | VB NP S [0.00020] | VBG SBARNOM [0.00040] | VBN S SBARPRP [0.00040] | VBG NP SCLR [0.00040] | VBN NP SBAR [0.00020] | VBD NPEXT PPDIR

PPLOC [0.00020] | VBN NP NPTMP PPLOC [0.00020] | VBG ADVPMNR PPCLR [0.00040] | VBD ADVPPRD [0.00059] | VB NPPRD 2 SBARADV [0.00020] | VB PPPRD [0.00079] | VBD PPTMP S [0.00020] | VBN NP PPCLR VP [0.00040] | VB [0.00554] | VBN NP SBARPRP [0.00020] | VBD NPEXT 2 SBARADV [0.00020] | VB ADVPDIR PPDIR [0.00020] | VBZ RB NP [0.00059] | VP 2 NP [0.00040] | VBD PRT [0.00020] | VBZ ADVPCLR [0.00040] | VBZ 66 ADJPPRD [0.00020] | VBZ SPRD [0.00119] | VBZ [0.00079] | VBD NP 2 SADV [0.00119] | VB NPTMP [0.00139] | VBN ADVPDIR PPTMP [0.00040] | VBD PPCLR [0.00475] | VBN NP NP [0.00277] | VBP PPPRD [0.00099] | VBG PP [0.00119] | VBD NP SBAR [0.00099] | VBD NPPRD [0.00534] | VBD PRT NP PP [0.00040] | VBD ADVPMNR PPCLR [0.00020] | TO VP [0.08429] | VBP NP ADVPTMP [0.00059] | VBD SBARTMP [0.00059] | VBP NP NP [0.00020] | VBD ADVPCLR [0.00040] | VBD NP PP [0.00218] | VBG VP [0.00158] | VBG NPTMPCLR [0.00059] | VBN VP [0.00613] | VBZ RB NPPRD PPLOC [0.00020] | VBZ SBARPRD [0.00099] | VB ADVP PPTMP [0.00020] | VBZ ADJPPRD [0.00693] | VB NP PPCLR [0.00851] | VBZ ADJPPRD SBARPRP [0.00020] | VP 2 RB VP [0.00020] | VBD PPLOC [0.00020] | VBD NP SBARADV [0.00020] | VBD NP 2 SBARADV [0.00040] | VBD NP 2 PP [0.00099] | VBD NP PPTMP 2 ADVP [0.00040] | VB S SBARPRP [0.00040] | VBD NPEXT PPCLR SCLR [0.00020] | VBN PPCLR [0.00198] | VBD SBARPRD [0.00059] | VBD ADVPDIR [0.00040] | VBD NP PP PPTMP [0.00020] | VBZ NP SBARADV [0.00040] | VB NP SBARTMP [0.00040] | VBN NP PPTMP [0.00396] | VB NPTMPCLR [0.00020] | VBN NP PPPRP [0.00059] | VBD NPPRD 2 ADVP [0.00059] | VBG NP [0.02948] | VBG NP SBARTMP [0.00040] | IN S [0.00020] | VBN ADJPPRD PPTMP [0.00059] | NONE PPLOC [0.00020] | VBD NPPRD PPTMP [0.00040] | VBN PRT NP [0.00040] | VBN NP SCLR [0.00277] | VBD VP [0.02473] | PPTMP VBZ VP [0.00020] | VBZ S [0.01840] | VBN NP PPCLR PP [0.00079] | VBP RB ADJPPRD [0.00020] | VBP NP PPCLR [0.00178] | VBD NPEXT PPDIR PPDIR [0.00198] | VBG NPTMP [0.00020] | VB PPMNR [0.00020] | VBG CC VBG NP [0.00059] | VBZ PRT NP [0.00059] | VBP

ADVPLOC [0.00020] | VBG NP NPTMP PP [0.00020] | VBD NPTMPCLR [0.00099] | VBZ PP [0.00059] | VBP ADVP ADJPPRD [0.00040] | VBD NPEXT PPCLR PP [0.00059] | VBD ADVP PPLOC [0.00040] | VBD PRT S [0.00020] | VBN NP PPLOC PPTMP [0.00020] | VBP ADVPTMP VP [0.00158] | VBD PRT NP [0.00099] | VBN NP ADVPMNR [0.00119] | VBD PPLOCPRD [0.00040] | VB CC VB NP [0.00040] | VBG SBAR [0.00237] | VB NP 2 PP [0.00020] | VB NP SBARADV [0.00218] | VBG PPDIR [0.00079] | JJ [0.00020] | VBD PPCLR PPCLR [0.00020] | VBD S PPTMP [0.00020] | VB NPTMP SBARADV [0.00040] | VB PRT PPCLR [0.00059] | VBN PPDIR PPDIR [0.00040] | VB PPCLR PP [0.00020] | VBP NPPRD 2 SBARADV [0.00020] | ADVPMNR VB NP SBARADV [0.00020] | VB NPPRD [0.00356] | VBN NP PP PPTMP PP [0.00040] | VBN ADVP VP [0.00040] | MD 66 VP [0.00020] | VBN NP PPCLR 2 PP [0.00020] | VBN NP PPLOC [0.00455] | VB NP SBARPRP [0.00040] | VBD NP NPTMP ADVPTMP [0.00020] | VBZ ADVP ADJPPRD [0.00040] | VB ADVP VP [0.00040] | VB NP [0.05936] | VBD 5 66 S [0.00040] | VB PP [0.00059] | VBG NP PPMNR [0.00059] | VB NP PPDIR [0.00079] | VBD ADJP [0.00020] | ADVP VBD SBAR [0.00040] | ADVPMNR VBN NP PPCLR [0.00020] | VBD NPTMP [0.00059] | VB NP PPMNR [0.00257] | MD RB ADVP VP [0.00020] | NN S [0.00020] | VBZ NP 2 SADV [0.00040] | VBN NP ADVPLOC [0.00059] | VBG PRT [0.00059] | VB PPLOC [0.00099] | VBD SBAR PPTMP [0.00020] | VBD ADJPPRD [0.00594] | VBG NP SBARPRP [0.00040] | VBP NP [0.01464] | VBZ RB SPRD [0.00020] | VBZ NPPRD [0.01286]

WHADVP -> IN [0.01818] | WRB [0.50909] | WRB JJ [0.03636] | NONE [0.43636]

WHNP -> WP [0.23875] | IN [0.03114] | WRB JJ [0.00346] | WP4 NNS [0.00692] | WDT NN [0.00346] | WHNP PP [0.00346] | WDT [0.50173] | NONE [0.21107]

WHPP -> IN WHNP [1.00000]».