А
П-71 25
5/8

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР

Проблемная группа по экспериментальной
и прикладной лингвистике

ПРЕДВАРИТЕЛЬНЫЕ ПУБЛИКАЦИИ

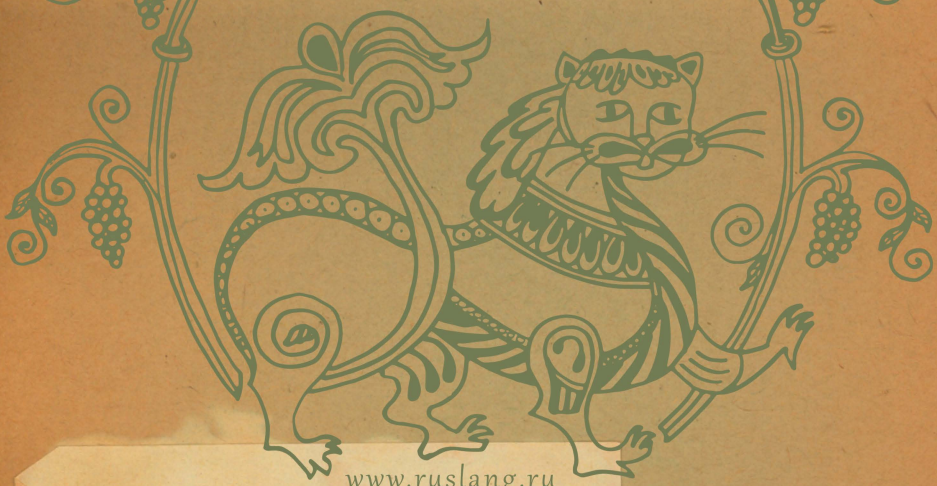
Выпуск 25

Б. В. СУХОТИН

ИССЛЕДОВАНИЕ СТРУКТУРЫ ПРОСТОГО ПРЕДЛОЖЕНИЯ С ПОМОЩЬЮ ЭВМ

Москва — 1972

ИРЯ
институт русского языка
им. в. в. виноградова



www.ruslang.ru

66467

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР

Проблемная группа по экспериментальной и
прикладной лингвистике

ПРЕДВАРИТЕЛЬНЫЕ ПУБЛИКАЦИИ

Выпуск 25

Б.В.Сухотин

ИССЛЕДОВАНИЕ СТРУКТУРЫ ПРОСТОГО ПРЕДЛОЖЕНИЯ С ПОМОЩЬЮ ЭГМ

Москва - 1972

4
П-7/25
БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР
66467

Введение.

В лингвистических работах прикладного характера, в особенности в работах, тяготеющих к формализации употребляемых понятий, нередко проводится разграничение выражений на "правильные" для данного языка и "неправильные" (по другой терминологии "допустимые" и "недопустимые", "отмеченные" и "неотмеченные"). При этом так или иначе происходит отсев "неправильных" выражений.

Лингвист, стоящий на более традиционных позициях, пользуется большей свободой. Так, он может сказать, что некоторое выражение более правильно (более литературно) чем другое, хотя подобное суждение покажется слишком неопределённым лингвисту-формализатору. Эта неопределённость, повидимому, является причиной того, что словари, алгоритмы машинного перевода и порождающие грамматики используют лишь две градации правильности.

Можно думать, что используя лишь двузначную оценку мы выигрываем в простоте описания. Это сомнительно. Отличие "очень правильных" выражений от "не очень правильных", полагаемых неправильными, может быть очень небольшим, и значит различающие правила должны быть очень тонкими, чтобы это различие уловить.

Кроме того, по самой сути дела точные "двоичные" правила не могут быть пригодными для определения явлений различных языков-ведь то, что допустимо для одного языка, в общем, недопустимо для другого.

Поэтому представляется достойной внимания другая стратегия в определении лингвистических явлений, которая основывается на следующих трёх принципах:

I. Мы отказываемся от попыток точно определить исследуемое явление с помощью бинарной системы оценок, при которой некоторый признак либо присутствует у этого явления, либо, отсутствует. Двоичную систему оценок мы будем использовать для определения гораздо более широкого обьекта: "области допустимых решений". Главное отличие этого объекта состоит в том, что он представляет собой множество лингвистических явлений в некотором смысле однотипных, принадлежащих не к одному языку, а к совокупности языков. Поэтому на долю двоичных правил остаются лишь весьма общие

и содержательно бедные свойства, которые, однако, отличаются стабильностью и существенностью.

2. В области допустимых решений вводится оценочная функция. Обычно мы задаем эту функцию на основе вероятностных соображений. Значения оценочной функции определяются в результате подсчётов по тексту на исследуемом языке, поэтому именно значения оценочной функции содержат специфику данного языка.

3. Определяемое явление данного языка есть элемент области допустимых решений, которому в данной системе значений оценочной функции соответствует максимум или минимум (т.е. экстремум, или "оптимум").

Алгоритмы, отыскивающие экстремум числовой функции на множестве другой природы, обычно называются оптимизационными. Они широко употребляются в математической экономии.

Чрезвычайно благодарной областью применения оптимизационных алгоритмов указанного вида являются задачи лингвистической дешифровки. Действительно, обнаружение элементов области допустимых решений обычно не составляет труда; вычисление значений оценочной функции и поиск её экстремума завершают определение данного явления.

Алгоритмы лингвистической дешифровки имеют ещё одну особенность, которую следует учитывать при оценке качества их работы: используемая ими информация должна быть по возможности скудной, поэтому "выход" удовлетворителен и тогда, когда он даёт значительный прирост знаний по сравнению с исходными, хотя сами по себе результаты могут показаться не особенно впечатляющими.

В качестве примеров лингвистических оптимизационных алгоритмов можно сослаться на работы (3,4). В данной статье развиваются соображения из (3), касающиеся синтаксиса.

Синтаксическая связь слов в простом предложении (двоичная характеристика).

Слова в текстах на "естественных" языках располагаются последовательно. Всеобщность этого факта, однако, никоим образом не свидетельствует о его существенности. Проще всего его можно объяснить тем, что человек обладает лишь одним генератором речи

и не может произносить два слова сразу. Этот факт был бы существенным, если бы такое "параллельное" генерирование было ненужным.

Нетрудно убедиться в обратном. Мы ощущаем иногда определённую близость слов, не расположенных рядом; в то же время, соседние слова порой не выглядят близкими. Дело в том, что мы интуитивно оцениваем близость слов, исходя из их связанности по смыслу, что, конечно, является очень существенным критерием. Большая степень связанности по смыслу лишь приблизительно отражается в текстуальной близости слов.

Как известно, смысловая близость обнаруживается простым приёмом: про испытуемую пару слов мы спрашиваем, образует ли она осмысленное сочетание; если да, то слова связаны по смыслу непосредственно, если нет, то значит, они удалены. Конечно, этот приём имеет свои недостатки: он не годится для испытания связанности слов, принадлежащих разным предложениям; в случае связанных слов испытуемая пара выглядит порой хотя и осмысленной, но не законченной и т.д., но для не слишком детального анализа его вполне хватает.

Столь же несложным образом можно обнаружить и направление смысловой связи. Если из фразы удалять по одному слову, сохраняя осмысленность фразы, то можно считать, что смысловая связь ориентирована к удаляемому слову от слова, связанного с ним, но ещё не удалённого. Подобным образом для предложения

Над омрачённым Петроградом
дышал ноябрь осенним хладом.

мы получим следующую схему смысловой связи:

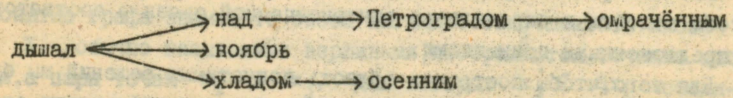


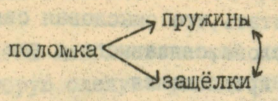
Схема (иначе, граф) смысловой связи играет важную роль в понимании текста. Каждый, кто пытался разобраться в тексте на мало знакомом языке (в особенности, на языке со сложным порядком слов, например, латинском), знает, как важно разобраться в связанности слов по смыслу.

Хорошо известно, что для правильного установления смысловой связи не обязательно обладать полным пониманием исследуемой фразы. Для этого достаточно знать принадлежность слов к так называ-

емым синтаксическим классам и правила, сообщающие, к каким синтаксическим классам должны принадлежать слова, связанные по смыслу. Таким образом, правила описывают "сочетаемостные" свойства синтаксических классов.

Схемы связи слов, установленные с помощью синтаксических классов и правил, в общем, отличаются от схем смысловой связи, поскольку они зависят от качества исследовательской работы лингвиста, выделившего эти классы и правила, от возможности опознавания принадлежности к тому или иному классу в тексте, и, конечно, от возможности (зависящей от языка) предложить хорошие классы и правила. Связь слов, установленную с помощью синтаксических классов и правил мы будем называть (в согласии с обычной практикой) синтаксической связью.

Если рассмотреть схему синтаксической связи простого предложения, то обнаруживается её своеобразное строение: в ней нет замкнутых контуров (циклов). Наличие циклов говорило бы о неоднозначности понимания. Например, фраза "поломка пружины защёлки" может быть понята двояко, если не только защёлка имеет пружину, но и пружина - защёлку. Этому соответствует следующая схема:



Отсутствие циклов в схеме простого предложения является весьма общим свойством синтаксической связи.

Все эти факты хорошо известны. Их полезно было напомнить, чтобы подойти к определению синтаксической связи в соответствии с предложенными принципами.

Для того, чтобы построить область допустимых решений, мы будем исключать из определения синтаксической связи информацию, специфичную для отдельного языка. Прежде всего мы исключаем информацию о сочетаемостных свойствах синтаксических классов, т.е. правила типа "существительное в именительном падеже подчиняет прилагательное в именительном падеже". Как мы увидим несколько позже, сочетаемость лингвистических единиц можно с большой степенью единообразия описывать вероятностными средствами.

Далее следует исключить и какое бы то ни было упоминание о классах, поскольку набор синтаксических классов специфичен для данного языка. Конечно, это не означает, что мы выбрасываем понятие класса из определения синтаксической связи. Речь идёт лишь об исключении его из определения области допустимых решений, т.е. "квинтэссенции" понятия синтаксической связи.

Таким образом, мы можем опереться лишь на свойства схемы синтаксической связи. Мы уже упоминали одно из этих свойств - отсутствие циклов. Ниже мы дадим определения некоторых математических объектов, безусловно, знакомых читателю, с целью устранения терминологической несогласованности.

Отображение. Пусть X и Y два данных множества. Закон согласно которому каждому элементу x ($x \in X$) ставится в соответствие элемент y ($y \in Y$), называется однозначным отображением X в Y или функцией, определённой на X и принимающей значения в Y. Многозначное отображение S множества X в множество Y есть закон, по которому каждому элементу x ($x \in X$) ставится в соответствие некоторое подмножество Sx ($Sx \subset Y$); при этом не исключается возможность $Sx = \emptyset$. В дальнейшем термином "отображение" как правило, будет именоваться многозначное отображение.

Граф. Граф считается заданным, если указаны

1. Непустое множество X;
2. Отображение S множества X в .

Понятие графа знаменито своей наглядностью. Элементы множества X, обычно называемые вершинами, изображаются точками плоскости, а пары точек x_i и x_j, для которых $x_j \in Sx_i$ соединяются стрелками, направленными от x_i к x_j. При этом вершину x_i назовём началом, а вершину x_j - концом стрелки x_i, x_j. Две стрелки назовём смежными, если они имеют общую граничную точку. Начало и конец стрелки называются её граничными точками.

Путём в графе называется последовательность стрелок $s_1, s_2, \ldots\ldots s_{k-1}, s_k$ в которой любые две стрелки, расположенные друг за другом, смежны. Если при этом конец предыдущей стрелки является началом следующей, то путь будет назы-

ваться ориентированным.

Циклом мы будем называть путь, в котором последняя стрелка смежна с первой. Граф, между любыми двумя вершинами которого имеется путь, называется связным. Связный граф без циклов называется деревом. По другому дерево можно определить как граф, между любыми двумя вершинами которого имеется единственный путь. Наконец, ориентированным деревом или "прадеревом" мы назовём дерево со следующими свойствами: оно имеет выделенную вершину, именуемую корнем дерева; существует ориентированный путь от выделенной вершины к любой другой вершине этого дерева.

Теперь мы готовы к тому, чтобы описать схему синтаксической связи предложения с помощью понятия "граф".

а) Множество X. В качестве множества X естественно взять множество слов, составляющих предложение. Но что такое множество слов? Если, как обычно, считать словом последовательность букв между двумя пробелами и не различать последовательности, совпадающие при наложении, то в предложении "скалы громоздятся на скалы" имеются три слова, хотя естественно полагать, что в этом предложении их четыре. Поэтому в качестве элемента множества X мы будем использовать более сложное понятие, т. наз. "вхождение слова". Вхождение слова - это пара: {слово, номер слова} (от начала текста или от начала простого предложения).

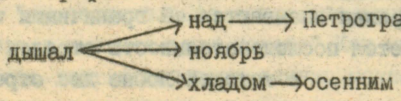
Итак, множество X - это множество вхождений слов, составляющих предложение.

б) Любое отображение множества X в себя, образующее дерево, есть синтаксическая связь слов в данном предложении.

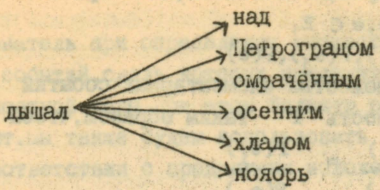
В силу этого определения следует признать изображением синтаксической связи в предложении

Над омрачённым Петроградом
Дышал ноябрь осенним хладом

не только граф



но, например, и такой граф:



Объясняется это кажущаяся неразумие тем, что, по сути дела, мы описали лишь "допустимый вид" синтаксической связи, но не единственное её изображение.

Действительно, в соответствии с принципом I мы описали допустимое решение в задаче об определении синтаксической связи, причём описали с помощью "логических", бинарных признаков. Таким образом, множество допустимых решений при нахождении синтаксической связи слов в данном предложении состоит из всевозможных деревьев, построенных на множестве вхождений слов этого предложения.

Дальнейшие наши действия будут состоять в том, чтобы отыскивать в этом множестве наилучшее, т.е. правильное дерево.

Вероятностная интерпретация синтаксической связи.

Введение функции качества на множестве допустимых решений - допустимых синтаксических связей - основывается, в конечном счёте, на естественном, хотя и смутном представлении о неодинаковости степени или силы этой связи. Возможность привлечения вероятностных соображений для оценки силы синтаксической связи кажется более или менее очевидной.

Напомним основные понятия. Множеством элементарных событий ε_i называется некоторое произвольное множество E ($E \ni \varepsilon_i$), а событиями e - его подмножества. Вероятностная мера (вероятность) есть числовая функция P, определенная на событиях и обладающая следующими свойствами:

1) $P(e) = 0$ тогда и только тогда, когда $e = \emptyset$;

2) $0 \leq P(e) \leq 1$ для любого e;

3) $P(e_i \cup e_j) = P(e_i) + P(e_j)$ тогда и только тогда, когда $e_i \cap e_j = \emptyset$.

Из этих определений следует, в частности, что мера всего равна I. Пусть $e_i \subset e, e_j \subset e, e \subset E$, и определены значения $P(e_i), P(e_j), P(e)$.

Если взять в качестве множества элементарных событий e и определить на нём вероятность P' таким образом, чтобы соблюдалась пропорция

$$\frac{P'(e_i)}{P'(e_j)} = \frac{P(e_i)}{P(e_j)}$$

для любых e_i и e_j, входящих в e, то произойдёт следующее: значения $P'(e_i)$ будут больше соответствующих значений $P(e_i)$ в одно и то же число раз k. Это число легко определить, взяв в качестве e_i множество e: $P'(e) = k \cdot P(e)$; $P'(e) = 1$; $k = 1/P(e)$.

Поэтому значение $P'(e_i)$ $(e_i \subset e)$ определяется следующим образом:

$$P'(e_i) = \frac{P(e_i)}{P(e)},$$

или, если $e_i \not\subset e$, то $P(e_i \cap e)/P(e)$.

Мы видим, что для вычисления вероятности $P'(e_i)$ используется и некоторая информация о другом событии, e. Вероятность $P'(e_i)$, определённая с помощью события e называется условной вероятностью события e_i при условии e. Трудно представить более естественный инструмент для определения степени связи двух событий, чем условная вероятность. Обозначается условная вероятность события e_i при условии e выражением $P(e_i/e)$. "Неразрывности" связи событий e_i и e соответствует $P(e_i/e) = 1$ это значит, что события e_i и e состоят из одних и тех же элементарных событий; $P(e_i/e) = 0$ говорит о полном различии событий e_i и e. Например, если e есть множество облачных дней в году, а e_i множество дней с осадками за тот же период, то $P(e_i/e)$ показывает долю дней с осадками относительно всех облачных дней. (8).

Эта теоретически ясная схема может применяться весьма по разному и давать различные результаты даже при изучении одного и того же явления. Всё дело в большой свободе, которой пользуется

исследователь при определении множества элементарных событий, выборе событий, связь которых подлежит изучению, и определении вероятностной меры. При этом большую роль играют интуиция и эксперимент. Мы также будем использовать оба эти фактора.

В соответствии с предыдущим изложением, текст T есть множество вхождений i_{xy} т.е. пар вида v_x, n_y, где v_x принадлежит алфавиту слов, а n_y есть номер слова от начала текста (предполагается, что слова следуют друг за другом).

Допустим, проведен синтаксический анализ, т.е. определено отображение GT или, другими словами, построен граф синтаксической связи на всем тексте. Тем самым, считается заданным множество пар вхождений вида

$$i_{xy} \longrightarrow i_{rt},$$

где вхождение i_{xy} отображается во вхождение i_{rt}. Такую пару мы будем называть стрелкой s_{xyrt}.

Появление (выбор) стрелки s будет считаться элементарным событием. Вероятности всех стрелок мы будем считать равными друг другу; если GT содержит N стрелок, то вероятность одной стрелки будет равна

$$P(s_{xyrt}) = \frac{1}{N}.$$

Особый интерес представляют некоторые неэлементарные события. Они получаются из элементарных с помощью теоретико-множественного сложения.

1. Верхнее вхождение слова v или u(v). u(v) есть множество стрелок вида s_{vxyz}, где v - фиксированный символ, x, y, z - переменные.
2. Нижнее вхождение слова w или l(w). l(w) определяется как множество стрелок вида s_{xywz}, где фиксирован символ w.
3. Обобщённая стрелка s_{vw} есть пресечение u(v) и l(w).

Степень (силу) синтаксической связи слов v и w можно характеризовать условной вероятностью

$$P(u(v))/l(w)) = \frac{P(u(v) \cap l(w))}{P(l(w))}$$

т.е. вероятностью верхнего вхождения v при условии появления

нижнего вхождения w в той же обобщённой стрелке. Смысл этой условной вероятности в том, что зависимое слово "сильно нуждается" в управляющем, тем сильнее, чем теснее эта зависимость. Вто же время величина условной вероятности

$$P(1(w)/u(v)) = \frac{P(u(v) \cap 1(w))}{P(u(v))}$$

в меньшей степени характеризует связь слов v и w поскольку главное слово может обходится без подчинённого, а может и иметь его.

Осветив таким образом вопрос о силе (степени) синтаксической связи, мы можем перейти к определению понятия "наилучшая синтаксическая связь" (наилучшее дерево).

Очень естественно считать, что хорошая связь - это сильная связь и наилучшее дерево предложения есть то, стрелки которого характеризуются наибольшей силой связи. Если сила связи стрелки s_{xyrt} обозначается через $q(s_{xyrt})$, то функция, оценивающая качество дерева будет выражаться следующим образом:

$$Q = \sum_{y} \sum_{t} q(s_{xyrt}),$$

где суммирование ведётся по всем стрелкам дерева. Тем самым, оказывается, что цель процедуры состоит в максимизации функции Q.

Функция Q (т.е. просто сумма условных вероятностей) сама по себе лишена вероятностного смысла, т.к. она может быть больше единицы. По этой причине лучше средняя условная вероятность

$$Q' = Q/n$$

где n - количество стрелок в предложении. Однако оптимизационно Q' и Q совпадают, поскольку их максимум имеет место для одного и того же дерева. Дело в том, что n - величина постоянная; она на единицу меньше числа вершин в дереве (числа слов в предложении). Поэтому максимум Q' обеспечивается максимумом Q. Вычислительно же функция Q проще, так как не содержит ненужного деления.

Прежде чем приступить к обсуждению процедуры, максимизирующей Q нужно обсудить проблему вычисления условных вероятностей

вида $P(u(v)/l(w))$. Чтобы оценить непростоту и неочевидность этой проблемы следует вспомнить, что предыдущие рассуждения относились к тому случаю, когда отображение GT уже известно.

Если мы не найдём способа оценить эти вероятности без нахождения GT то все эти рассуждения окажутся бесплодными. Далее мы попытаемся предложить несколько способов такой оценки и сделаем некоторые выводы.

Основная гипотеза методов оценки состоит в том, что можно, повидимому предложить легко определимое отображение $G'T$ и с его помощью вычислить условные вероятности, значения которых в некотором смысле подобны значениям условных вероятностей, вычисленных в случае известности отображения GT.

Конечно, выражение "в некотором смысле подобны" нельзя считать достаточно ясным. Однако, добиваться полной ясности в этом вопросе нет надобности, поскольку нам не придётся определять степень подобия (ведь истинные условные вероятности нам не известны). Критерий подобия может иметь значение в методах, использующих сходимость значений условных вероятностей к некоторой "идеальной серии значений".

Одно из наиболее простых отображений, используемых для приближённой оценки истинных условных вероятностей (назовём его G_2T) содержит все стрелки вида s_{xryr+1} и s_{xtyt-1}, где x и y - произвольные слова, r - произвольный непоследний, t - произвольный непервый номер слова в тексте. Условные вероятности для отображения G_2T вычисляются обычным образом, т.е.

$$P(u(v)/l(w)) = \frac{P(u(v) \cap l(w))}{P(l(w))}$$

Использование этого отображения равносильно предположению о том, что сильно связанные между собой слова обычно располагаются в непосредственной близости друг к другу в тексте.

Другое, тоже довольно естественное предположение состоит в том, что сильно связанные слова сильно предсказывают появление друг друга в пределах простого предложения. Эта гипотеза соответствует

отображению G_1T, содержащему стрелки вида s_{vrwt}, где вхождения i_{vr} и i_{wt} принадлежат к одному и тому же простому предложению. Для определения этого отображения, следовательно и соответствующих условных вероятностей, необходимо заранее определить границы простых предложений.

Экспериментальная часть настоящей работы будет состоять именно в проверке этих двух гипотез, оценки результатов с последующим обсуждением и выдвижением новых предположений. Может возникнуть вопрос: зачем нужно было вводить представление об истинных условных вероятностях и об их приближённом вычислении с помощью G_1T и G_2T? Вернее, в чём смысл выделения отображений G_1T и G_2T как особых объектов? Ведь их можно считать лишь средством для вычисления значений функции качества. На это можно ответить следующее.

Представляется вполне возможным описать класс простых отображений типа $G'T$. В то же время функция качества полностью определяется заданным отображением. Поэтому можно поставить вопрос об отыскании наилучшего "вспомогательного" или "проектирующего отображения GT_{opt} - наилучшего в том смысле, что оно даёт условные вероятности, наиболее согласующиеся с истинными. Это отображение можно будет оценить либо величиной вероятностей, определяющих силу синтаксической связи, которая должна быть по возможности большой для синтаксически связанных слов, либо из каких-нибудь "внутренних" соображений.

Кроме того, вопрос о способе проектирования дерева синтаксической связи нельзя считать лишённым самостоятельного лингвистического интереса.

Роль синтаксических классов.

Содержание предыдущей главы будет неполным и даже неверным если не осветить роли синтаксических классов для определения синтаксической связи.

Выше мы описали текст как множество вхождений слов, т.е. пар вида v_i, n_j, где v_i принадлежит алфавиту слов, а n_j есть номер от начала текста. Для целей синтаксического анализа такое представление неудовлетворительно. Нетрудно догадаться,

что при таком изображении текста условные вероятности $P(u(v)/l(w))$ получаются очень малыми по величине и ненадёжными (если текст не слишком велик),а кроме того (и это ещё важнее) - не соответствующими существу дела.Они,например,будут велики для слов,не всегда связанных синтаксически,но встречающихся в одном контексте,например:

попался,как ..ур во ..щи.

Здесь редкое слово сильно прогнозирует своё окружение. Подобная ситуация будет иметь место,конечно,не только для фразеологизмов,но и для любых слов,часто встречающихся в данном тексте.

Синтаксические классы,как мы уже говорили,специфичны для данного языка и поэтому не могут быть включены (в конкретном виде) в определение множества допустимых решений для синтаксического анализа.Поэтому их можно считать лишь средством для получения хорошего значения оценочной функции.В этом смысле их роль аналогична роли отображений G'T, цель которых - обеспечить получение приближённых значений условных вероятностей.Оказывается, между этими понятиями есть не только аналогия,но и глубокая связь

Хорошо известно,что есть языки богатые синтаксическими классами - русский,латинский,греческий,санскрит и другие,т.наз. синтетические языки,и языки,бедные этими классами,такие,как английский,китайский,датский (аналитические).Это различие очень заметно проявляется в категории падежа.В синтетических языках существует определённое количество падежных форм имени (существительного и прилагательного).Разные падежи одного и того же слова нельзя относить к одному и тому же классу:они не обладают взаимной заменимостью и "играют различную синтаксическую роль" - выступают различными членами предложения.Различия в личных формах глагола в пределах простого предложения менее существенны - они не влияют на роль слов в предложении.Наличие категории рода и числа существенно:обнаружение связи прилагательного и существительного облегчается,если можно установить совпадение их родов (чисел)

Аналитические языки обладают более твёрдым порядком слов.Это

значит, что для них необходимо определение отображения $G'T$ более специфического вида, чем для синтетических языков. Например, в случае синтетических языков можно не делать различия между стрелками s_{vw} и s_{wv} из $G'T$, но для аналитических языков порядок расположения слов в тексте может иметь очень большое значение. При этом сопоставление стрелке s_{vw} в дереве предложения того или иного веса должно зависеть от порядка расположения слов v и w в тексте.

В настоящей работе мы не ставим себе целью подробно исследовать этот важный вопрос и ограничиваемся его постановкой. Конкретные эксперименты мы ведём на русском, т.е. синтетическом языке; поэтому мы применяем отображения $G'T$ очень простого вида; впрочем, предлагаемые синтаксические классы тоже введены в небольшом числе, для получения приближённых результатов.

Полезно разобраться, какова формальная природа синтаксических классов. Вполне естественно определять их в соответствии с предложенными принципами, т.е. описывая множество допустимых решений и определяя функцию качества. В определении множества допустимых решений (наборов синтаксических классов) основной вопрос состоит в следующем: являются ли синтаксические классы пересекающимися множествами слов, или они образуют разбиение алфавита слов (словоформ)?

Для этого надо ясно представить наши потребности. Обычные алгоритмы анализа опираются опираются на сочетаемостные свойства слов, зависящие от описания слов в терминах синтаксических признаков (категорий). Категории образуют систему пересекающихся подмножеств алфавита слов. Однако, в предлагаемых алгоритмах не используются правила, зависящие от сочетаемостных свойств категорий. В то же время классы слов, используемые в наших алгоритмах должны совпадать с классами слов, эквивалентных с точки зрения их описаний в терминах категорий. Например, некоторый класс может содержать слова, описываемые как Существительное, им. падежа, мужск. рода, Ед. числа. Иными словами, можно сказать, что нам нужны классы синтаксически взаимозаменимых слов, в смысле (9).

Простота множества непересекающихся синтаксических классов

свойство очень ценное и с точки зрения дешифровочной стратегии. Повидимому, наилучшее разбиение найти легче, чем трудно определимое множество категорий. Кроме того, обладая множеством синтаксических классов, мы можем с большей легкостью приступить к нахождению синтаксических категорий.

Вопрос о качестве выбранного множества синтаксических классов мы оставляем открыты. Предлагаемый набор синтаксических классов не носит принципиального характера. Он содержит следующие классы (рядом дана буквенная и цифровая кодировка):

Именительный падеж существительного	ИС	I
Родительный " "	РС	2
Дательный " "	ДС	3
Винительный " "	ВС	4
Творительный " "	ТС	5
Предложный " "	ПС	6
Глагол в личной форме	Г	7
Предлог	П	8
Деепричастие	Д	9
Наречие	Н	I0
Именительный падеж прилагательного	ИП	II
Родительный " "	РП	I2
Дательный " "	ДП	I3
Винительный " "	ВП	I4
Творительный " "	ТП	I5
Предложный " "	ПП	I6
Инфинитив	ИН	I7
Союз	С	I8
Частица	Ч	I9

Местоимение распределено по различным классам в зависимости от его синтаксической роли.

Оптимизационная процедура.

Оптимизирующий алгоритм излагается согласно

Алгоритмы для нахождения дерева с максимальной или минимальной суммой весов стрелок носят название алгоритмов поиска максимального (минимального) стягивающего графа и неоднократно переоткры-

www.rustang.ru

66467

БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР

вались.

Наиболе простой из этих алгоритмов можно изложить следующим образом: до начала процедуры все вершины графа считаются изолированными. Дальнейшие действия таковы:

1) Стрелкой s_{vw} соединяются такие две вершины v и w, для которых условная вероятность P(u(v)/1(w)) максимальна. Вершины v и w объявляются выбранными, остальные - невыбранными.

2) Проверяется наличие невыбранных вершин. Если их нет, то процедура заканчивается.

3) Рассматриваются стрелки вида s_{rx} или s_{xr}, где r - выбранная вершина, x невыбранная. Среди стрелок такого вида выбирается стрелка с наибольшей условной вероятностью P(u(r)/1(x)) или P(u(x)/1(r)). Вершина x объявляется выбранной.

4) Выполняется пункт 2).

В наших экспериментах этот алгоритм дополняется предварительным вычислением условных вероятностей. Подробнее об этом рассказывается в главе "Программы".

Описание и оценка экспериментов.

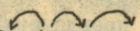
Для экспериментов были выбраны следующие тексты: А.С.Пушкин, "Пиковая дама" (целиком) и "Выстрел" (начало), объёмом около 10000 символов. Таблицы условных вероятностей подсчитывались по всем этим предложениям. Примеры выбирались из анализов первых 1500 предложений. Анализы первых 200 предложений помещены в приложении.

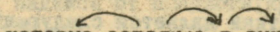
Нетрудно видеть, что результаты экспериментов в первом и во втором случае в значительной степени противоположны. Отображение G_2T даёт анализы, в которых почти все слова непосредственно подчинены глаголу. Это, в общем, неудивительно, если принять во внимание, что почти в каждом предложении кроме слова произвольного класса есть ещё и глагол. Вот некоторые исключения из этого правила:

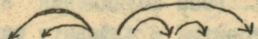
$$P(П/ПС) > P(Г/ПС) \text{ и } P(П/ПП) > P(Г/ПП),$$

что даёт правильные анализы предложений типа ИС, Г, ЗС, ДС, П, ПС (в этом наборе могут быть пропущены какие-нибудь падежи

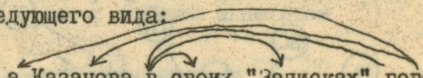
существительного. ПС пропускается вместе с предлогом). Например:

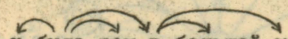






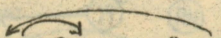
При этом, однако $P(П/ИП) > P(ПС/ПП)$
что даёт графы следующего вида:

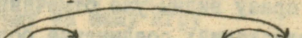




Далее, $P(ИС/ИП) > P(Г/ИП)$, $P(РС/РП) > P(Г/РП)$ и $P(ДС/ДП) > P(Г/ДП)$.

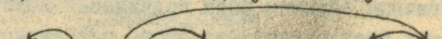
Это также даёт ряд правильных анализов:





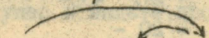




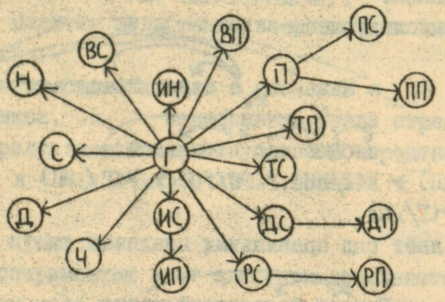




Почти во всех остальных случаях отношение подчинения непосредственно подчиняет зависимое слово глаголу. Если предложение именное, то корнем дерева является ИС. В таких предложениях возможно правильное отражение зависимости ВС → ВП и ТС → ТП, потому, что $P(ВС/ВП) > P(ИС/ВП)$ и $P(ТС/ТП) > P(ИС/ТП)$:

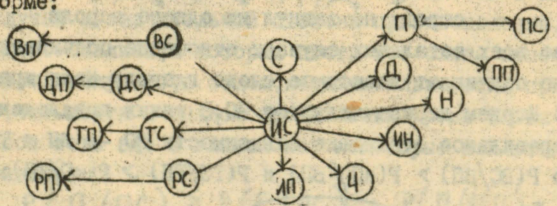


Отображение G'T, набор синтаксических классов, таблица условных вероятностей и оптимизирующий алгоритм представляют собой некоторую систему анализа предложений данного языка, т.е. по сути дела, его грамматику. Очень наглядной и полезной характеристикой данной грамматики является следующий "граф-характеристика" или "роза подчинения". "Роза подчинения" получается применением оптимизирующего алгоритма к алфавиту синтаксических классов. Отображение G_1T дает следующую розу подчинения:



На этом рисунке сразу видны уже отмеченные особенности в анализе простых предложений: соединение большинства слов непосредственно с глаголом; подчинение предложных падежей существительного и прилагательного предлогу; правильное подчинение именительного, родительного и дательного падежей прилагательного соответствующим существительным.

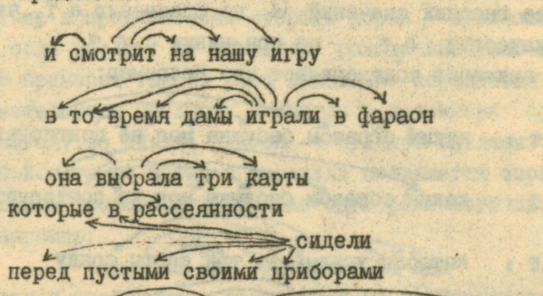
Аналогичным образом роза подчинения может быть получена для именных предложений. Для этого оптимизирующий алгоритм следует применить к алфавиту синтаксических классов без глагола в личной форме:



Этот рисунок неплохо отражает структуру именных предложений. Ошибки анализа можно подытожить следующим образом: имеет место
I)отрыв предложных падежей от предлога(кроме собственно предложного падежа).Например:



Таких примеров очень много.
2)Неправильное соотнесение прилагательных в винительном и творительном падежах непосредственно с глаголом,а в предложном падеже - с предлогом:



3)В предложениях с "вторичным подчинением",т.е.с подчинением инфинитиву,деепричастию,прилагательному или существительному, происходит связывание "вторично подчинённых"слов непосредственно с главным словом.Анализы такого рода имеют особенно нелепый вид





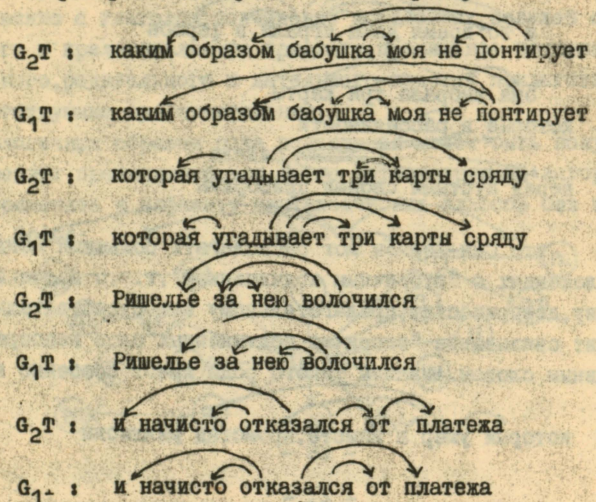
(и много других).

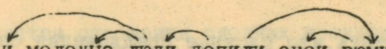
Оценивая работу этого алгоритма следует заметить, что он различает лишь самые грубые закономерности. Это и не удивительно: он не только использует весьма бедный набор классов, но и крайне мало содержательное отображение G_1T, в котором не используется даже признак непосредственной близости слов.

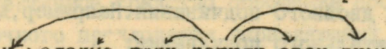
Отображение G_2T.

Уже беглый взгляд на таблицу T_2, соответствующую G_2T, позволяет отметить её важное отличие от T_I - наличие пустых или почти пустых клеток, наряду с клетками, содержащими очень большие числа. Эта большая контрастирующая способность обеспечивает получение более высоких значений Q по сравнению с T_I, что говорит о лучшем качестве G_2T по сравнению с G_1T.

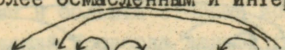
Анализ примеров подтверждает это. Например:

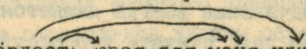


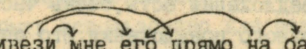
G_2T : 

G_1T : 

В этих примерах мы видим значительное улучшение анализа связи прилагательное ← существительное и глагол —→ предлог → существительное. В некоторых случаях новый анализ, также ошибочный, выглядит гораздо более осмысленным и интересным:



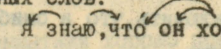




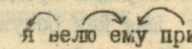
Известно большое логическое сходство между предлогом и глаголом, в некоторых языках они очень похожи и грамматически (китайский, вьетнамский). Несомненно, что отражением этого факта является появление в предложениях второго корня - предложного - в том случае, если предлог не управляет родительным или дательным падежом. В примерах такого рода правильно передается подчинение существительного предлогу, а иногда и подчинение существительного глаголу, но связь глагол → предлог выглядит сравнительно слабой. В то же время в деревьях появляется характернейшая формальная особенность - наличие одновременного двойного подчинения, например:



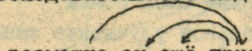
Эта формальная особенность, совершенно не свойственная правильным графам простого предложения, чрезвычайно характерна для предложений сложных, в особенности в языках, избегающих употребления соотносительных слов:



Подобный разбор может считаться правильным и для предложений с вторичным подчинением:



Следует заметить,что сложные предложения соединяются не только способом двойного подчинения.Например,для них допустим способ последовательного подчинения:

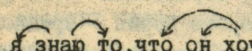


Общая часть двух предложений может быть управляющей,и это не приводит к двойному подчинению (однородность):

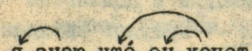


Однако,двойное подчинение всегда является признаком сложности предложения.Можно даже считать,что отсутствие двойного подчинения в сложном предложении есть,очень часто,результат некоторого опрощения сложного предложения.Это опрощение может происходит следующими способами:

I)введение соотносительных слов.В этом случае предложения-компоненты приобретают разъединённый вид:



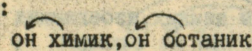
2)отнесение дважды подчинённого слова к одному из предложений-компонентов:



3)Подчинение корня придаточного предложения управляющему слову главного:



4)повторение общей части предложения с однородными управляющими группами:



Этот обзор позволяет заключить,что в ошибках типа

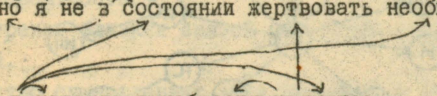


сказывается ложность соответствующих предложений,или,точнее, не улавливается простота,возникшая как результат опрощения сложного предложения.Очевидно,что ошибки такого рода можно ис-

править, введя средства опрощения сложных предложений в систему анализа. Очень эффективное средство опрощения состоит в подчинении корня придаточного предложения управляющему слову главного. Осуществление этого приёма достигается весьма просто: запретом двойного подчинения. Иначе говоря, мы уточняем понятие дерева синтаксической связи, вводя следующее определение элемента множества допустимых решений: граф простого предложения есть прадерево. Это определение в большей степени использует факт простоты разбираемых предложений. Проверке данной гипотезы посвящён третий эксперимент, обсуждение которого приводится далее.

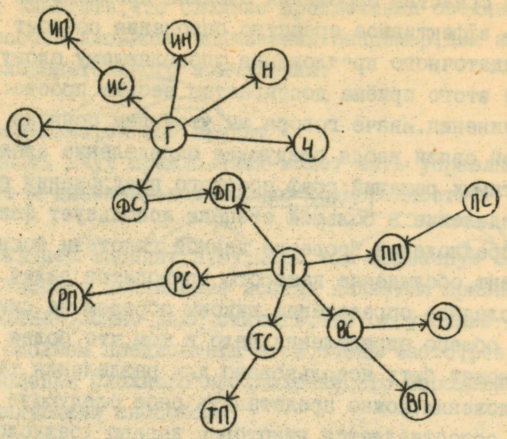
Однако, последнее определение никоим образом не снижает интереса более общего определения. Дело в том, что более общее определение может быть использовано для разделения текста на простые предложения. Можно представить себе следующую стратегию: всему тексту сопоставляется некоторое дерево (безусловно, обладающее большим количеством неоднократно подчинённых вершин), а затем над ним проводятся процедуры опрощения типа перечисленных выше.

Возвращаясь к обсуждению эксперимента с T_2, следует отметить столь же запутанные и неправильные анализы обширных предложений, как и в случае с T_2, например:

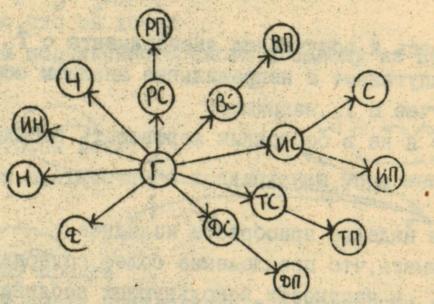


Можно думать, что предложение более содержательных отображений G'T, в частности, использующих неодинаковость прогнозирования в зависимости от порядка расположения слов в тексте, позволит значительно улучшить и анализы простых предложений большой длины.

Особенности грамматики хорошо представляются розами подчинения. В первом рисунке хорошо видна многовершинность предложений с предлогами; второй рисунок ясно подчёркивает правильность анализа связи прилагательное ← существительное.



То же самое, без предложной группы:



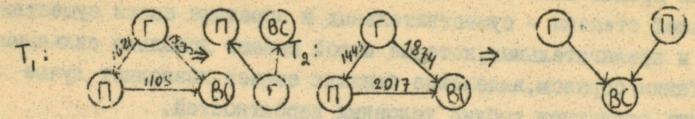
Близкие и далёкие связи.

Сравнивая результаты анализов, полученных с помощью T_1 и T_2 можно заметить, что во втором случае хорошо передаются связи слов, близко расположенных в тексте: существительных и прилагательных, предлогов и существительных; в первом случае, напротив, лучше отражаются связи удалённых слов: глагол -- союз, или глагол -- предлог.

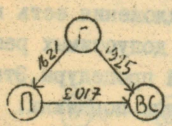
Можно попытаться скомбинировать обе таблицы, построив новую таблицу $T' = \| p'_{ij} \|$; $p'_{ij} = \max(p_{ij}, p'_{ij})$, где p_{ij}

есть элемент T_I ($T_1 = \| p_{ij} \|$, а p'_{ij} есть элемент T_2.

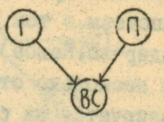
Рассмотрим, как при этом будет анализироваться одна из основных конструкций: Г, П, ВС.



Вычислим значения p'_{ij} в соответствии с предложенным правилом:



что даёт



Как видим, новое правило не улучшает анализа, проведённого в соответствии с требованиями второго эксперимента. Однако, сравнение чисел p_{ij} и p'_{ij} позволяет формально описать интересные понятия "близкой" и "далёкой" связи.

Близкой связью назовём пару v,w (v,w -элементы алфавита синтаксических классов) в том случае, если $p'_{vw} > p_{vw}$, в противном случае эту пару назовём далёкой связью. В T_2 все близкие связи выделены кружками, окружающими соответствующие условные вероятности.

Рассмотрение близких связей позволяет сделать ряд любопытных наблюдений.

I) Большинство существенных (т.е. входящих в деревья предложений) связей - близкие. Таковы связи ИС ← Г, все связи существительное → прилагательное, все связи существительное ← предлог, связь Н ← Г.

К далёким существенным связям относятся связи падежей существительного и глагола (кроме ИС и ДС), связь глагола и предлога, Д ← Г.

2) Близкие связи обычно сильные (т.е. отличаются высокими значениями условных вероятностей). Возможно, это объясняется уже отмечавшейся большей "контрастностью" T_2 по сравнению с T_I.

3)Очень интересны некоторые "несущественные"близкие связи. К ним прежде всего относятся связи однородных прилагательных,в меньшей степени - существительных и обратные связи существительных и прилагательных,которые порой бывают довольно сильными.

Таким образом,выделение близких связей позволяет лучше понять структуру таблиц условных вероятностей.

Поиск наилучшего прадерева.

Третий эксперимент был посвящен проверке сформулированной выше гипотезы о том,что граф простого предложения есть наилучшее прадерево.Новое определение множества допустимых решений требует несколько отличной оптимизационной процедуры.Эта процедура базируется на следующих вспомогательных понятиях:

1)Вершина r некоторого дерева называется корневой,если в нее не входят никакие стрелки.

2)Вершина w называется своей корневой для вершины v ,если w есть корневая вершина и существует путь из w в v.

В процедуре применяются следующие правила:

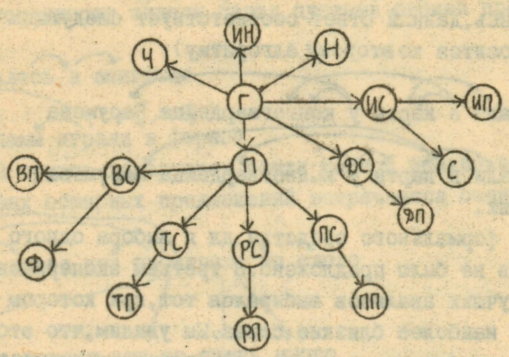
1)Первоначально все вхождения слов,составляющих простое предложение считаются изолированными.Тем самым , каждое слово v есть корневая вершина и,кроме того,своя корневая вершина для этого же слова v .

2)На очередном этапе процедуры произвольная вершина x может соединяться стрелкой с вершиной y , идущей в вершину y ,где y - корневая вершина,не являющаяся своей корневой для вершины x.

3)Стрелкой s_{ab} соединяются вершины a и b такие, что P(u(a)/l(b)) максимальна в множестве условных вероятностей для стрелок,определяемых правилом 2.

Процедура заканчивается,когда множество стрелок,определяемых правилом 2,пусто.

Далее приводится роза подчинения, вычисленная для данной процедуры.

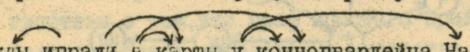


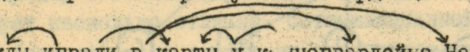
В третьем эксперименте по другому решалась проблема многозначности ответов. В первых двух алгоритмах ответ выдавался в форме, содержащей все наилучшие анализы (хотя и не в явном виде), например:

№п/п	Слово	Код	Код связанного слова
I	Однажды	Н	Г
2	играли	Г	О
3	в	П	-ВС
4	карты	ВС	Г
5	у	П	-РС
6	конногвардейца	РС	П
7	Нарумова	РС	П

В графе "код связанного слова" без знака дан код управляющего слова, а подчиненное слово дается со знаком "минус". Выдача программ не содержала естественной записи слова (графа "слово") и порядкового номера, а только две последние графы.

Ввиду того, что разные вхождения слова одного и того же класса не различались, данный ответ соответствует следующим анализам (пример относится ко второму алгоритму):





и ряду других.

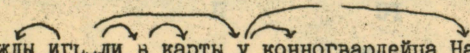
Никакого формального средства для выбора одного из наилучших анализов не было предложено. В третьем эксперименте из множества наилучших анализов выбирался тот, при котором стрелками соединяются наиболее близкие слова. Мы увидим, что этот метод не свободен от ошибок, но во многих случаях дает удовлетворительные ответы. В то же время получение однозначного ответа есть несомненное удобство.

Например, в третьем эксперименте предыдущее предложение было выдано в таком виде:

I	Однажды	Н	2
2	играли	Г	0
3	в	П	2
4	карты	ВС	3
5	у	П	2
6	конногвардейца	РС	5
7	Нарумова	РС	5

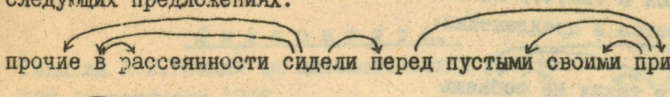
Здесь в последней графе дается номер управляющего слова по порядку. 0 означает, как и в предыдущем примере, что данное слово не имеет управляющего. Реальная выдача содержала лишь две последние графы.

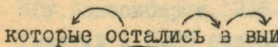
С помощью стрелок этот ответ записывается так:



Этот ответ верен.

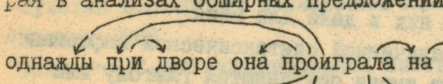
Как и следовало ожидать, в третьем эксперименте связь глагол→предлог→существительное в большинстве случаев передается правильно. Помимо предыдущего, это отражается, например, в следующих предложениях:





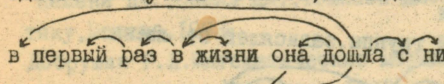


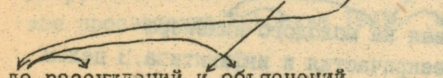
Следующий пример характерен почти полной правильностью, которая в анализах обширных предложений встречается очень редко:



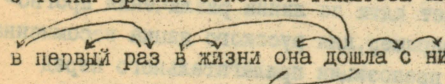


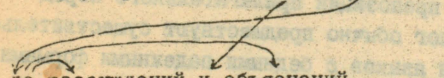
Несущественная неправильность, объясняющаяся выдачей лишь одного наилучшего ответа, наблюдается в следующих примерах:





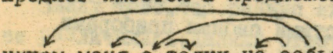
Здесь слово "раз" ближе к второму вхождению предлог "в". С точки зрения основной гипотезы анализ



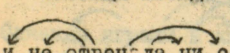


так же является наилучшим. Существенной ошибкой здесь является отнесение "и" к слову "она".

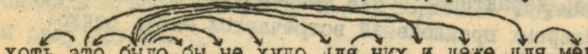
Любопытно,что в языках с постпозицией прилагательных ошибок такого рода было бы существенно меньше.Остальные ошибки неплохо отражаются розой подчинения.Такова например,неизбежная ошибка, заключающаяся в том,что ВС всегда подчиняется предлогу,если предлог имеется в предложении:



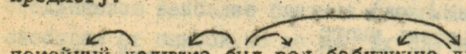
Частицы всегда подчиняются глаголу:



Союзы подчиняются либо ИС,либо,при отсутствии ИС,глаголу:

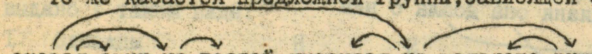


Как это неизбежно при ограниченной синтаксической информации, косвенные падежи,зависящие от имени,подчиняются глаголу или предлогу:





То же касается предложной группы,зависящей от имени:



Плохо отражаются связи деепричастия и инфинитива.В целом очевидно значительное улучшение результатов по сравнению с первым и вторым алгоритмами.

Дальнейшее улучшение может идти по линии различения прогнозирования вправо и влево.Например,для русского языка несомненна большая предпочтительность препозиции прилагательного перед постпозицией;глагол и предлог обычно предшествуют существительному в косвенном падеже.Для языков с бедными падежными формами подобное различение может иметь тем большее значение.Могут быть полезны и детализация и использование понятий близкой и далекой связи;наконец,всегда есть возможность использовать

более богатую исходную синтаксическую или лексическую информа - цию.

Впрочем, использование более богатой синтаксической информации выдвигает более сильные требования к типологически предшествующим алгоритмам классификации.

Литература.

Гладкий А.В. Лекции по математической лингвистике для студентов НГУ. Новосибирск, 1966.

Гладкий А.В. Об описании синтаксической структуры предложения, Новосибирск, 1971.

Сухотин Б.В. Алгоритмы лингвистической дешифровки, Проблемы структурной лингвистики, Москва, 1963.

Сухотин Б.В. Методы дешифровки сообщений, Внеземные цивилизации, Москва, 1969.

Фитиалов С.Я. О моделировании синтаксиса в структурной лингвистике, Проблемы структурной лингвистики, Москва, 1961.

Берж К. Теория графов и ее применение. Москва, 19[illegible].

Прим Р.К. Кратчайшие связывающие сети и некоторые обобщения, Кибернетический сборник 2, ИЛ, Москва, 1961.

Кемени Дж., Снелл Дж., Томпсон Дж., Введение в конечную математику, Москва, 1963.

Добрушин Р.Л. Математические методы в лингвистике. Математическое просвещение, 6, Москва, 1961.

www.ruslang.ru

А 06735 от 9.3.72, 25 п.л. 100 экз. Зак.№ 111

Тип. МГПИИЯ

30k.