*И.М. Богуславский, Л.Л. Иомдин, Л.Г. Митюшин, В.Г. Сизов*

# ДЛИНА СИНТАКСИЧЕСКИХ СВЯЗЕЙ В РУССКОМ АННОТИРОВАННОМ КОРПУСЕ[1]

## 1. Введение

В данной работе проводится статистический анализ корпуса СинТагРус (Syntactically Tagged Russian text corpus – синтаксически аннотированный корпус русских текстов). Корпус представляет собой совокупность текстов различных жанров (художественных, научно-популярных, информационных и др.), где каждому предложению приписана морфо-синтаксическая структура в виде дерева зависимостей в соответствии с формализмом, используемым в многофункциональном лингвистическом процессоре ЭТАП[2] и основанным на модели «Смысл ⇔ Текст»[3]. Работа над корпусом была начата в 2000 г.[4] и продолжается в настоящее время. В начале 2008 г. корпус содержал около 32 000 предложений общим объемом около 460 000 слов.

В компьютерно-лингвистической литературе неоднократно отмечалось, что в текстах на естественном языке предпочитаются

[1] Работа выполнена при финансовой поддержке РФФИ, гранты 07-06-00339 и 08-06-00373.

[2] *Apresjan Ju., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L.* ETAP-3 linguistic processor: a full-fledged NLP implementation of the MTT // Proceedings of the First International Conference on Meaning – Text Theory (MTT-2003). Paris, 2003. P. 279–288.

[3] *Мельчук И.А.* Опыт теории лингвистических моделей «Смысл ⇔ Текст». М.: Наука, 1974.

[4] *Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N.* Dependency treebank for Russian: concept, tools, types of information // Proceedings of the 18th Conference on Computational Linguistics (COLING-2000). Vol. 2. Saarbrücken, Germany, 2000. P. 987–991.

более короткие синтаксические связи[1]. В данной работе делается попытка дать статистическое подтверждение этому наблюдению. Для различных типов связей анализируется характер зависимости частоты связей от длины. Рассматривается также условная вероятность наличия связи в ситуации, благоприятной для ее возникновения, и зависимость этой вероятности от длины связи. Анализ статистических данных показывает, что для большинства типов связей в обоих случаях имеет место устойчивое экспоненциальное убывание рассматриваемых величин.

## 2. Данные в корпусе СинТагРус

Морфо-синтаксическая структура предложения представляет собой дерево зависимостей – ориентированное дерево, узлы которого соответствуют словам предложения, а дуги помечены именами синтаксических отношений. Узлы считаются упорядоченными в соответствии с порядком слов в предложении. Для каждого узла указывается лемма, часть речи слова и набор морфологических характеристик, выражающих значения таких категорий как вид, время, залог, лицо, число, падеж, род и т.п. В процессоре ЭТАП и корпусе используются обычные русские части речи: глагол (V), существительное (S), прилагательное (A), наречие (ADV), числительное (NUM), предлог (PR), союз (CONJ), частица (PART), междометие (INTJ).

В корпусе используется более 60 синтаксических отношений. Многие из них соответствуют очень частым и простым ситуациям (например, связь между некоторым словом и его «главным» дополнением представляется 1-м комплетивным отноше-

[1] См., например: *Митюшин Л.Г.* Длина синтаксических связей и индуктивные структуры // Семиотика и информатика. Вып. 26. М., 1985. С. 34–51; *Lin D.* On the structural complexity of natural language sentences // Proceedings of the 16th Conference on Computational Linguistics (COLING-1996). Vol. 2. Copenhagen, 1996. P. 729–733.

нием, связь между существительным и модифицирующим его прилагательным – определительным отношением, связь между глаголом и модифицирующим его наречием – обстоятельственным отношением). С другой стороны, есть и редкие отношения – например, аппроксимативно-количественное, связывающее существительное и числительное в выражениях со значением приблизительного количества (*минут десять*).

### 3. Статистические данные о длине связей

Под длиной связи (или расстоянием) *d* между узлами *A* и *B* понимается число узлов дерева, расположенных между *A* и *B*, плюс 1. Таким образом, для соседних слов *d* = 1. Прежде всего приведем некоторые общие статистические показатели, характеризующие длину связей без учета маркирующих связи синтаксических отношений.

| | |
|---|---|
| Средняя длина связей, идущих влево: | 2,18 |
| Средняя длина связей, идущих вправо: | 1,99 |
| Доля связей, идущих влево: | 0,38 |
| Средняя длина всех связей: | 2,06 |

Теперь рассмотрим распределение длины для наиболее частотных типов связей. Мы будем рассматривать только проективные связи. Напомним, что связь от *A* к *B* называется проективной, если все узлы, линейно расположенные между *A* и *B,* прямо или косвенно подчинены *A*. Более 99% связей в корпусе являются проективными.

Под связью определенного типа мы понимаем определенное синтаксическое отношение, соединяющее слова определенных частей речи в определенном направлении. Мы рассмотрим 40 наиболее частотных типов связей; в совокупности они покрывают около 77% всех связей корпуса. Наиболее частый тип – связь A ← S с определительным отношением, которая встречается 53 252 раза (12,5% от общего числа связей в корпусе). В

табл. 1 для этого типа приведено распределение частот по длине (частоты меньше 10 не приводятся, так как считаются статистически недостоверными). Около 84% общей частоты приходится на $d = 1$; при $d \geq 2$ числа убывают приблизительно экспоненциально.

*Таблица 1.* Определительная связь A ← S.
(О содержании столбцов 3 и 4 см. раздел 5.)

| длина связи *d* | частота *F(d)* | число допустимых ситуаций | оценка Pr**(d)* |
|---|---|---|---|
| 1 | 44707 | 47540 | 0,9404 |
| 2 | 6059 | 7155 | 0,8468 |
| 3 | 1580 | 2768 | 0,5708 |
| 4 | 561 | 1337 | 0,4196 |
| 5 | 194 | 783 | 0,2478 |
| 6 | 76 | 560 | 0,1357 |
| 7 | 40 | 439 | 0,0911 |
| 8 | 13 | 372 | 0,0349 |

Подобная картина характерна для большинства рассмотренных типов связей. Начиная с $d = 2$ (иногда с $d = 3$), имеет место приблизительно экспоненциальное убывание; при $d = 1$ и 2 возможны разного рода «индивидуальные отклонения».

В Приложении для каждого рассматриваемого типа указывается средняя длина связи, а также коэффициент регрессии $a$, характеризующий скорость убывания логарифма частоты связей данного типа при возрастании длины. Значение $a$ вычисляется следующим образом. Определяется отрезок $[d_1, d_2]$ рассматриваемых значений $d$ по следующим правилам:

$d_1$ – точка максимума $F(d)$ при $d \geq 2$ (чаще всего $d_1 = 2$).
$d_2$ – наибольшее число, такое, что $F(d) \geq 10$ на отрезке $[d_1, d_2]$.

Далее строится линейная функция $ad + b$, которая является наилучшим приближением $-\ln F(d)$ на $[d_1, d_2]$ в смысле суммы квадратов отклонений. Таким образом, на $[d_1, d_2]$ частоты $F(d)$ ведут себя приблизительно как $C\,e^{-ad}$, где $C$ – константа.

### 4. Условные вероятности связей

Деревья зависимостей, как правило, можно собирать методом «снизу вверх» из отдельных узлов (тривиальных поддеревьев), последовательно проводя синтаксические связи между построенными к данному моменту соседними поддеревьями. Реальные деревья зависимостей, не допускающие такую сборку, редки – в корпусе всего 30 таких предложений.

Предположим, что нам известны части речи $P_1, \ldots, P_n$ всех слов некоторого предложения, и мы хотим определить вероятность того, что данное предложение имеет дерево зависимостей $T$ (имеющее $n$–1 связь). Предположим, что $T$ можно собрать из отдельных узлов методом «снизу вверх». Зададим некоторый порядок сборки, т.е. порядок связей $l_1, \ldots, l_{n-1}$ дерева $T$ так, что каждая связь $l_i$ проводится между соседними уже построенными поддеревьями (это могут быть либо исходные изолированные узлы, либо поддеревья, образованные некоторыми из связей $l_1, \ldots, l_{i-1}$). Тогда

$$\Pr(T / P_1, \ldots, P_n) = \Pr(l_1, \ldots, l_{n-1} / P_1, \ldots, P_n)$$
$$= \Pr(l_1 / P_1, \ldots, P_n) \cdot \Pr(l_2 / l_1, P_1, \ldots, P_n) \cdot \ldots \cdot \Pr(l_{n-1} / l_1, \ldots, l_{n-2}, P_1, \ldots, P_n).$$

Вместо этой точной формулы для вероятности $T$ при заданных $P_1, \ldots, P_n$ будем рассматривать приближенную оценку правдоподобия $T$, получаемую заменой сомножителей $\Pr(l_i / l_1, \ldots, l_{i-1}, P_1, \ldots, P_n)$ на их «контекстно-свободные» аппроксимации вида $\Pr(R / P_A, P_B, d, s)$, где последнее выражение означает вероятность того, что от узла $A$ к узлу $B$ идет связь с синтаксическим отношением $R$ при условии, что известны части речи $P_A$ и $P_B$ узлов $A$ и $B$, расстояние $d$ между ними и направление

связи $s$ (0 = влево, 1= вправо). При этом подразумевается, что, в некотором смысле, $A$ и $B$ являются в $T$ вершинами некоторых соседних поддеревьев (более точно это условие обсуждается в разделе 5). Новая оценка имеет вид

$$\Pr_{appr}(T / P_1, \ldots, P_n) = \prod_{i=1}^{n-1} \Pr(R_i / P_{Ai}, P_{Bi}, d_i, s_i),$$

где $R_i$, $P_{Ai}$, $P_{Bi}$, $d_i$, $s_i$ – соответственно синтаксическое отношение, часть речи хозяина, часть речи слуги, длина и направление связи $l_i$.

Заметим, что задание определенных значений $R$, $P_A$, $P_B$, $d$ и $s$ эквивалентно выбору определенного типа связи. Оценим значения $\Pr(R / P_A, P_B, d, s)$ для наиболее частотных типов связей, используя данные корпуса. Прежде всего необходимо внести одно ограничение на «контекстную независимость», связанное с так называемыми неповторимыми синтаксическими отношениями. Смысл «неповторимости» состоит в том, что из узла может идти не более одной связи, помеченной данным отношением. Например, 1-е комплетивное отношение является неповторимым (что означает, в частности, что глаголу разрешается иметь не более одного прямого дополнения), тогда как обстоятельственное отношение повторимо (в частности, глагол может иметь несколько модифицирующих его адвербиальных групп). Большинство отношений, используемых в процессоре ЭТАП и корпусе, являются неповторимыми.

Ограничения, налагаемые неповторимостью, соблюдаются во всех деревьях корпуса. Поэтому для связей с неповторимым отношением $R$ в $\Pr(R / P_A, P_B, d, s)$ присутствует неявное дополнительное условие: из узла-хозяина $A$ не идет связь с отношением $R$ в какой-либо узел, отличный от $B$. Это учитывается при вычислении оценок $\Pr(R / P_A, P_B, d, s)$ по данным корпуса.

### 5. Оценка условных вероятностей по данным корпуса

Чтобы оценить Pr ($R$ / $P_A$, $P_B$, $d$, $s$), для некоторых $R$, $P_A$, $P_B$, $d$ и $s$, рассматриваются все пары узлов $A$ и $B$ в предложениях корпуса, имеющие части речи $P_A$ и $P_B$ и расположенные в порядке $s$ на расстоянии $d$ друг от друга. Для каждой такой ситуации проверяется, что $A$ и $B$ являются в $T$ вершинами некоторых двух соседних поддеревьев. Для этого удаляются все связи в $T$, идущие из $A$ в $B$ или «дальше $B$», и идущие из $B$ в $A$ или «дальше $A$», и в оставшемся графе рассматриваются полные поддеревья $T_A$ и $T_B$ с вершинами $A$ и $B$. Если они граничат, т.е. покрываемые ими отрезки не пересекаются и расстояние между ближайшими узлами $T_A$ и $T_B$ равно 1, ситуация считается допустимой. Если отношение $R$ является неповторимым, из числа допустимых исключаются ситуации, когда в $T_A$ из $A$ выходит связь с отношением $R$. За оценку Pr ($R$ / $P_A$, $P_B$, $d$, $s$) (обозначаемую Pr*) принимается отношение числа $F$ тех допустимых ситуаций, в которых в $T$ имеется связь с отношением $R$ из $A$ в $B$, к числу всех допустимых ситуаций.

Для наиболее частого типа связи – A ← S с определительным отношением – соответствующие данные приведены в табл. 1. Значения Pr*$(d)$, как и $F(d)$, убывают приблизительно экспоненциально, хотя и с меньшей скоростью, чем $F(d)$. Такой характер убывания Pr*$(d)$ типичен для большинства рассмотренных типов связей.

В Приложении приводятся коэффициенты регрессии $a_1$ для –ln Pr*$(d)$.Для 33 из 40 типов связей выполняется $a_1 \geq 0{,}1$. Среди остальных семи выделяется тип 2 (PR → S с предложным отношением): у него с ростом длины абсолютная частота связей быстро убывает, а вероятность наличия связи в допустимой ситуации остается близкой к 1. Остальные шесть типов (3, 7, 9, 14, 25 и 29) интуитивно ощущаются как допускающие весьма далекие связи. Небольшое отрицательное значение $a_1$ для типа 29 – скорее всего случайная ошибка, возникшая в ситуации малого числа наблюдений.

**Приложение. 40 наиболее частотных типов связей**

| № | части речи и направление | сокращенное имя отношения | частота | средняя длина | $a$ | $a_1$ |
|---|---|---|---|---|---|---|
| 1 | A <-- S | опред | 53252 | 1,24 | 0,992 | 0,513 |
| 2 | PR --> S | предл | 50503 | 1,45 | 1,122 | 0,024 |
| 3 | S <-- V | предик | 24969 | 2,59 | 0,318 | 0,039 |
| 4 | V --> S | 1-компл | 17339 | 1,89 | 0,703 | 0,449 |
| 5 | S --> S | квазиагент | 16911 | 1,43 | 1,066 | 0,693 |
| 6 | S --> S | 1-компл | 13447 | 1,53 | 0,963 | 0,595 |
| 7 | ADV <-- V | обст | 10399 | 2,23 | 0,367 | 0,086 |
| 8 | V --> S | предик | 9101 | 1,96 | 0,697 | 0,425 |
| 9 | PR <-- V | обст | 8819 | 5,23 | 0,289 | 0,037 |
| 10 | CONJ --> V | соч-союзн | 8125 | 3,34 | 0,326 | 0,122 |
| 11 | V --> PR | обст | 7740 | 2,08 | 0,521 | 0,234 |
| 12 | PART <-- V | огранич | 7033 | 1,22 | 0,657 | 0,105 |
| 13 | CONJ --> S | соч-союзн | 6615 | 1,57 | 0,829 | 0,612 |
| 14 | S --> CONJ | сочин | 6475 | 1,63 | 0,423 | 0,011 |
| 15 | S --> PR | атриб | 6439 | 1,56 | 0,558 | 0,141 |
| 16 | S --> S | аппоз | 6193 | 1,52 | 0,508 | 0,214 |
| 17 | CONJ --> V | подч-союзн | 5436 | 4,08 | 0,305 | 0,100 |
| 18 | V --> PR | 1-компл | 4841 | 1,26 | 0,897 | 0,446 |
| 19 | S <-- V | 1-компл | 4744 | 2,54 | 0,430 | 0,118 |
| 20 | NUM <-- S | количест | 4566 | 1,23 | 1,317 | 0,604 |
| 21 | V --> PR | 2-компл | 4490 | 1,64 | 0,719 | 0,318 |
| 22 | S --> S | сочин | 4378 | 2,32 | 0,398 | 0,121 |
| 23 | V --> V | 1-компл | 4224 | 1,73 | 0,540 | 0,400 |
| 24 | S --> S | атриб | 3588 | 1,68 | 1,053 | 0,675 |
| 25 | V --> CONJ | сочин | 3444 | 3,61 | 0,339 | 0,041 |
| 26 | S --> PR | 1-компл | 3363 | 1,29 | 0,783 | 0,204 |
| 27 | V <-- S | опред | 3355 | 1,80 | 0,621 | 0,414 |
| 28 | S --> V | релят | 2831 | 4,18 | 0,383 | 0,109 |
| 29 | V --> CONJ | сент-соч | 2654 | 4,49 | 0,296 | -0,016 |
| 30 | S --> V | опред | 2653 | 1,87 | 0,488 | 0,153 |
| 31 | V --> ADV | обст | 2623 | 1,76 | 0,711 | 0,483 |
| 32 | S <-- A | предик | 2171 | 2,63 | 0,412 | 0,166 |
| 33 | CONJ --> A | соч-союзн | 2166 | 1,92 | 0,407 | 0,280 |
| 34 | ADV <-- A | огранич | 2163 | 1,08 | 1,710 | 1,057 |
| 35 | V --> S | 2-компл | 2064 | 1,88 | 0,733 | 0,490 |
| 36 | V --> V | сент-соч | 1907 | 6,52 | 0,227 | 0,151 |
| 37 | V --> V | сочин | 1900 | 3,89 | 0,324 | 0,204 |
| 38 | PART <-- S | огранич | 1847 | 1,32 | 1,295 | 0,303 |
| 39 | PART <-- ADV | огранич | 1822 | 1,04 | 3,765 | 0,791 |
| 40 | V --> CONJ | 1-компл | 1793 | 1,18 | 0,927 | 0,789 |