“ТИПОВОЙ” КОНТЕКСТ: СЛУЧАЙНОСТЬ ИЛИ ЗАКОНОМЕРНОСТЬ?[1]

“PATTERN” CONTEXT: RANDOMNESS OR REGULARITY?

О.А. Митрофанова *(alkonost@om12520.spb.edu)*
Санкт-Петербургский государственный университет, Санкт-Петербург

С.А. Крылов *(krylov-58@mail.ru)*
Институт востоковедения РАН, Москва

Исследование посвящено определению статуса “типовых” контекстов, отражающих употребление лексем в том или ином значении и их сочетаемостные характеристики. Особое внимание уделяется сопоставлению данных из толковых словарей и корпусов русского языка. Результаты проведенного эксперимента позволяют скорректировать процедуры синтагматического анализа и извлечения семантической информации из текста.

> Наука имеет дело лишь с наблюдаемыми вещами …
> мы можем наблюдать объект, лишь заставляя его
> взаимодействовать с чем-нибудь внешним…
>
> *П.А.М. Дирак. Принципы квантовой механики.*
> *М., 1960. С. 18.*

1. Мотивация

Поводом для проведения настоящего исследования явились некоторые наблюдения над контекстной информацией об употреблении слов, представленной в иллюстративной части лексикографических описаний и в корпусах текстов. Данные наблюдения были сделаны при подготовке экспериментов по определению тесноты семантических связей лексических единиц на основе близости их сочетаемостных свойств. Прежде чем приступать к практическому решению этой задачи, необходимо выяснить следующее:

1) какие данные о синтагматике слов можно получить при работе с разноплановыми источниками (словарями и корпусами текстов) и в каком отношении эти данные отличаются друг от друга;
2) в какой мере словари и корпуса текстов совместимы как носители синтагматической информации, в чём их преимущества и недостатки, каким источникам надо отдавать предпочтение.

Ответы на эти вопросы может дать анализ “типовых” контекстов в иллюстративных примерах толковых словарей и их верификация с помощью корпусных данных. На первый взгляд, более естественным и лингвистически оправданным является переход от текста к словарю (что в действительности происходит в процессе формирования словарных статей, при отборе примеров употребления слов в том или ином значении). Однако с учётом стоящей перед нами цели, выбор противоположного направления движения – от словаря к тексту – оказывается столь же оправданным, так как лексикографические описания регистрируют реальные лингвистические факты, при этом обладают эксплицитной структурой и являются обозримыми. Тем самым, словарь (а особенно его иллюстративная часть) оказывается одновременно и “выходом”, и “входом” по отношению к корпусу текстов. Это означает, что исследование природы, формальных и содержательных характеристик “типовых” контекстов, представленных в иллюстративной части словаря, может быть полезным как в фундаментальном аспекте (в связи с созданием комплексного описания синтагматических связей единиц текста, а также дифференциации и изучения различных типов сочетаний, возникающих на их основе), так и в прикладном (возможно применение данных описаний при снятии лексической неоднозначности, определении тесноты семантических связей, автоматической классификации лексических единиц на основе близости их синтагматических свойств).

2. Исходная гипотеза и метод её верификации

В обсуждении принципов создания иллюстративной части словаря особое внимание уделяется тем приёмам, которые целесообразно использовать при подборе примеров употребления слов в том или ином значении [Семёнова 2003]. С одной стороны, лексикограф вправе воспользоваться готовыми цитатами, которые

[1] Данная работа выполнена при финансовой поддержке гранта Президента РФ для поддержки молодых российских ученых № МК-9701.2006.6

отражают реальные факты употребления лексических единиц, но могут содержать избыточную, случайную информацию. С другой, допустимо создание модельных примеров, наиболее удачные из которых согласуются с узусом, не перегружены элементами частного характера, и в то же время, позволяют демонстрировать спектр семантических / синтаксических валентностей слова. В обоих случаях подготовка примеров требует особой тщательности. Качество иллюстративной части в немалой степени определяет ценность лексикографического источника как такового, а также то, может ли он служить инструментом компьютерного понимания естественно-языкового текста [Леонтьева, Семёнова 2003].

Вероятно, по отношению к корпусу текстов иллюстративный блок словаря выступает в качестве особой модели. Учитывая тот факт, что примеры могут поступать в словарь непосредственно из реальных речевых произведений или создаваться лексикографом на основе исследовательских наблюдений, есть некоторые основания утверждать, что множество "типовых" контекстов, из которых и состоит иллюстративный блок лексикографического источника, может оказаться неоднородным. Если это верно, тогда часть множества "типовых" контекстов будет максимально приближена к репрезентативной выборке из корпуса текстов, другая часть должна напоминать модель подобной выборки. Насколько же велика дистанция между "типовыми" контекстами – модельными примерами из словаря и контекстами из корпуса? Чтобы ответить на этот вопрос, предлагается сформулировать исходную гипотезу H_0 и проверить ее экспериментальным путём.

Исходная гипотеза H_0: Между синтагматическими связями лексических единиц в реальных текстах и в "типовых" контекстах, приводимых в словаре, существует качественное различие.

Метод верификации исходной гипотезы H_0: Следует выбрать экспериментальный материал – лексические единицы X, Y,.. и определить круг их синтагматических соседей (элементов, встречающихся в контекстах совместно с ними) на основе словаря с модельными иллюстративными примерами и корпуса текстов, а затем провести сопоставительный анализ сформированных таким образом синтагматических полей лексических единиц X, Y,..

Если состав синтагматических полей, сформированных на основе словаря и корпуса, будет однороден, тогда гипотеза H_0 должна быть признана несостоятельной. Если окажется, что между синтагматическими полями данных типов есть значимые расхождения, тогда гипотеза H_0 находит подтверждение.

3. Планирование эксперимента

В ходе исследования предлагается использовать следующие источники контекстной информации: Словарь русского языка С.И. Ожегова [Ожегов 1989] в формате базы данных Starling (CO-Starling) [Крылов, Старостин 2005] и корпус Бокрёнок [Азарова, Синопальникова 2004], подключённый к корпус-менеджеру Bonito (BK-Bonito) [Rychly, Smrž 2004]. CO-Starling используется в проекте «Вавилонская башня» (http://starling.rinet.ru/), BK-Bonito – на кафедре математической лингвистики СПбГУ.

Необходимость обращения к Словарю русского языка С.И. Ожегова как к источнику контекстных данных обуславливается тем, что в его иллюстративный блок входят преимущественно модельные примеры (минимальные синтагмы), а также устойчивые сочетания. Кроме того, и толковая, и иллюстративная части словаря компактно представляют основной состав активной лексики русского языка, имеющей широкую сферу функционирования. Представление словаря в формате базы данных Starling открывает прямой доступ к его иллюстративному блоку и позволяет вести поиск "типовых" контекстов не только в отдельно взятой статье, но и в пределах словаря целиком. Тем самым, обращаясь к CO-Starling, можно выявить множество словарных контекстных примеров, в которых встречаются исследуемые лексемы, и сформировать их синтагматические поля.

Корпусные данные об употреблении слов в тех или иных контекстах удобно извлекать из корпуса текстов русского языка Бокрёнок объемом 21 млн с/у. В отношении словарного состава корпус ориентирован на общеупотребительную лексику. Тексты, включённые в корпус Бокрёнок, характеризуются тематическим разнообразием (40% – газетные тексты, 30% – научно-популярные тексты, 20% – художественные тексты, 10% – тексты нормативных актов). Хронологические рамки корпуса – с середины 80-х годов XX в. по настоящее время. Доступ к корпусу Бокрёнок обеспечивается корпус-менеджером Bonito, с помощью которого можно осуществлять поиск контекстов для лексем или словоформ, производить некоторые операции с выборочными совокупностями контекстов и определять их статистические характеристики. Таким образом, обращение к BK-Bonito позволяет формировать синтагматические поля исследуемых слов по данным корпуса.

4. Выбор экспериментального материала

В ходе эксперимента предлагается исследовать синтагматические поля высокочастотных слов, имеющих достаточно высокий индекс полисемии. Для эксперимента выбраны глагол *говорить* и существительное *год*. Такое решение было принято по следующим причинам.

- Слова *говорить* и *год* имеют высокую частотность: в списке 5000 наиболее частых слов С.А. Шарова (http://www.artint.ru/projects/frqlist.asp) глагол *говорить* приводится с показателем частотности 2059,32 ipm, а существительное *год* – с показателем 2042,67 ipm. Высокая частота встречаемости данных слов может означать и то, что круг их синтагматических соседей весьма широк.

- Слова *говорить* и *год* имеют достаточно высокий индекс полисемии (по СО глагол *говорить* имеет 7 значений, существительное *год* – 5). Как правило, за отдельным значением лексической единицы закреплена особая схема сочетаемости с контекстом, что обеспечивает качественное разнообразие синтагматических соседей рассматриваемых слов.
- Слова *говорить* и *год* характеризуются неравноценностью свойственных им значений, в тексте данные слова могут употребляться и как полнозначные, и как частично делексикализованные элементы (ср. употребление глагола *говорить* и существительного *год* в таких контекстах, как: ***говорить*** *полным голосом, марсианский* ***год****; собственно / вообще / кстати* ***говоря****; 1 января 2006* ***года***).

Приведённые аргументы подтверждают, что материал для экспериментов вполне представителен.

Итак, основное содержание эксперимента – это сопоставление множеств синтагматических соседей слов *говорить* и *год*, полученных на основе выборок иллюстративных примеров из CO-Starling и случайных выборок контекстов из BK-Bonito.

5. Качественный и количественный состав синтагматических полей глагола говорить и существительного год по данным CO-Starling

Синтагматические поля глагола *говорить* и существительного *год* по данным CO-Starling объединяют те лексемы, которые являются их синтагматическими соседями в иллюстративных примерах. Особенно важно, что в сформированные таким образом поля попадают данные не только из словарных статьей для *говорить* и *год*, но и из других статей, в иллюстративной части которых присутствуют исследуемые лексические единицы. По всей видимости, ядрами данных синтагматических полей следует считать элементы "типовых" контекстов, включённых в словарные статьи для глагола *говорить* и существительного *год* (например, ***говорить*** *по телефону*, *круглый* ***год***); на периферии полей находятся лексемы, употребление которых иллюстрируется типовыми контекстами, содержащими глагол *говорить* и существительное *год* (например, *Не путай,* ***говори*** *с толком; Дети умнеют с* ***годами***). Иными словами, по отношению к исследуемым лексическим единицам элементы ядерной зоны синтагматических полей выступают в качестве иллюстративного фона, тогда как элементы периферийной зоны – это объект иллюстрации в контексте исследуемых единиц.

Как в ядерной, так и в периферийной зонах полей содержатся синтагматические соседи, соответствующие активным и пассивным валентностям исследуемых слов. Представляется удобным противопоставлять два вида синтагматических соседей: это "типовые" хозяева (например, *Больному трудно* ***говорить****; времена* ***года***) и "типовые" слуги (например, ***говорить*** *правду; календарный* ***год***). Кроме этого, в поля могут попасть как закономерные элементы, семантически и / или синтаксически связанные с исследуемыми словами (например, ***говорить*** *шепотом* или ***годы*** *пронеслись*), так и более или менее случайные (например, глагол *говорить* упоминается в примере из словарной статьи наречия *вообще*: *Я* ***говорю*** *о людях вообще, а не о тебе*; существительное *год* приведено в иллюстрации из словарной статьи существительного *водность*: *Средний по водности* ***год***).

Синтагматические соседи из ядерной зоны поля – это носители основной информации о семантических и синтаксических валентностях исследуемых слов, а также о прототипической характеристике заполнения этих валентностей. Конечно, на основе типовых контекстов с ядерными элементами можно создать картину употребления того или иного слова в текстах, однако такая картина будет неполной. Восстановить недостающие фрагменты позволяют синтагматические соседи из периферийной зоны поля – элементы, нежёстко связанные с исследуемым словом и не всегда предсказуемые. А это означает, что более реалистичную картину употребления слова в текстах отражает его синтагматическое поле, не замыкающееся на контекстах отдельной словарной статьи, а охватывающее словарь целиком.

5.1. Говорить – *CO-Starling*

Синтагматическое поле глагола *говорить* по данным CO-Starling включает 228 синтагматических соседей с учетом дифференциации их значений (для ЛСВ) (из них 27 контекстов с ядерными синтагматическими соседями, остальные – с периферийными) или 190 – без учета дифференциации значений (для лексем).

Среди элементов "типовых" контекстов для глагола *говорить* есть его соседи слева (*строго* ***говоря***), соседи справа (***говорить*** *загадками*), соседи слева / справа (*о многом* ***говорить***, ***говорить*** *о многом*).

Синтагматические соседи слова *говорить*, встречающиеся в "типовых" контекстах, можно распределить по таким группам: "типовые" хозяева, "типовые" слуги, служебные слова:

Синтагматические соседи – "типовые" слуги:

существительные: ("предметные" актанты) *об успехах, о верёвке* и пр.; ("предметно-предикатные" актанты) *банальности, гадости, гоп, двусмысленности, дерзости, колкости, наглости, пакости, правду*, (сирконстанты) *аллегориями, баском, загадками, недомолвками, парадоксами, с акцентом, с апломбом, с издёвкой, с вывертами, с выкрутасами, без вычур, в глаза/за глаза, на диалекте, с достоинством, с желчью, без закавык, с заминками, сквозь зубы, с пафосом, с присвистом, в рифму, по телефону, с увлечением, по-французски* и пр.; (сирконстанты – именные группы) *вызывающим тоном, полным голосом, с победоносным видом, на ломаном языке, с отчаянием в голосе, с чужих слов, с чужого голоса* и пр.;

наречия: *внушительно, возмущённо, выше* (*об этом **говорилось** выше*), *медленно, по-русски, непонятно, кичливо, заносчиво, искренне, конкретно, коротко, льстиво, наперекор, общо, одухотворённо, серьёзно* и пр.;

компаратив: *яснее* и пр.;

деепричастия: *не таясь* и пр.;

местоимения: *сам за себя* (что-н. ***говорит** само за себя*) и пр.

Синтагматические соседи – "типовые" хозяева:

глаголы: *начать, заказать (правду **говорить** никому не закажешь)* и пр.;

предикативы: *трудно, не место (здесь не место **говорить** о пустяках), не пристало* и пр.

Синтагматические соседи – служебные слова:

предлоги: *о, от, по, с* и пр.;

союзы: *как, что* и пр.;

частицы: *вот, не* и пр.

В числе иллюстративных примеров употребления глагола *говорить* присутствуют "типовые" контексты со специфическим субъектом (например, *в нём **говорит** гордость, факт **говорит** о многом, **говорящий** попугай*) и специфическим объектом – субъектом при пассиве (***говоримая** речь*), идиоматизированные выражения (например, *у кого что болит, тот о том и **говорит**; в доме повешенного не **говорят** о верёвке; не **говори** гоп, пока не перепрыгнешь; **говорить**, бия себя в грудь / не закрывая рта / как по писаному; двадцать / сто раз **говорил*** и пр.), вводные конструкции (***говорят**, что; **говорю** Вам; как **говорят**; строго / короче / иначе **говоря***), окказиональные контексты (*вдругорядь **говорю***).

Надо заметить, что в составе "типовых" контекстов присутствуют не только синтагматические соседи глагола *говорить*, но и некоторые потенциальные парадигматические корреляты (*говорить – думать, подумать:* (*думает одно, а **говорит** другое; сперва **подумай**, а потом говори*); *говорить – кричать:* (*не кричи, **говори** спокойно*); *говорить – замолчать* (*долго **говорил** и наконец замолчал*); *говорить – подразумевать* (***говорит** о других, а подразумевает себя*). Это бывает, как правило, в контексте сочинительных (или, шире, паратактических) отношений между глаголом *говорить* и его коррелятами. В таких контекстах эксплицируется не только "собственно" антонимия в строгом смысле слова (ср. антонимическую пару *говорить – молчать*), но и отношение потенциального смыслового контрастирования (так называемая "речевая" или "контекстная" антонимия).

5.2. Год – *CO-Starling*

Синтагматическое поле существительного *год* по данным CO-Starling включает 149 синтагматических соседей с учетом дифференциации их значений (для ЛСВ) (из них 26 контекстов с ядерными синтагматическими соседями, остальные – с периферийными) или 106 – без учета дифференциации значений (для лексем).

С точки зрения линейного расположения в контекстах для слова *год* выделяются его соседи слева (*утро **года***), соседи справа (***годы** учёбы*), соседи слева / справа (*спустя **год**, **год** спустя*).

Синтагматические соседи слова *год*, встречающиеся в "типовых" контекстах, естественным образом распределяются по следующим группам: это "типовые" хозяева, "типовые" слуги, служебные слова:

Синтагматические соседи – "типовые" слуги:

прилагательные: *академический, голодный, грядущий, детский (детские **годы**), истёкший, календарный, круглый, незабываемый, незапамятный (в незапамятные **годы**), новый, плодородный, сельскохозяйственный, текущий, урожайный, учебный, финансовый, хлебный, хозяйственный, учебный, юный (юные **годы**)* и пр.;

порядковые числительные: *первый, второй, сорок пятый* и пр.;

наречия: *назад, спустя* и пр.;

существительные: *юность* (***годы** юности*) и пр.

Синтагматические соседи – "типовые" хозяева:

глаголы: (*год* – в субъектной позиции): *минуть, начаться, пройти, щадить (**годы** не щадят никого);* (*год* – в позиции сирконстанта): *видеться (не видеться **годами**), намыкаться (намыкаться за **годы** лишений), оставить (оставить на второй **год**), остаться (остаться на второй **год**), отработать, перенестись (перенестись в детские **годы**), проработать, просидеть, прослужить, проходить, расползтись (расползтись с **годами**), сесть (сесть на три **года**), умнеть (умнеть с **годами**)* и пр.;

существительные: *время, канун (Нового **года**), круговорот (времени **года**), начало, утро (**года**), чреда (**годов**), задел (задел на следующий **год**)* и пр.

Синтагматические соседи – служебные слова:

предлоги: *в (разница в **годах**, в прошлом **году**), за (год за **годом**, заработок за **год**), на (план на **год**), от (на сороковом году от **рождения**), под (в ночь под Новый **год**), с (с **годами** вкусы меняются), через / чрез (**годы**)* и пр.;

союзы: *как (прошёл **год**, как мы виделись)* и пр.

Ряд иллюстративных примеров – это идиомы (*Живёт такой **год**, что на день семь погод*) и контексты с нерегулярными элементами (*выработать продукции больше против прошлого **года***). Кроме того, в некоторых "типовых" контекстах вместе с существительным *год* встречаются его парадигматические корреляты (*год – май: май холодный – **год** плодородный; год – зима: в этом **году** зима ранняя*).

Большинство элементов этого синтагматического поля составляют закономерные элементы, доля случайных элементов оказалась ничтожно мала.

6. Количественный и качественный состав синтагматических полей глагола *говорить* и существительного *год* по данным BK-Bonito

Для определения состава синтагматических полей глагола *говорить* и существительного *год* были проанализированы случайные выборки контекстов из BK-Bonito. Объем выборок контекстов для слов *говорить* и *год* составил 200 контекстов при ширине контекстного окна [–10 … 10]. Некоторые наблюдения показывают, что выборки такого объема репрезентативны и достаточно объективно отражают распределение лексических единиц в корпусе [Азарова 2004]. В каждом из контекстов выделялась наиболее информативная часть – минимальные синтагмы, которые должны быть достаточны для идентификации значения, в котором употреблено исследуемое слово. При обработке выборок из корпуса был сделан акцент на систематизацию тех элементов контекста, появление которых закономерно (и предопределяется активными и пассивными валентностями слов), при этом случайные, "инородные" фрагменты отбрасывались. В пределах минимальных синтагм определялись непосредственные и опосредованные соседи слева и справа. Например, контекст *{, а ее поэзия сама – вне мер. Но европейская культура <**говорила**> и её устами. Переводя Бодлера, Цветаева как бы измеряла}* содержит в своем составе минимальную синтагму глагола *говорить* с непосредственным соседом слева и с опосредованным соседом справа (*культура **говорила** # устами*); контекст *{что триумф может сослужить плохую службу и в олимпийский <**год**> соперницы будут каждый раз настраиваться на наших чемпионок}* содержит минимальную синтагму существительного *год* с непосредственным и опосредованным соседями слева (*в олимпийский **год***). Множества синтагматических соседей слов *говорить* и *год*, сформированные таким образом, и составляют синтагматические поля исследуемых лексических единиц по данным корпуса.

В процессе работы с BK-Bonito возможно получение некоторых данных о корреляции между употреблением исследуемых слов и их соседей (в частности, статистический блок Bonito позволяет вычислять коэффициент взаимной информации [Азарова, Синопальникова, Смрж 2005]). Как и следовало ожидать, в корпусе лексемы *говорить* и *год* проявляют сильную корреляцию со служебными словами, вспомогательными глаголами, личными и указательными местоимениями, со знаками препинания (а это свидетельствует о том, что в синтагмах слова *говорить* и *год* нередко являются пограничными элементами). Однако эти высокочастотные элементы контекста не несут существенной лексико-семантической информации, и потому не представляют особого интереса.

6.1. Говорить – *BK-Bonito*

Синтагматическое поле глагола *говорить* по данным BK-Bonito отражает употребление рассматриваемого слова в случайной выборке объемом 200 контекстов, которые распределяются по следующим группам:

38% – парентетические контексты (с прямой речью, с придаточными дополнительными);

11% – полипредикативные конструкции (с личными формами глагола – *мешать / начать / продолжать **говорить*** и пр.; с предикативами – *можно / надо / нужно / приходится **говорить*** и пр.);

24% – контексты со стандартными актантами (существительными или местоимениями – например, *мы уже с вами **говорили** на эту тему*);

20% – вводные конструкции (*иначе / короче / вообще / кстати / собственно **говоря**, как **говорят*** и пр.);

4,5% – контексты со специфическим субъектом (существительным или указательным местоимением: например, *о подлинности **говорит** сам дух Велесовой книги; это **говорит** о наличии у них зачатков разума*);

3,5% – безобъектные конструкции (например, *собеседник не соглашается и **говорит** больше, чем больной*).

Во всех упомянутых группах присутствуют контексты с разнообразными сирконстантами (***говорить** естественно, успокоительно, восхищённо, вяло, неуверенно, равнодушно, тихо, в старину, в прошедшем времени, на родном языке, по-немецки, по-хорошему* и пр.). Зарегистрирован единичный случай распространения глагола *говорить* "предметно-предикатными" актантами (***говорить** намёками*). Хотелось бы обратить внимание на то, что в двух контекстах из 200 в качестве чистых актантов используются лексемы *слово* (***говорить** добрые слова*) и *притча* (***говорить** притчу*). Среди служебных слов в непосредственном синтагматическом окружении доминируют предлоги *о*, *об* и союзы *что*, *как*.

При обработке контекстных данных с помощью BK-Bonito был определен высокий коэффициент взаимной информации для сочетаний слова *говорить* с существительными в субъектной позиции (например, *Заполь* – 9,437), с модификаторами (например, *прикладывая* – 9,285), с существительными в объектной позиции (например, *матросу* – 8,607), что согласуется с представлением о насыщенности синтаксических валентностей глагола.

6.2 Год – *BK-Bonito*

Синтагматическое поле существительного *год* по данным BK-Bonito отражает употребление рассматриваемого слова в случайной выборке объемом 200 контекстов, которые распределяются по следующим группам:

60% – контексты с числительными (даты, периоды времени, измеряемые в годах);

21% – контексты с прилагательными (3/4 с таксисно-дейктическим значением – *последний, следующий, текущий, нынешний, прошлый, ближайший* и пр., 1/4 с относительным значением (*олимпийский, детский (детские **годы**), юношеский (юношеские **годы**), марсианский* и пр.);

8% – контексты с местоимениями (*тот (в те **годы**, в тот же **год**), этот, наш (где наши **годы**), весь* и пр.);

6% – контексты с существительными в постпозиции (*правление, оккупация, осмысление, реформа, появление, расцвет* и пр.));

3% – контексты с существительными в препозиции (*начало, конец, половина*);

1,5% – контексты с идиомами (*из **года** в **год**, **год** от **года***);

0,5% – контексты с существительным *год* в значении меры (*запасы не в **годах**, а в тоннах*).

Наиболее типичные предлоги, встречающиеся в контекстах – *в, с, за, в течение*. Зарегистрированы единичные сочетания с союзами – *и, что*. В синтагматическое поле слов *год* по корпусным данным не попали глаголы: основная причина – их дистанцированность от исследуемой лексической единицы.

При обработке контекстных данных с помощью BK-Bonito был определен высокий коэффициент взаимной информации для сочетаний слова *год* с существительными в постпозиции (например, *Обезьяны* – 9,356), с порядковыми числительными (например, *сороковом* – 8,619), однако лексико-семантическая интерпретация полученных данных не представляется возможной.

7. Выводы

Результаты проведённого исследования позволяют говорить о принципиальной совместимости двух разноплановых описаний – корпуса реальных текстов и корпуса “типовых” контекстов, существующего внутри словаря. Следовательно, гипотеза H_0 пока не находит убедительного подтверждения.

Как и реальные тексты, “типовые” контексты синкретично отражают то, что является случайным и закономерным в системе языка и в речи.

Во-первых, анализ корпуса “типовых” контекстов позволяет с высокой точностью определить структурную организацию и лексическое наполнение реальных синтагм, в которых исследуемые слова могут оказаться либо главными, либо зависимыми элементами. В этом смысле, в “типовых” контекстах проявляется закономерное.

Во-вторых, в корпусе “типовых” контекстов могут встретиться некоторые маргинальные единицы, имеющие слабые семантические и синтаксические связи с исследуемыми словами. По всей видимости, так в “типовых” контекстах проявляется случайное.

В-третьих, выявлен класс контекстов, совмещающих случайное и закономерное: это контексты – идиоматичные выражения, контексты с редкоупотребляемыми и стилистически маркированными словами, частотность которых в реальных текстах достаточно низка. Для исследуемых слов связь с такими синтагматическими соседями является малопредсказуемой, и поэтому случайной. Вместе с тем, для данных синтагматических соседей восстановление контекстов является предсказуемым, тем самым, их связь с контекстными партнёрами является закономерной.

Итог эксперимента: информация о синтагматике исследуемых слов по данным CO-Starling оказывается

более разнообразной, чем можно было бы ожидать исходя из “типовых” контекстов в отдельно взятых словарных статьях, и при этом отражающей основные закономерности употребления данных слов в текстах;

более концентрированной, чем сведения о сочетаемости, почерпнутые из корпуса текстов BK-Bonito, и при этом более отфильтрованной, чем в случайных выборках контекстов из корпуса;

более сбалансированной, чем в реальном корпусе (в силу сглаживания различий по частотности, существующих между единицами текста).

Итак, синтагматические поля, формируемые на основе словарных баз данных в формате Starling, можно использовать как надёжный источник данных о сочетаемостных возможностях лексем при определении тесноты семантических связей между ними, что позволяет решать задачи автоматической классификации лексики и тезаурусного моделирования.

Главный итог: есть все основания рассматривать корпус “типовых” контекстов, формируемый на основе иллюстративной части словаря, как метаописание по отношению к большому корпусу текстов, сходное с ним по качественному наполнению, но отличающееся компактностью, гибкостью и удобством в обращении.

Список литературы:

1. *Азарова И.В.* Выявление лексикализованных понятий в RussNet с использованием контекстной информации из корпуса // XXXIII Международная филологическая конференция. Секция прикладной и математической лингвистики. Т. 1. СПб., 2004. С. 3– 10.

2. *Азарова И.В., Синопальникова А.А.* Использование статистико-комбинаторных свойств корпуса современных текстов для формирования структуры компьютерного тезауруса RussNet // Труды международной конференции “Корпусная лингвистика – 2004”. СПб., 2004. С. 5–15.
3. *Азарова И.В., Синопальникова А.А., Смрж П.* Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet // Труды международной конференции “Диалог – 2005”. М., 2005. С. 11–16.
4. *Крылов С.А., Старостин С.А.* Металингвистическая разметка текстовых баз данных в системе Starling и современные задачи корпусной лингвистики // Международная конференция “MegaLing – 2005”: Прикладная лингвистика в поиске новых путей. Симферополь, 2005. С. 33.
5. *Леонтьева Н.Н., Семёнова С.Ю.* Семантический словарь РУСЛАН как инструмент компьютерного понимания // Понимание в коммуникации: Материалы научно-практической конференции, 5–6 марта 2003 г. М., 2003. С. 41–46.
6. *Ожегов С.И.* Словарь русского языка / Под ред. Н.Ю. Шведовой. – 20-е изд. М., 1989.
7. *Семёнова С.Ю.* Примеры в компьютерном семантическом словаре: некоторые наблюдения над процессом подбора // Труды международной конференции “Диалог – 2003”. М., 2003. С. 593–598.
8. *Rychly P., Smrž P.* Manatee, Bonito and Word Sketches for Czech // Труды международной конференции “Корпусная лингвистика – 2004”. СПб., 2004. С. 116–121.