ИСПОЛЬЗОВАНИЕ СХЕМЫ НАСЛЕДОВАНИЯ РАМОК ВАЛЕНТНОСТЕЙ В ТЕЗАУРУСЕ RUSSNET ДЛЯ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТА

RUSSNET VALENCY FRAME INHERITANCE IN AUTOMATIC TEXT PROCESSING

И. В. Азарова *(azic@bsr.spb.ru)*
Санкт-Петербургский государственный университет

В. Л. Иванов *(artifex.i@gmail.com)*
ООО "Идеограф"

Е. А. Овчинникова *(e.ovchinnikova@gmail.com)*
ООО "Идеограф"

В докладе рассматривается процедура автоматического анализа текста проекта ИДЕОГРАФ, которая использует формально-грамматическое описание русского текста Rus4IR и ресурсы компьютерного тезауруса RussNet. Рамки валентностей RussNet позволяют разрешать лексическую и грамматическую неоднозначность. Рамки содержат спецификацию контекстных маркеров для валентностей, которые определены на материале текстового корпуса.

Семантическая интерпретация текста задается в виде структур пропозиций, ядерными элементами которых являются субъектно-объектные позиции, соотнесенные с семантическими деревьями тезауруса RussNet. На примере трех семантических деревьев описывается схема наследования рамок валентностей в RussNet, которая используется для уточнения анализа предложения, ранжирования вариантов анализа, унификации позиций при семантическом выводе.

1. Введение

Успех систем автоматического анализа текста, которые могут иметь довольно различные цели: от усложненного информационного поиска до извлечения из текстов фактической и оценочной информации, во многом зависит от того, насколько точным и надежным является грамматический анализатор текста. Несмотря на то, что на уровне человеческого понимания фразы, предложения и тем более тексты достаточно редко неоднозначны, результаты статистических версий грамматик дают невероятно большое количество вариантов анализа. Например, грамматика Penn Treebank, опробованная на случайном наборе предложений, дает в среднем 7.2×10^{27} разборов на предложение (Moore 2000). Для русского языка, обладающего свободным порядком слов, как и для других славянских языков, количество вариантов анализа еще больше увеличивается.

На сайте ACL Anthology[1] (архива исследовательских работ по компьютерной лингвистике), где представлены исследовательские материалы начиная с 1979 г., можно выделить две основные тенденции разработок: (1) системы с традиционными лингвистическими модулями, которые ориентированы на всеохватывающий анализ текста, причем анализ осуществляется «снизу-вверх» от наименьших единиц типа морфем или основ к наибольшим (предложениям, текстам); (2) эвристические системы с «распределенной» архитектурой и комплексным описанием, объединяющим данные нескольких лингвистических уровней. Очевидны недостатки этих подходов. Для первого типа характерна высокая неоднозначность анализа на каждом из последовательных уровней, которая дополнительно возрастает при переходе от одного уровня к другому. Второй тип систем может показывать высокую эффективность в некоторой ограниченной области, однако отсутствие понимания глубинных механизмов и реализация типа «ad hoc» не дает возможности расширения частных случаев (чаще всего, предметных областей) на другие области.

В подходе ИДЕОГРАФ[2] мы постарались совместить преимущества двух типов анализа: система имеет модульный принцип, при этом модули «низшего» лингвистического уровня встроены в модули следующего уровня, так что внешним модулем является семантический. Эта система «встраивания» структур низшего уровня

[1] http://ucrel.lancs.ac.uk/acl/
[2] http://www.ideograph.ru

в высшие позволяет существенно понизить неоднозначность анализаторов и облегчает передачу данных между модулями.

2. Системы ИДЕОЛОГ

Проект ИДЕОГРАФ разрабатывается совместно Кафедрой прикладной лингвистики Санкт-Петербургского государственного университета и компанией «Идеограф» (Азарова, Иванов, Овчинникова, 2005). В работе используются результаты предыдущих исследовательских проектов Кафедры прикладной лингвистики СПбГУ (Азарова 2002; www.phil.pu.ru/depts/12/ AGFL; Азарова и др. 2002; Азарова, Синопальникова 2004; http://www.phil.pu.ru/depts/12/RN).

В рамках проекта ИДЕОГРАФ используются

- AGFL формализм (Koster 1991) для грамматического описания текста;
- отдельные компоненты формализма HPSG (Pollard & Sag 1994) для синтаксического и семантического описания текста;
- структуры типизированных признаков (Carpenter 1992) для внутреннего представления лингвистических объектов;
- компьютерный тезаурус RussNet, словарь типа wordnet, который дополнен рамками валентностей;
- семантический модуль для разрешения грамматической и лексической неоднозначности и создания результирующих пропозиций;
- платформа ИдеоЛог для логического вывода.

Платформа ИдеоЛог является эффективной реализацией абстрактной машины (Takaki et al. 1997), которая обеспечивает процедуру унификации, определенную на структурах типизированных признаков (TFS). Платформа интерпретирует правила формального синтаксиса AGFL и HPSG, создает описание лингвистических объектов в терминах TFS, обеспечивает связь с внешними данными, в частности с тезаурусом RussNet.

3. Особенности грамматического и семантического представлений

Опишем кратко особенности структур грамматической и семантической информации, которая используется при анализе текста в проекте ИДЕОГРАФ.

Грамматический анализатор построен на базе контекстно-свободной грамматики AGFL, дополненной уровнем «признаков» с заданными значениями (в терминологии К. Костера они называются аффиксами). В качестве признаков используются бесспорные грамматические категории (например: вид и залог глаголов, падеж существительных и проч.), различные вспомогательные словоизменительные признаки (например: подтипы склонения, словоизменительные классы глаголов, возвратность и проч.), а также более сложные характеристики фразовых структур (например: отрицательный/ неотрицательный предикат, признак малого и большого количества для описания сочетаний количественных числительных с существительными и проч.). Хотя генеративное описание считается прямой противоположностью статистического подхода к анализу текста, мы исходим из концепции «гибридного разбора» (“hybrid parsing” – Beinema, Koster 2004), при котором предполагается использовать частотные параметры синтаксических конструкций в корпусе современных текстов с тем, чтобы ускорить вычисления и придать грамматическому анализу стереотипный характер. Таким образом, мы добиваемся того, чтобы в качестве «первого» разбора в подавляющем количестве случаев (95%) выдавались наиболее правдоподобные варианты.

Морфологический модуль использует словарь основ, который охватывает все слова, включенные в RussNet. Опознанная основа морфологического словаря отсылает к синонимическим рядам (синсетам) тезауруса RussNet, в которые входит данное слово. Такие связи являются «проекцией» леммы на тезаурус. Если грамматическая форма слова не совместима с каким-либо значением, то оно не будет входить в ее «проекцию». Например, леммы слова *гусеница* (1) 'личинка бабочки' и (2) 'замкнутая металлическая цепь, состоящая из звеньев, служащая вместо колес у тракторов, танков и т.п.' морфологически различаются формой вин. падежа мн. ч., поэтому во фразе «*его кидают под гусеницы танков*» лемма однозначно указывает на значение (2), напротив, во фразе «*выращивать гусениц плодожорки можно на среде…*» – на значение (1), а во фразах «*место на коже, к которому прикасалась гусеница, надо обдуть…*» или «*из темноты послышался грохот и лязг гусениц*» отбор подходящего значения будет осуществляться на других этапах анализа.

Выявленный через проекцию синсет в тезарусе активирует стандартную для wordnet-словаря окрестность – набор синсетов с установленными семантическими связями: родовидовыми (гипонимы-гиперонимы), связями часть-целое (меронимы-холонимы), анотонимы и проч. Последовательность гиперонимов синсета определяет его принадлежность к тому или иному семантическому дереву. Так значение (1) в предыдущем примере входит в дерево «животные», а значение (2) – дерево «артефакт» (предмет, созданный человеком). В семантическом блоке можно задавать объединения деревьев, регулярно используемые группировки получают стандартные обозначения совокупности, например «одушевленные» («человек», «животные»). Вопрос о самостоятельности некоторого дерева, во-первых, связан с набором образующих его синсетов (деревья имеют сопоставимые объемы), и во-вторых, с тем, насколько часто данная совокупность синсетов может выступать в качестве семантической спецификации валентной позиции (см ниже).

Поскольку в тезаурусе RussNet заданы значения только для существительных, глаголов, прилагательных и наречий, для слов других частей речи, в частности местоимений, в семантическом блоке определяется проекция

на структуру тезаурусных значений: личные местоимения 1 и 2-го лица указывают на вершину дерева «человек», а местоимение 3-го лица в м. и ж. роде ед. ч. или во мн. ч. имеет проекцию на ряд деревьев из более широкой совокупности «сущность».

Словарь основ может быть расширен за счет деривационного модуля, который порождает новые основы от имеющихся при помощи продуктивных префиксов и суффиксов. Сгенерированная основа получает «привязку» к синсетам RussNet посредством семантико-деривационных отношений. Например, префикс «анти-» образует новые основы для прилагательных и существительных, которые присоединяются к членам имеющихся синсетов отношением DER_ANTONYM_OPPOSITE (деривационный антоним комплементарного типа). Слова с данным префиксом, которые регулярно встречаются в корпусе (*антисоветский* 10.2 ipm, *антивоенный* 1.48 ipm, *антитело* 2.14 ipm и проч.), входят в RussNet, а следовательно, и в словарь основ, но такие образования как *антигерой* считаются потенциальными словами и получают грамматическую и семантическую интерпретацию лишь в деривационном блоке. Использование процедуры деривационного анализа носит ограниченный характер, поскольку она увеличивает время обработки одного слова на 10%, уменьшая количество неопознанных слов лишь на 3 %.

3. Синтактико-семантический компонент: рамки валентностей

Синтактико-семантический компонент обеспечивает взаимодействие грамматических и лексико-семантических данных, полученных на соотвествующих этапах анализа, для снятия неоднозначности на обоих уровнях. Информационным ядром описываемой процедуры являются рамки валентностей. Поскольку под этим термином понимают совершенно разные структуры, определим их как семантическое и грамматическое описание контекстных маркеров в RussNet, регулярно встречающихся в корпусе современных текстов при реализации значения данного синсета. Число валентностей в рамке варьируется.

В нашей системе рамки валентностей выявляются и описываются на этапе подготовки данных для тезауруса RussNet. Чтобы разграничить значения полисемантичного слова, размечается случайная выборка его контекстов из корпуса современных текстов. В качестве «нулевой» гипотезы используется структура значений в толковом словаре МАС. Количество контекстов в выборке на каждое из значений используется для новой нумерации значений в тезаурусе RussNet (это часть стандартной процедуры подготовки wordnet-словарей). Перечисляются грамматические и семантические параметры маркеров, занимающих одну и ту же функционально-синтаксическую позицию. Регулярно встречающиеся параметры задают валентности. Таким образом, валентности в RussNet определяются через частотность реализации маркеров. Минимальным порогом частотности является 35% от совокупности контекстов, реализующих данное значение. Те валентности, которые встречаются с высокой частотностью (66–100%), считаются обязательными, менее частотные – факультативными.

Мы провели исследование количества контекстов[3], которые необходимы для разграничения значений и определения контекстных маркеров, и выяснили, что 100 случайных контекстов из корпуса дают такую же схему валентностей, что и 1000 контекстов; минимальным набором является 25 контекстов. Проблемным для описанного подхода являются редко встречающиеся в корпусе значения, возможность приписывания валентностей для них связана со схемой наследования рамок валентностей в семантических деревьях, которая будет описана ниже.

Разметка частотности употребления значений в выборках контекстов показала, что в довольно большом числе случаев (около 80%) распределение частот носит весьма четкий характер: первое значение (которое, правда, не всегда совпадает с первым значением в толковом словаре) представлено в 50-70% контекстов, напротив, низкочастотные значения довольно плохо противопоставлены по частоте и регулярно составляют долю от 1 до 5% контекстов.

3.1. Спецификация рамок валентностей

Валентности характеризуются нескольким параметрами (см. схему 1). Один из них (*obligatory*) – параметр обязательности/факультативности был описан выше. Следующий признак (*active*) связан с тем, выступает ли характеризуемое значение в качестве грамматически главного или подчиненного слова. В зависимости от этого параметра рамки валентностей подразделяются на активные и пассивные. Активные рамки регулярно встречаются у предикативных слов (как правило, глаголов и прилагательных, а также их дериватов), они «предсказывают» появление определенных типов синтаксически связанных зависимых слов. Пассивные рамки оформляют грамматическую форму зависимых слов (чаще всего существительных), в которой реализуется отдельное значение. В качестве примера пассивной рамки можно привести употребление слова *лицо* при глаголах говорения в конструкции «в» + «лицо», что означает 'без церемоний'. Если эта конструкция присоединяется к другим группам глаголов (*ударить в лицо, заглянуть в лицо, дунуть в лицо* и проч.), то слово употребляется в своем первом значении 'передняя часть головы человека'.

Для общей характеристики рамки валентностей используется параметр (main_segment), задающий конструкцию, в рамках которой разрешаются валентности: пропозициональная или референциальная структура.

[3] Сходные оценки наборов предложений (25 vs 200), которые можно использовать в качестве обучающей совокупности для разграничения значений полисемантичного слова, были получены в работе (Leacock & Chodorow, 1998).

Семантическая характеристика валентности (sem_type) указывает тип сегмента. В конструкции пропозиции это может быть объект, атрибут пропозиции, встроенная пропозиция, а в референциальной структуре – объект и атрибут объекта. В том случае когда нет ясности в отношении этих параметров, они могут опускаться.

```
<VALENCY_FRAME active="yes" main_segment="proposition">
        <VALENCY obligatory="yes" role="arg0" sem_type="object">
                <VARIANTS>
                        <VARIANT>
                                <morph_data POS="noun|pron" CASE="nom"/>
                                <sem_data TYPE="group" ID="RUS-nAnimate"/>
                        </VARIANT>
                </VARIANTS>
        </VALENCY>
        <VALENCY obligatory="no" role="arg3" sem_type="object">
                <VARIANTS>
                        <VARIANT>
                                <morph_data POS="noun|pron" PREP="по" CASE="dat" />
                                <sem_data TYPE="top" ID="RUS-nLocation"/>
                        </VARIANT>
                </VARIANTS>
        </VALENCY>
</VALENCY_FRAME>
```

Схема 1. Пример xml-представления рамки валентностей

Ролевая характеристика валентности (role) регулярно встречается в различных концепциях описания валентности. Однако вместо традиционного набора (объектив, результатив и проч.) мы используем аргументную (субъектно-объектную) структуру, которая привязана к определенному семантическому дереву RussNet. Например, для дерева глаголов движения, которое рассматривается ниже, выделяются следующий набор аргументов: arg0 – одушевленный или неодушевленный субъект движения; arg1 – конечная точка движения; arg2 – начальная точка движения; arg3 – пересекаемое пространство; arg4 – средство транспорта, которое используется для движения; arg5 – лицо, которое направляется при движении; arg6 – объект, который переносится при движении и т. п. Нумерация аргументов показывает, насколько часто признак пропозиции уточняется в контекстах членов синсета, относящихся к данному дереву.

Семантические ограничения, накладываемые на заполнение валентной позиции задаются в блоке (sem_data) путем отсылки на семантические деревья RussNet, при этом задается тип отсылки (TYPE): значение "top" указывает на вершину дерева (например «человек»); "group" задает стандартную группировку деревьев (например, «одушевленный»), значение "synset" является отсылкой к определенному синсету RussNet. Параметр ID конкретизирует адрес в структуре тезауруса.

Аналогом данного блока являются уточнения в скобках в традиционных словарных описаниях, например: 'двигаться, вращаясь (о круглых предметах'. Однако реальные контексты употребления слов зачастую дают более широкий спектр возможностей для заполнения валентной позиции. Например, для значения *катиться1* позицию субъекта заполняют не только *шар, колобок, клубок, колесо*, но и *камни, тела людей, булыжник, комок теста*.

Грамматическая спецификация валентности (*morph_data*) включает указание части речи (возможны объединения) и значения грамматических категорий, которые существенны для оформления позиции (например, предложно-падежная форма существительного, видовая характеристика инфинитива, разряд наречий и проч.). Среди многообразия способов выражения валентности выбираются те варианты (variant), которые имеют статистическую устойчивость. Низкочастотные заполнения валентной позиции будут обсуждаться ниже, при описании схемы наследования валентностей.

Параметр позиции (*place*) используется в том случае, когда валентная позиция устойчиво (>50%) в контекстах занимает позицию, которая не совпадает с нейтральным порядком слов в словосочетании или предложении.

3.2. Схема наследования рамок валентностей

Исследовательской задачей настоящей работы является описание того, как и в какой степени параметры рамок валентностей наследуются в семантических деревьях RussNet в синсетах, связанных родовидовым отношением (синсет A – гипероним, B – гипоним). Мы рассматриваем данную проблему применительно к трем разным деревьям: глаголам движения, глаголам принятия положения в пространстве и глаголам изменения местоположения объекта. Для иллюстрации описываемой схемы на рис. 2 приведен фрагмент дерева глаголов движения.

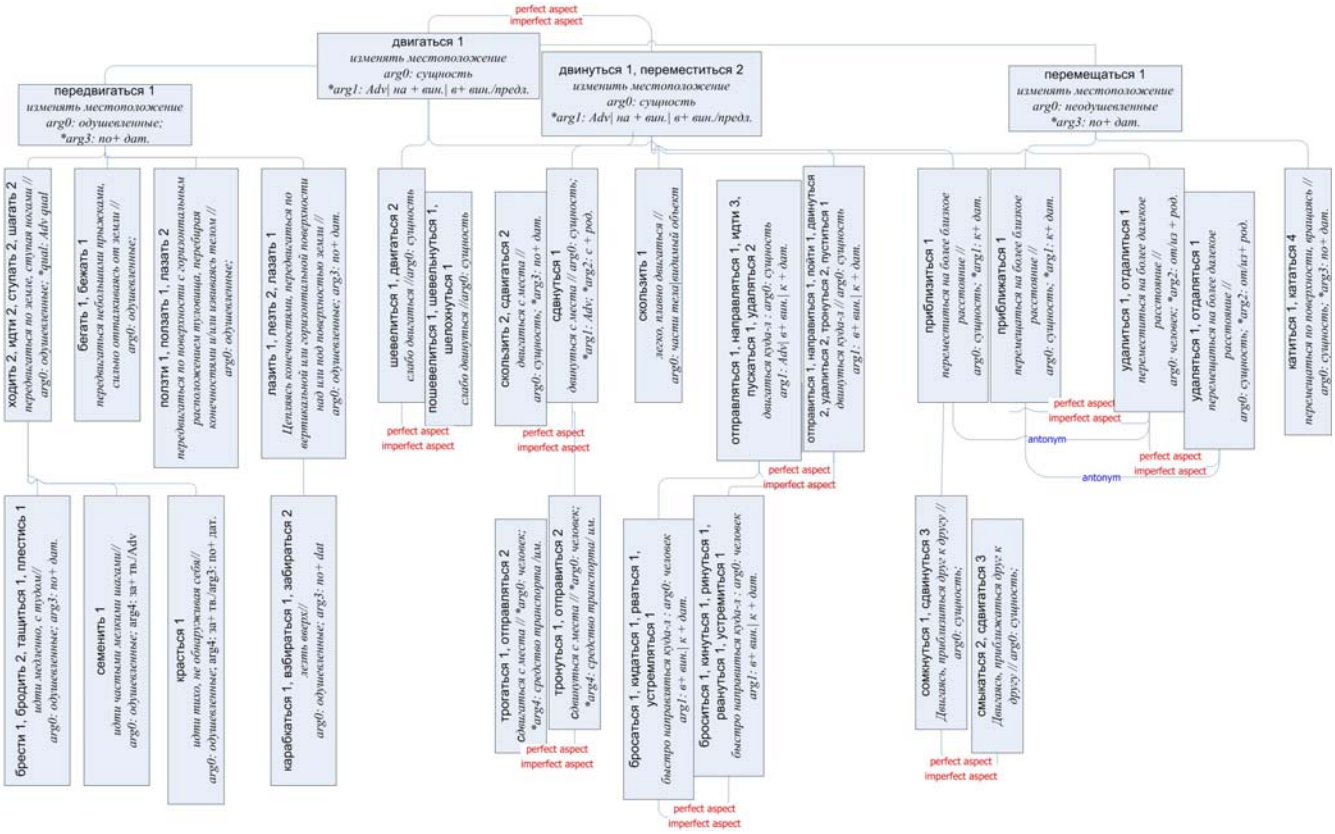

Рис. 2. Фрагмент семантического дерева глаголов движения

Фрагмент семантического дерева представлен в виде двух параллельных гипонимических структур для глагольных синсетов совершенного/несовершенного вида (СВ/НСВ). Соотносимые по значению глаголы связаны семантико-грамматическим отношением perfect/imperfect aspect. Гипонимические структуры не полностью идентичны, поскольку некоторые значения синсетов чаще реализуются в корпусе глаголами НСВ, а другие – СВ. На рисунке частотно доминирующий синсет слегка «приподнят» по отношению к синсету с видовыми парами. При определении «доминирования» учитывается, что общее число глаголов несовершенного вида примерно в 1,5-2 раза превышает число глаголов совершенного вида в корпусе (что объясняется грамматической маркированностью значения совершенного вида). В отдельных случаях, например для синсета {*трогаться1, отправляться2*}, частота употребления глаголов НСВ столь незначительна по сравнению с их видовыми коррелятами 0,1 vs 7 ipm, что они присоединяются в тезаурусе лишь аспектной связью к синсетам коррелятов {*тронуться1, отправиться2*}. Такой же способ присоединения будет использован для «потенциальных» видовых глагольных пар, которые не были зарегистрированы в корпусе, но которые могут встретиться окказионально при обработке текста. Что касается схемы наследования рамок валентностей для синсетов видовых коррелятов, то они регулярно совпадают. Однако при низкой частотности членов синсета полная картина дистрибуции контекстных маркеров бывает не видна. Опираясь на адекватно представленные дистрибуции синсетов разных видовых пар, мы считаем, что валентная рамка более частотного видового синсета наследуется низкочастотным коррелятом.

Будем рассматривать наследование рамок по отдельным параметрам, которые будут указываться в подзаголовке.

(I) Обязательность/факультативность (obligatory)

Параметр факультативности валентности обозначен на рисунке звездочкой. Наследование этих параметров проходит по-разному в субъектной позиции и остальных аргументных позициях.

(Ia) Субъектная позиция (arg0)

В субъектной позиции возможны два типа соотношения между значениями параметра obligatory для гиперонима A и гипонима B (после слеша указан процент реализации соотношения в рассмотренных деревьях):

(1) A: obligatory="yes" => B: obligatory="yes" / 99%

(2) A: obligatory ="yes" => B: obligatory ="no" / 1%

Процент сохранения обязательности субъектной позиции является практически абсолютным. Ослабление субъектной валентности происходит при появлении другого «агентивного» участника в рамке, например для глагола *везти1* 'двигаясь, перемещать кого-/что-либо каким-либо средством транспорта' (*В машине я спросил, куда меня везут? Он не хотел, чтоб меня везли в другую больницу*). При этом формы 1-2-го л. дают возможность установить субъекта действия, хотя он и не задан в поверхностной структуре, ср. *Важных гостей везу!* (arg0: «я»).

(Ib) Несубъектные позиции (arg1…)

В других аргументных позициях факультативной валентности гиперонима может соответствовать обязательная валентность гипонима. Например: *передвигаться1 – лазить1* (arg3: 40% vs 80%).

(3) A: obligatory ="no" => B: obligatory ="yes" / 43%

Факультативная валентность гиперонима может преобразоваться в окказиональную, которая не указывается в рамке валентности, например: *передвигаться1 – ползти1* (40% vs 20%).

(4) A: obligatory ="no" => B: no valency / 44%

Сохранение параметра обязательности аргументной позиции встречается окказионально в рассматриваемых деревьях:

(5) A: obligatory ="X" => B: obligatory ="X" / 13%

Вариант соотношения (4) значения параметра обязательности в рамках гиперонима-гипонима может быть использован в процедуре разрешения неоднозначности для идентификации низкочастотных «следов» валентностей гиперонима в контекстах гипонима. Еще один важный аспект наследования схемы валентностей – то, что для низкочастотных синсетов в структуре RussNet, у которые нет в корпусе достаточно данных для задания рамки валентностей, ее можно экстраполировать по рамке гиперонима, используя (3).

(II) Семантические ограничения (sem_data/"group")

Первоначально было замечено, что семантические группировки рамки прилагательного-гиперонима являются тождественными или более общими, чем у гипонима (Azarova, Sinopalnikova 2004). Это положение вполне подтверждается для рассмотренных деревьев. В частности, вершиной дерева глаголов движения глагол можно выбрать глаголы *двигаться (двинуться), передвигаться*, *перемещаться (переместиться)*, в корпусе соответствующие значения представлены в объеме 42 ipm[4] (10 ipm), 12 ipm (1,3 ipm), 7 ipm. В словарном определении МАС для *двигаться* 'совершать движение, передвигаться, перемещаться', глагол представлен как синоним *передвигаться*, *перемещаться*. Однако различия в частотности глаголов подкрепляются разными типами заполнения позиции arg0. Для *двигаться* это любой наблюдаемый объект: человек, группа людей, средства транспорта, части тела человека, животные, механизмы, естественные объекты, то есть группировкой "RUS-nEntity". Соотношение одушевленных и неодушевленных заполнений позиции 3:2. Для глагола *передвигаться* эта группировка сокращается до RUS-nAnimate (человек, животные), а для *перемещаться* – до RUS-nInanimate (средства транспорта, естественные объекты, механизмы, части тела человека). Поскольку эти соотношения носят статистический характер, нельзя сказать, что не бывает противоположных случаев, но они укладываются в границы окказиональных флуктуаций. Более того, даже если речь идет о человеке при глаголе *перемещаться* контекст весьма характерно представляет действие, например: *Гора мышц и мускулов перемещалась на столе*. Таким образом, в отношении параметра sem_data чаще наблюдается схема наследования, реже – сужения:

(6) A: sem_data=X => B: sem_data=X / 85%

(7) A: sem_data=X => B: sem_data=Y, X⊂Y / 15%

(III) Морфологическое оформление (morph_data)

Для данного параметра соотношения рамок гиперонима-гипонима зависят от аргументной позиции.

(IIIa) Субъектная позиция (arg0)

В субъектной позиции наблюдается доминирование схемы наследования формы именительного падежа

(8) A: CASE="nom" => B: CASE="nom" / 100%

(IIIb) Несубъектные позиции (arg1…)

Наследование грамматического оформления аргументной позиции, помимо субъектной, носит окказиональный характер (например *направиться вперед, в сторону, к дому* & *идти вперед, в сторону, к дому*), причем наследоваться может схема не только непосредственного гиперонима (ср. *передвигаться по залу* & *красться по коридору*).

(9) A:variants = B:variants / 23%

Чаще всего грамматические варианты оформления валентности гипонима представляют собой пересечение с вариантами гиперонима. Например, валентность конечной точки движения (arg1) для синсета {*двигаться1*} оформляется наречием; предложно-падежной конструкцией «в + В./П.п.» или «на + В.п.» (*двигаться вперед, в сторону, в направлении, на север*); два варианта – наречие и конструкция «в + В.п.» – совпадают с морфологическим оформлением гипонима {*отправиться1, направиться1, …*}, при этом у гипонима есть и другие частотные варианты (*направиться вперед, в сторону, к дому*). остальные варианты различаются.

(10) A:variants ∩ B:variants ≠ ∅ / 70%

Существенно реже наблюдается сокращение морфологических вариантов оформления аргументной позиции гиперонима у гипонима (ср. *направиться вперед, в сторону, к дому* vs *броситься в сторону, к дому*).

(11) A:variants ⊂ B:variants / 7%

Следует отметить, что даже в тех случаях, когда морфологическое оформление валентности у гипонима-гиперонима не наследуется частотно, на уровне окказиональных употреблений они составляют приблизительно один и тот же набор. Поэтому у корневых синсетов {*двигаться1*} наблюдается самый широкий диапазон

[4] ipm (items per million) – единиц на 1 миллион словоупотреблений в корпусе.

окказиональных вариантов грамматического оформления валентности, которые тем не менее столь разнообразны и малочисленны, что не представляется возможным их перечислить. Поэтому более реальным является перечисление частотных морфологических вариантов у гипонимов, но тогда надо признать наследование в обратном направлении, как «просачивание» морфологической формы аргументной позиции от гипонима к гипериониму.

В качестве частного замечания относительно морфологического оформления хотелось бы отметить, что среди предложных вариантов оформления валентностей регулярно встречаются многозначные исконные предлоги *к, в, по, на, с* в отличие от мотивированных предлогов типа *вдоль, сквозь* и т. п. Этот факт интерпретируется следующим образом: «семантически определенные» предлоги относительно самостоятельны, их интерпретация в отношении аргументной структуры более-менее однозначна, кроме того, их использование отчасти факультативно: они задают «фокусное» (конкретизированное) заполнение аргументной позиции. Такую же точку зрения высказывала Е.С. Скобликова (Скобликова 1990, с. 87), указывая на синонимию предлогов: *у стола, около стола, возле стола, подле стола, рядом со столом, близ стола, недалеко от стола*.

(IV) Набор аргументных позиций

В отношении набора заполненных аргументных позиций полное наследование встречается на уровне факультативности (52%), а с учетом наследования не только от непосредственного гиперонима доходит до уровня обязательности (66%). Все остальные случаи подпадают под соотношение типа «пересечения» (типа 10), если учитывать субъектную позицию (arg0).

Появление новой аргументной позиции регулярно соотносится с наличием в морфемной структуре глагола аффикса (чаще префикса) (см. *сдвинуться1*). Можно предположить, что сходные наборы аргументов указывают на однотипность семантической структуры глаголов в каких бы частях семантического дерева они ни находились (ср. *семенить1* и *красться1*).

Контексты слов общей семантики типа *двигаться* дают максимальный набор возможных аргументов, правда, контекстные маркеры столь низкочастотны и разнообразны, что возникают проблемы как с идентификацией типа аргументов, так и с их нумерацией. Более реальным является исчисление аргументов по рамкам семантического дерева, порядок их нумерации определяется частотой употребления аргумента в дереве. Контекстные маркеры, которые не попали ни в одну рамку валентности семантического дерева, являются сирконстатами, они определяют общую ситуацию безотносительно к специфике его значений. Естественно, что аргументы одного дерева могут быть сирконстантами для другого. Между списками аргументов можно установить соответствия.

4. Заключение

Описанная выше схема наследования рамок валентностей в тезаурусе RussNet может уточнить процедуру разрешения неоднозначности в проекте ИДЕОГРАФ и процедуру семантического анализа в терминах аргументных структур пропозиций.

Рассмотрение схемы наследования валентных рамок показывает, что наследуются лишь отдельные параметры, а не вся структура целиком. Расширяя набор обследованных деревьев RussNet, мы в дальнейшем постараемся уточнить описанную схему.

Список литературы

1. *Advances in Probabilistic and Other Parsing Technologies / Blunt H., Nijholt A. (eds.) Kluwer Academic Publishers, 2000.*
2. *Azarova I. The matching of AGFL subcategories to Russian lexical and grammatical groupings // Proceedings of the Second AGFL Workshop on Syntactic Description and Processing of Natural Language. Radboud University Nijmegen, the Netherlands, www.cs.ru.nl/agfl/papers/ 2002.*
3. *Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I. RussNet: Building a Lexical Database for the Russian Language // Workshop on WordNet Structures and Standardisation, and how these affect Wordnet Application and Evaluation. 28th May 2002. Las Palmas de Gran Canaria, 2002. P. 60–64.*
4. *Azarova I., Sinopalnikova A. Adjectives in Russnet // Proceedings of the Second International WordNet Conference, GWC 2004, Brno, Czech Republic, January 20–23, 2004. P. 251–259.*
5. *Beinema P., Koster C.H.A. AGFL Grammar Work Lab: Manual for the AGFL system. URL: http://www.cs.ru.nl/agfl/ papers/manual.pdf. 2004. 62 p.*
6. *Carpenter B. The Logic of Typed Feature Structures. Cambridge University Press, Cambridge, 1992. 270 p.*
7. *Kay M. Parsing in Functional Unification Grammar // Readings in Natural Language Processing / B. J. Grosz, K. Spark Jones & B. L. Webber (ed.) Morgan Kaufmann Publishers, Inc.: Los Altos, California, 1986. P. 125–138.*
8. *Koster C.H.A. Affix Grammars for natural languages // Attribute Grammars, Applications and Systems, International Summer School SAGA, Prague, Czechoslovakia, June, 1991.*
9. *Leacock C., Chodorow M. Combining Local Context and WordNet Similarity for Word Sense Identification // WordNet: An Electronic Lexical Database / C. Fellbaum (ed.) MIT Press, 1998. P. 265–283.*
10. *Makino T., Torisawa K. and Tsujii J.* ***LiLFeS – Practical Programming Language For Typed Feature Structures*** *// Proceedings of Natural Language Pacific Rim Symposium '97". 1997.*

Pantel P., Lin D. Word-for-Word Glossing with Contextually Similar Words // Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, May 27 – June 1, Edmonton, 2003.

11. *Pollard C. & Sag I. Head-Driven Phrase Structure Grammar. Chicago: University of Chicago Press, 1994. 440 p.*
12. *Voorhees E.M. Using WordNet for Text Retrieval // WordNet: an Electronic Lexical Database / Ch. Fellbaum (ed.) MIT Press, 1998. P. 285–303.*

Азарова И.В. Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003 (Протвино, 11-16 июня 2003 г .) М., 2003. С . 51-55.

Азарова И.В., Митрофанова О.А., Синопальникова А.А. Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003 (Протвино, 11-16 июня 2003 г .) М., 2003. С . 43-50.

Азарова И.В., Секликов Ю. В., Иванов В. Л. Интерпретация текстовых документов с использованием формальной грамматики AGFL и компьютерного тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 (“Верхневолжский”, 2-7 июня 2004 г .) М., 2004. С. 1-6.

Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 (“Верхневолжский”, 2-7 июня 2004 г .) М., 2004. С. 542-547.

Скобликова Е.С. Очерки по теории словосочетания и предложения. Изд-во Саратовского университета: Куйбышевский филиал. 1990.