Международная конференция «Диалог 2009»

Круглый стол «*Net — лексические ресурсы в Интернете» (ведущая Н.В. Лукашевич)

Н.В. ЛУКАШЕВИЧ (НИВЦ МГУ): Название круглого стола сформулировано общим образом и охватывает не только WordNet и FrameNet, которые вчера обсуждались, но и еще какие-то ресурсы для исследований в области компьютерной лингвистики.

Какие вопросы можно обсудить? Во-первых, какова значимость таких общественно доступных словарных ресурсов — понимая, что они не идеальны, что они не отвечают всем нашим потребностям, были сделаны не конкретно для нашего применения. Другой вопрос, который может быть интересен, заключается в следующем: если реально ресурс создается чаще всего на английском языке и для английского языка, и нам, исследователям, интересен этот ресурс, и мы хотим иметь что-то такое для русского языка. Как правильнее поступать? Стараться делать так, как сделано на английском языке, и тем самым примыкать к некоторому сообществу, или учиться видеть какие-то ошибки и недочеты, исправлять их и тем самым делать следующие шаги? В России не очень много бесплатных ресурсов в области компьютерной лингвистики. Что должно измениться, чтобы таких ресурсов стало больше? Кто может их делать, и что нужно учесть при их создании? Более частные вопросы касаются конкретных ресурсов — может быть, у кого-то есть какой-то опыт, и он тоже поделится с нами опытом положительного общения с WordNet, FrameNet и другими ресурсами? Какие недостатки и не упомянутые мной во вчерашнем докладе достоинства этих словарей вы можете указать?

Повторю свои выводы из доклада. Я говорила о проблеме WordNet'а, говорила об описании единиц, что в приложениях вызывает проблемы раздельное описание частей речи, отдельные синсеты для разных наименований одной и той же сущности, недостаточная различимость синсетов — вот, мы, может быть, узнаем, как в WordNet'е различаются джемпера, пуловеры и свитера — есть у нас такая слишком большая многозначность. Проблема с описанием отношений. И, может быть, кому-то нравится или не нравится этот сайт. Значение WordNet'а, независимо от того, согласны мы, что он применяется в приложениях, или несогласны, но по крайней мере он дает нам возможность обсудить, как надо было делать, в чем были ошибки, и, может быть, если мы хотим делать похожие ресурсы — не наступить на те же грабли. И самые разные применения WordNet'а, которые связаны с созданием каких-то дополнительных ресурсов. Может быть, кто-то скажет про конкретные приложения и роли ресурсов в них.

В.П. СЕЛЕГЕЙ (ABBYY Software House): Я просто воспользовался правом члена программного комитета, потому что мы обдумывали, какие доминанты должны быть представлены на «Диалоге», и лексические ресурсы показались нам наиболее очевидной темой, которая объединяет интересы всех присутствующих.

Хотелось бы еще несколько расширить пространство обсуждения, включить в него не только WordNet, но и что-то еще. Поскольку у нас уже была секция, посвященная лингвистическим ресурсам, что-то уже было сказано, и, может быть,

пару слов стоит сказать о некоторых вопросах, которые возникли при слушании тех докладов.

Поскольку мы на конференции по компьютерной лингвистике, нас интересует достаточно полное формализованное описание, и чем эти ресурсы глубже, тем лучше. Но мы видим, что когда докладчики представляют свои ресурсы, они не всегда четко, как мне кажется, осознают или не всегда четко доводят до аудитории, о какого рода ресурсе идет речь. Понятно, что на «Диалоге» представляются ресурсы трех типов, и когда мы говорим о критериях ресурсов — хорош он или плох, все определяется тем, для чего он предназначен. **Идеологические** ресурсы довольно часто бывают представлены, то есть ресурсы, которые нацелены на демонстрацию работоспособности определенной теории. Адепты этой теории могут вкладывать серьезные усилия, чтобы создать такой ресурс, и такие ресурсы существуют, в том числе в Интернете. Есть ресурсы **функциональные**. Понятно, что разработчик любой крупной системы машинного перевода делает свое крупное семантическое описание, но, вообще говоря, это лексико-семантическое описание мало интересует пользователя, его интересует конечный результат. Поэтому предъявлять какие-либо претензии к тому, как устроено лексико-семантическое описание довольно глупо, потому что вам либо нравится, как оно работает, либо не нравится. И, наконец, есть третий тип ресурсов, которому и посвящено наше сегодняшнее обсуждение, — у Н.В. Лукашевич они называются **общественно-доступными**. Они бы могли быть не общественно-доступными, но важно, что они рассчитаны на многоцелевое применение большим количеством людей. К таким ресурсам предъявляются совершенно определенные критерии. На многих европейских конференциях, посвященных лингвистическим ресурсам, приглашенные докладчики начинают с того, что говорят: «Ну понимаете, что лексические ресурсы создаются такими колоссальными усилиями многих людей, что всякое неправильное решение влияет на успех или неуспех каких-то разработок». То есть как важно разрабатывать ресурс, в который вкладываются такие колоссальные средства и который предназначен для многоцелевого использования, как важно, чтобы исходные принципы были адекватными, и через год или два не оказалось, что действует инерция неправильного решения, когда вы продолжаете вкладывать в неудачную модель усилия людей.

Понятно, что универсальные лексические ресурсы образуют некую иерархию. С одной стороны, мы все знакомы с корпусами, и постоянно возникает вопрос, как сделать эти корпуса более функциональными или более богатыми с точки зрения приложений. И тут же возникает вопрос разметки — синтаксической разметки, семантической разметки. И хотя современные корпуса позволяют делать целую систему разных разметок на одном текстовом материале, и это считается правильным подходом к созданию корпуса, тем не менее для пользователя очень важно, какого рода семантическую и синтаксическую разметку вы предложили. То есть сами лингвистические технологии являются ресурсом. Вот WordNet является таким ресурсом, потому что в работе по компьютерной лингвистике он выступает как инвентарь лексических значений. Другого общепризнанного инвентаря лексических значений не существует. Нравится он вам или нет, но если вы хотите, чтобы ваши коллеги вас понимали, Наташа не даст соврать — она участвовала во многих конференциях, посвященных лексической неоднозначности… Вообще сами задачи разрешения неоднозначности зачастую формулируются в системе семантических понятий WordNet’а. Так же обстоит дело и с FrameNet’ом. То есть понятно, что если мы такие ресурсы разрабатываем, то волей-неволей, для

молодого исследователя, который приходит в эту область, для него это и есть лингвистика.

Когда мы на «Диалоге» слушаем доклады, посвященные лингвистическим ресурсам, всегда хочется, как мне кажется, чтобы какие-то требования были соблюдены. Ну глубина описания — понятно, концентрическая полнота — тут я пользуюсь словом, которым Сергей Константинович Крылов использовал некоторое время назад на «Диалоге», то есть это действительно необходимо для развития такого рода лексических интернет-ресурсов. Неважно, как вы определите этот концентрический принцип, будет ли он частотным, будет ли он каким-то девиационным. Понятно, что ресурс должен обеспечивать полноту относительно критерия, который вы задали.

Мы слышали доклад М.В. Яворской про RussNet, там назывался МАС (Малый академический словарь). Уже хорошо, что исследователи не выдумывают систему значений из головы, а рассматривают что-то в качестве базового ресурса. Правда, всякие девиации относительно этого МАСа — нужно понимать, есть они или нет, можно их увидеть в самом ресурсе, чтобы они были ясно формализованы. На самом деле, для всяких верифицирующих процедур, для обучения очень важно понимать, на каком лексико-семантическом материале вы строили новое колоссальное семантическое описание. Создавая лексический ресурс, вы обязательно связываете его с определенной системой понятий и передаете пользователю этот ресурс вместе с этой системой понятий. WordNet — это не только словарь, но и определенная концепция, которая всем нам сегодня предложена, и теперь она воспроизводится в новых языках, в том числе в русском. И я не знаю, какова степень свободы, когда вы называете что-то WordNet'ом, до какой степени вы связаны, какова конвенция с разработчиками WordNet'а. Вот в UNL о котором мы тоже слышали в докладе про ресурсы, — там ясно, какова конвенция, и понятно, является ли UNL ресурсом третьего типа — универсальным. И интересно, как воспринимают UNL люди, которые работают на других платформах.

Вот, собственно говоря, вопрос, который я предлагаю добавить в число обсуждаемых, — это те лексические ресурсы, о которых мы услышали? — готовы ли они на то, чтобы стать стандартом для решения задач в области компьютерной лингвистики и не только.

И последний вопрос — каково соотношение между лингвистикой и комьпютерной лингвистикой. WordNet — это что? Это лексическая семантика. Считаем ли мы этот ресурс адекватным с точки зрения того, что мы знаем о лексической семантике, как считают лингвисты? Считайте, что это вопрос от программного комитета, потому что «Диалог» — это взаимодействие компьютерных лингвистов и лингвистов формальных. Очень интересно, какова наша оценка этих ресурсов, можно ли сказать, что разработчики WordNet'а базируются на какой-то глубокой лексико-семантической концепции. Я думаю, что скорее нет.

Н.В. ЛУКАШЕВИЧ: Я много раз слышала, что английские и американские лексикографы очень ворчат по этому поводу.
Давайте послушаем представителей русского WordNet'а.

М.В. ЯВОРСКАЯ (СПбГУ): Меня озадачили глобальным образом, постараюсь ответить, на что смогу, но для начала чуть-чуть опущу планку и начну с малого и

коварно соглашусь с Натальей о некоторых ограничениях, которые налагает WordNet на меня как на разработчика. Я совершу некоторую самоубийственную, на мой взгляд, вещь, покажу группу, которая не до конца обработана и в связи с этим вызывает некоторый легкий ужас, поскольку мне же сейчас ее делать. Это группа цвета. [Демонстрирует слайд.] Как я вчера рассказывала, по всей видимости, существительное «цвет» будет гиперонимом. Там, где есть номер значения — «белый 1», «черный 1», это значит, прилагательное обработано, а остальные еще нет. И сразу начинаются проблемы. По Вежбицкой, например, прилагательные «белый» и «черный» очень сильно связаны с цветом. По MAC они определяются другим образом, поэтому я в данном случае беру такую иерархию, и получается трихотомия: «белый», «черный» и «цветной», который в MACe формулируется как ‘наделенный цветом, не белый и не черный’. Тут очень сложная система строиться начинается. Мы встречаемся сразными характеристиками, которые грозят преобразовать более-менее строгое дерево в сеть. Понятно, что у «цветного» будет куча гипонимов с конкретными названиями цветов, начинаются совместные гипонимы типа «серый», который тоже определен как нечто среднее между «белым» и «черным», и, возможно, у «белого» и «черного» тоже появятся какие-то гипонимы.

Дальше в тот же самый цвет, но сверху, мы видим, что существует такая характеристика, как «насыщенность». И опять начинается проблема, потому что, возможно, я встречусь с определением, которое будет гласить, что, допустим, цвет глубокий и зеленый одновременно. Дальше начинается та самая категория света. Появляется значение «белый 2» как самый светлый, и такая категория, как «пропускающий свет — не пропускающий свет». Поскольку мы говорим о перцептивных прилагательных, то она влияет. Дальше мы встречаемся с градуальностью в цвете, и вот у меня вопрос — «светло-зеленый» и «темно-зеленый» — куда это все девать? В результате получается, что количество стрелок, входящих в одно и то же прилагательное, несколько превышает то, что обычно ожидают от WordNet’a. В данном случае мне становится немного тесно в этой структуре, когда я пытаюсь реально описать, с чем сталкиваюсь. Как соединиться цвет, свет и насыщенность в этой системе, для меня пока совершенно непонятно, и я надеюсь, что, может быть, полная обработка всех этих прилагательных даст мне какой-то ответ.

Что касается замечания к докладу Натальи. Во-первых, уже даже в английском WordNet’e существуют взаимоотношения между частями речи, в RussNet’e они тоже есть.

Теперь поговорим о плюсах. Я считаю, что WordNet — это прежде всего инструмент, лежащий в открытом доступе. Как любой инструмент, его можно не использовать, использовать, использовать более оптимально или менее оптимально. Если кому-то что-то не нравится — пожалуйста, всегда можно что-то добавить.

РЕПЛИКА ИЗ ЗАЛА: А RussNet не лежит!

М.В. ЯВОРСКАЯ: Лично я за доступность RussNet’a. Там лежит объективное опасение выкладывать столь маленький, непредставительный, недоделанный продукт. На мой взгляд, его лучше выложить и дополнять, но в данном случае меня пока еще никто не спросил. В связи с тем, что это инструмент, можно сказать, что, во-первых, он в связи с тем, что мы в RussNet вводим валентности и

актанты для глаголов, то он несколько сближается с FrameNet'ом, хотя я сейчас вряд ли готова сказать, насколько они становятся от этого похожи. Безусловно, он может являться базой для создания онтологии. Как связаны WordNet'ы между собой — вы, наверно, знаете: существует Interlingva Index (ILI) как метаязык. В действии я его еще не видела, но как сверхцель он, тем не менее, у всех стоит. Также мы все слышали, что WordNet используют как базу для создания семантического языка, ну и, как Наталья сказала, при информационном поиске не мытьем так катаньем его все-таки научились худо-бедно использовать.

Н.В. ЛУКАШЕВИЧ: И для исследовательской работы.

М.В. ЯВОРСКАЯ: Да, это тоже можно.

ВОПРОС ИЗ ЗАЛА: Вот видно ли, что между «светло-зеленый» и «темно-зеленый» одно и то же отношение?

М.В. ЯВОРСКАЯ: Я в своем докладе говорила, что гипонимия бывает трех типов — похожая на существительное, похожая на тропонимию глаголов и смешанная для. В данном случае это никак не различается, поскольку, если мы еще будем создавать дерево возможных гипонимий, то это еще более усложнит задачу.

В.П. СЕЛЕГЕЙ: А вот есть ли какая-то методика создания WordNet'а, общая для всех разработчиков в разных странах? Мы сейчас обсуждаем вопросы, которые кажутся основополагающими для такого рода ресурсов. Глубина филиации, где остановиться, где критерии? Я верю, что вы очень хороший лингвист, вы решаете это по своей интуиции, опираясь на МАС. У вас группа из трех, может быть, человек, вы как-то между собой это решили и подарили человечеству, с большим риском, что человечество не согласится. Вы об этом как-то думатее?

М.В. ЯВОРСКАЯ: Думаем ли мы обо всем человечестве? Думаем, безусловно. Есть формальный признак, к которому должны прийти все WordNet'ы — они все должны встроиться в этот ILI. Я практически уверена, что, допустим, испанский и итальянский WordNet вместе не обсуждали ничего.

Н.В. ЛУКАШЕВИЧ: По-моему, этот ILI базируется на старой версии WordNet'а, типа 1.6, а сейчас уже есть WordNet.3.

В.П. СЕЛЕГЕЙ: Я знаю оценку китайского WordNet'а самими китайцами, и это очень низкая оценка. Это переведенный на китайский язык английский тезаурус.

М.В. ЯВОРСКАЯ: Если говорить о второй версии WordNet'а, которая обладает этими 122 или сколько там тысячами, то Сергей Яблонский пошел по пути перевода английского на русский.

В.П. СЕЛЕГЕЙ: То есть у нас как минимум два WordNet'а?

М.В. ЯВОРСКАЯ: У нас три WordNet'а. Один в Москве.

В.П. СЕЛЕГЕЙ: Три, четыре, пять, и каждый может называть себя русским WordNet'ом? Есть ли какие-то формальные требования, может быть, подписать какую-то конвенцию, войти в какой-то комитет…

Н.В. ЛУКАШЕВИЧ: На самом деле, в Италии тоже два WordNet'а. Это просто вопрос менеджмента.

В.П. СЕЛЕГЕЙ: То есть WordNet — это такая абсолютно десинхронизированная структура, и в каждой стране можно создать сколько угодно WordNet'ов, и каждый будет иметь право на существование?

М.В. ЯВОРСКАЯ: Да.

Д.О. ДОБРОВОЛЬСКИЙ (ИРЯ РАН): Если я правильно понял, то сами семантические категории, которые являются основой WordNet'а, они никак не обсуждались. И, скажем, те семантические категории, которые будет выделены в русской версии, никак не настроены на английские или немецкие.

М.В. ЯВОРСКАЯ: Ну, существует top onthology, как это называется в WordNet'е.

Д.О. ДОБРОВОЛЬСКИЙ: Она как бы является обязательным компонентом, и ее соблюдают все?

М.В. ЯВОРСКАЯ: На нее все равняются, да, безусловно, хотя бы потому, что такой уровень абстракции…

В.И. БЕЛИКОВ (ИРЯ РАН): «Равняются» — надо буквально понимать? Или «следуют»? Стропроцентно следуют, или только равняются?

М.В. ЯВОРСКАЯ: Вот клясться не буду, но по крайней мере были разговоры, что мы держим руку на пульсе.

Д.О. ДОБРОВОЛЬСКИЙ: Просто единственный, но самый естественный путь использования такого ресурса, как мне кажется, — это всякие сопоставительные исследования. То есть если я знаю, что для группы языков, которые мне интересны, есть такой ресурс, я могу взять ту или иную семантическую категорию, которая мне интересна, и посмотреть, как она наполнена в том или ином языке, и сделать какие-то выводы. А если оно сделано как-то не совсем так, если все эти категории не со всем совпадают, то этот путь закрывается напрочь.

ВОПРОС ИЗ ЗАЛА: Кем закрывается?

М.В. ЯВОРСКАЯ: Закрываются — совсем другим образом. Например, тот же испанский WordNet закрыт, он за деньги.

Д.О. ДОБРОВОЛЬСКИЙ: Но мне хочется знать, за что я буду платить.

М.В. ЯВОРСКАЯ: Хорошо. Я как разработчик, который не собирается платить деньги, не могу взять испанский WordNet и посмотреть, что из этого получается.

Д.О. ДОБРОВОЛЬСКИЙ: Но хоть с английским. Если английский — как бы мать, то все дочери смотрят на мать…

Н.В. ЛУКАШЕВИЧ: А можно, я здесь поясню? На самом деле, в разных WordNet'ах разных языков было принято две основные концепции. Либо держится исходным некоторый базовый набор синсетов, а все остальное делается в соответствии, как

с языком, и может пойти совсем другим образом, чем в английском языке, хотя каждый разработчик своего WordNet'а все-таки смотрит, что сделано в английском WordNet'е. Но много представителей другого направления, когда берется английский WordNet' как он есть, применяется некоторые автоматизированные, часто автоматические процедуры по компьютерным словарям, и как-то это переводится. И это как раз два других WordNet'а, один из которых в Новосибирске, а один в Санкт-Петербурге. Я не думаю, что это кого-нибудь порадует, хотя сходство с английским очень тесное, но то, что там получается в результате автоматического применения словарей, очень страшно и никого не порадует.

Д.О. ДОБРОВОЛЬСКИЙ: Это никому не нужно, это понятно. Но понятно, что, наверно, неправильно делать каждую версию — русскую, английскую, немецкую, французскую — изолированно. Конечно, можно говорить о том, что каждый язык уникален, и культура накладывает ограничения, и семантические категории не совпадают, и, как мы знаем по работам Вежбицкой, в языках есть эмоции, о которых мы не догадываемся, и все это продукт культуры. Но все-таки, по большому счету, много чего совпадает. Если у меня, допустим, есть категория «растения» или «мебель» или глагол движения, все-таки хочется быть уверенным, что я могу их сопоставлять.

М.В. ЯВОРСКАЯ: С глаголами движения все-таки тяжело будет. Про мебель — я могу вас уверить, что вам все понравится.

А.С. НАРИНЬЯНИ (НИИ Искусственного интеллекта): Мне кажется, разговор приобретает несколько абстрактный характер. Обсуждается русская версия WordNet'а. И что мы будем сейчас обсуждать — можно будет ли объединить, навязать что-то испанцам, вообще собрать весь мир, как детей лейтенанта Шмидта, и вообще договориться? Гораздо интереснее вопрос: а те три группы, которые есть сейчас в России…

Н.В. ЛУКАШЕВИЧ: Они не договорились.

А.С. НАРИНЬЯНИ: Они не только не договорились, они и не будут договариваться. Потому что они действительно взяли вот этот лобовой… видимо там закрыли какой-то грант, получили зарплату…

Н.В. ЛУКАШЕВИЧ: Нет, они сами.

А.С. НАРИНЬЯНИ: Сами? Тогда не буду вешать, по Вежбицкой, какую-нибудь эмоциональную характеристику того, что они делали. Потому что как в прошлый раз говорили — масскультура это масскультура: захотели — и вспахали, например, Манежную площадь и посадили лимоны. Так захотелось, и все. Здесь мы видим все-таки осмысленный проект, который мы тут обсуждаем. У него есть свои плюсы и минусы. Может быть, учитывая такой важный фактор здесь, надо создать какую-то рабочую группу?

Н.В. ЛУКАШЕВИЧ: Помогите материально, как говорится.

А.С. НАРИНЬЯНИ: Мы живем в эпоху, когда, как я говорил, на «Спартак» денег хватает, и то в обрез. А уж на какой-то WordNet — кому он там нужен? Но мы-то тут есть, и нам он нужен, раз мы тут сидим и это обсуждаем. Рабочая группа —

это минимум, что мы можем сделать. Либо мы считаем, что это общая задача достаточна важна и образуем рабочую группу, которая собирается раз в месяц, раз в квартал. Тогда можем до какого-то стандарта договориться. А те, кто хотят это переводить калькой с английского, китайского или суахили — ну пускай переводят. Это не наука, а масскультура.

О.Н. ЛЯШЕВСКАЯ (Университет Тромсё, Норвегия): Я хотела бы начать свой рассказ о разных лексических ресурсах с того слайда, который я показывала в прошлом году: разные системы, которые в последнее время развиваются. Здесь было перечисление ресурсов, которые делаются на материале английского языка. Меня еще просили сказать про некоторые ресурсы, о которых я знаю лучше. А именно я буду говорить о четырех ресурсах: FrameNet, VerbNet, PropBank и NomBank.

Про FrameNet я уже коротко рассказывала: это разметка предикатных слов, глаголов, существительных, прилагательных и так далее в тексте, где мы имеем некоторую цепь предложений и указываем семантические роли, в которых сопоставляются некоторые фрагменты предложения и, даже может быть, предтекстов, посттекстов и так далее. Это подход прежде всего семантический, фреймы — это такая семантическая информация, и синтаксическая зависимость между словами, которые отвечают за те или иные роли. Второе важное свойство FrameNet'а — это то, что установлены связи между фреймами, такие как наследование, каузальная связь, временная, и даже такое понятие, как использование некоторого фрагмента одного фрейма в других фреймах.

PropBank — это проект, который разрабатывается группой Марты Палмер в Америке. Можно сказать, что это такой чуть более простой FrameNet, это тоже разметка ролей. Но на самом деле гораздо больше внимания здесь уделяется синтаксису. Во-первых, размечаются только глаголы. Аргументы на первом этапе были размечены просто номерами (аргумет № 0, аргумент № 1, аргумент № 2), и все эти аргументы должны быть синтаксически зависимы от глагола. И на самом деле, эта разметка была таким углублением. Ну и затем в PropBank была введена информация о типах ролей, вроде бы похоже на FrameNet [демонстрирует слайды]. Помимо синтаксических актантов были размечены некоторые адъюнкты, такие как, например, отрицание, которое стоит при глаголе, или модальные слова.

Ну вот NomBank — это аналогичный проект, но для имен существительных и прилагательных.

А в группе Марты Палмер параллельно разрабатывался проект VerbNet — это такая семантическая классификация глаголов. За ней стоит очень интересная история: Марта Палмер в университете училась вместе с Бет Левин, и они чуть ли не жили в одной комнате. Поэтому человек, который всю жизнь занимается прикладными разработками и, как она сама говорит, не имеет никакого лингвистического образования, она очень хорошо знакома с подходом Бет Левин, и это была такая попытка взять книгу Бет Левин о классах глаголов, эти классы глаголов были сформированы по принципу трансформации глагольного управления и на ее основе создать такой лексический ресурс. Там также размечены семантические роли, также имеется синтаксическая информация, поскольку там примеры размечены, и ограничение на заполнение валентности.

Кроме того, я хочу обратить ваше внимание на то, что был сделан такой ресурс, как SemLink — вначале автоматическими методами, потом с помощью постредактирования был построен индекс глаголов и глагольных значений, связывающий четыре системы — FrameNet, VerbNet, PropBank и NomBank. Это позволило сразу же найти много таких мест, про которые разработчики отдельных систем даже и не подозревали, потому что прямого соответствия между одним ресурсом и другим не было, но это было чрезвычайно полезно.

Вот некоторые уроки, которые имеет смысл извлечь из истории создания английских ресурсов. Первое: не все, но многие ресурсы открыты, и под лицензией их можно скачивать и использовать в тех или иных разработках, и я надеюсь, что развитие российских ресурсов тоже в конце концов пойдет по этому пути. Второе: помимо того, что есть группы, которые разрабатывают ресурсы, в других странах, университетах и лабораториях есть люди, которые проверяют связность информации, насколько в том же FrameNet'е связаны фреймы — с помощью своих разработок, им просто интересно понять, так ли хорош этот ресурс. И сами разработчики создавали некоторые новые классы элементов на основе этой информации. Кроме того, есть обратное движение, когда лингвисты могут как-то помочь компьютерщикам. Может быть, не только лингвисты, но, так сказать, теоретики. К нашей дискуссии о том, нужны ли компьютерщикам лингвисты или не нужны, я бы хотела сказать два слова: у лингвистов может быть еще одна роль — роль экспертов в объяснении некоторых количественных показателей. Мы можем сказать, что на некотором материале данная система дает 95% аккуратности, но эту информацию еще хорошо бы уметь интерпретировать качественно — попробовать объяснить, откуда берутся эти цифры, где какие слабые точки. Это та работа, которая ведется по отношению к английским ресурсам. В прошлом году мы говорили о таких состязаниях, когда на одном материале работают компьютерщики многих программ, и можно их как-то сопоставить и интерпретировать. Единственное отличие WordNet'а, который мы обсуждали, от FrameNet и так далее, то есть разметки ролей, состоит в том, что WordNet, несомненно, более всеобъемлющая система, и количество элементов, которые присутствуют, не сравнимо с тем, что есть во FrameNet и других системах.

Я специально сейчас посмотрела в системе Scholar.Google.Com — это поиск научных работ, я посмотрела, как часто упоминаются WordNet, FrameNet и так далее — везде различие на порядок. Например, просто запрос «WordNet» дает 34 000 цитат, а FrameNet — всего 3 000 с чем-то. Я посмотрела цитаты, когда WordNet используется в связи с разрешением многозначности — то же самое. Или что касается использования этих систем в машинном переводе — 11 000 и 1 000.

Возможно, самый важный урок состоит в том, что делаются попытки использовать несколько видов информации, то есть не только взять WordNet и посмотреть, как он работает, но и благодаря индексу, который связывает разные системы, можно взять WordNet и FrameNet, попробовать использовать их вместе и посмотреть, что это дает. И это есть некоторый ответ на вопрос, что есть три человека, и это такой их подарок миру. Если мы возьмем несколько видов информации, возможно, что степень субъективности станет уже меньше.

В.П. СЕЛЕГЕЙ: Можно, я тогда уточню? В какой мере разработчики FrameNet'а руководствуются системой WordNet'а? Есть какая-то конвенция в этом случае? Понятно, что если вы переходите на глубинный семантический уровень, все

верифицирующий, вы обнаружите определенные несоответствия. Что вы делаете — фиксируете это? Или просто делаете по-своему, не оглядываясь?

О.Н. ЛЯШЕВСКАЯ: В прошлом году я разговаривала с человеком, который своими руками делает статьи для FrameNet'а. Я поняла, что формальных критериев у них нет, но он мне говорил, что они в первую очередь смотрят, что сделано, в частности, он говорил, что разработчик прежде всего идет в WordNet, смотрит, что там есть, какие классы, допустим, глаголов, связанных с его глаголами, можно найти и так далее, то есть это такая отправная точка для его работы.

С другой стороны, я хотела ответить на вопрос о WordNet'е. Недавно я попала на такой workshop, посвященный созданию WordNet'ов для датского и некоторых других языков. И я поняла, что у них есть некоторые стандартные методики оценки того, что они сделали. В частности, поскольку есть top onthologies, за счет привязки к другим ресурсам. Не знаю, насколько хороша и полна эта методика, но какая-то работа в этом направлении ведется.

ВОПРОС ИЗ ЗАЛА: Почему раньше говорили о EuroWordNet, а сейчас уже нет?

О.Н. ЛЯШЕВСКАЯ: EuroWordNet — это конкретный проект, который начался в 1996 году и кончился в 1998 году. И было еще какое-то продолжение в 1999 и 2000 годах. Был еще и третий проект. На такие общие проекты три раза европейская комиссия выдала деньги. В результате образовалось два итальянских WordNet'а, потому что один раз договорились с одной группой, а второй раз — с другой. Даже европейская комиссия может так профинансировать два WordNet'а в каком-то конкретном языке. Три раза по три года. И в рамках этого были созданы WordNet'ы в нескольких языках, в частности голландский, испанский, два итальянских, немецкий. Потом был еще проект BalkaNet, когда несколько южных государств создали свои WordNet'ы. То есть это просто были проекты, которые позволили получить финансирование для WordNet'ов конкретных языков. Поскольку это было общее финансирование, то была какая-то общая концепция, причем отличная от WordNet'а, то есть развитие, попытка введения новых отношений. Эти ресурсы доступны, но, возможно, за плату, если обращаться к их правообладателям. Страны обрели свои WordNet'ы, причем если мы говорили о 100 000 синсетов английского WordNet'а, то, насколько я помню, то максимальный европейский WordNet, голландский, — порядка 70 000 синсетов, еще группа такая есть — немецкий, итальянский, возможно, оба итальянских — по 30 000—40 000 синсетов и так далее. Не знаю, является ли это их политикой, что они дальше не наращивают, или это вопрос финансирования.

ВОПРОС ИЗ ЗАЛА: А у существующих русских какой порядок?

О.Н. ЛЯШЕВСКАЯ: Тысячи, но это механистичный перевод.

М.В. ЯВОРСКАЯ: А у нас, может быть, 13 000.

Б.Л. ИОМДИН (ИРЯ РАН): Мне было дано задание проверить, как обстоят дела с моей лексикой, о которой я делал доклад. Я надеялся, что в WordNet'е как раз все обстоит хорошо. Оказалось следующее. [Демонстрирует слайд.] У меня было три категории — «одежда», «посуда» и всякие «кошельки». Вот что происходит с одеждой: вот три ряда, которые касаются свитеров. Первый ряд, который «sweater», «jumper», толкуются, я выделил красным оформленное слово, genos

proxima, они здесь тоже все разные. Значит, «sweater» и «jumper» описывается очень широко, примерно так, как я предлагал описывать упрощенное гиперонимическое употребление русского свитера — вязаный предмет для верхней части тела. Есть также слово «sweatshirt», которое вроде бы очень похоже на «свитер», но здесь оно в этом ряду отсутствует и толкуется через слово «pullover» и более подробно: во-первых, указано, что длинные рукава, а во-вторых, сказано, что это спортивная одежда. И, наконец, само слово «pullover» толкуется через «sweater», то есть снова получается круг, и входит в один ряд со словом «slipover», который по всем данным все-таки жилет, то есть без рукавов, но здесь они объединены в один ряд. «Джемперов» тут есть три: один — максимально неподробно описанная детская одежда, второе — это рабочая одежда широкая, толкуется через слово «jacket», и третий ряд — «jumper», «pinaford», «pinny» — это в общем такой фартук. Видно, что все три слова здесь разные. И одно общее — «coverall», другие более частные — «jacket» или «dress». Еще есть слово «jersey», которое, по некоторым данным, близко к «sweatshirt», два «jersey». Одно толкуется через слово «pullover», причем сказано, что это облегающий пуловер, при этом синоним — «T-shirt», хотя кажется, что это тоже не совсем правильное толкование для «T-shirt», что это такой пуловер. И второе «jersey» — это сама ткань джерси.

Все это пришлось искать вручную, потому что если слово «sweater» входит в толкование, а не в сам ряд, то не находится. Я здесь нашел то, что мог выявить по другим источникам.

Вот «водолазка», которая толкуется через два слова — «sweater» и «jersey» (в чем разница, не очень понятно). Есть слово «cardigan», которое толкуется через «jacket», «jacket» толкуется через «coat» и «coat» толкуется через «coverall». Причем про «coat» сказано, что это одежда для улицы, соответственно, получается, что и «jacket» — одежда для улицы, дальше мы возвращаемся к кардигану, и на самом деле непонятно, верно ли, что и кардиган, и все свитеры — это тоже одежда для улицы. Видимо, нет, но это ни из чего не следует.

Посуда. Здесь основной ряд — это «glass» и «drinking glass» — толкуется через слово «container». Какие есть «glasses»? «Wineglass», причем где-то он писался раздельно. Четыре ряда, описаны все по-разному: в одном случае указана форма — наличие ножки, в других случаях указывается, какие именно напитки используются. И, наконец, слово «goblet» — оно толкуется не через «glass», а через «drinking glass» почему-то, собственно, толкование очень широкое, это рюмка с основанием и ножкой. И «goblet» входит еще во второй ряд — такая чаша, причем толкуется не через слово «glass», не через «container», а через «vessel»

Третий ряд — это всякие бумажники и кошельки, галантерея. Основное слово для кошелька — это «purse». Оно есть отдельно без всяких синонимов, написано, что это маленькая сумочка для денег, и оно же есть в ряду «bag», «handbag», «pocketbook» и «purse» толкуются через слово «container». Мы видим различия: «purse» — единичный, там не указано, что туда можно положить еще что-то, кроме денег. А другое «purse» — это куда можно положить какие-то личные вещи и аксессуары, особенно женские, и это чуть ли не единственный случай, где я нашел пример, но на слово «bag». Не очень понятно, это все-таки разные значения слова «purse» или одно. И «wallet» с четырьмя синонимами — толкуется не через «container» или «bag», а через «case». Я не нашел несколько слов,

которые очень легко находятся в Интернете, в Википедии и других словарях, актуальных для этой темы, — «money clipper» ‘зажим для денег’, «manbag», что примерно соответствует русской барсетке, и «murse», что примерно соответствует слиянию «manbag» и «purse». Это то, что я сходу нашел в других местах, а в WordNet’е найти не смог.

В.П. СЕЛЕГЕЙ: Вы сейчас сравнивали английский WordNet с русскими толковыми словарями. Но существуют еще и английские толковые словари, сделанные достаточно аккуратно. Я, конечно, представляю себе, как будет выглядеть русский WordNet в исполнении Яблонского — перевод на русский язык таких английских описаний. У меня вопрос, почему разработчики WordNet’а не пользуются какими-то очевидными вещами, которые можно извлечь из толковых словарей? Или и словари врут?

Б.Л. ИОМДИН: В словарях, мне кажется, лучше, чем в WordNet’е.

ГОЛОС ИЗ ЗАЛА: Это говорит о том, что это халтура.

В.П. СЕЛЕГЕЙ: Я не хочу, чтобы круглый стол превращался в «Народ против WordNet’а», как вчера.

Б.Л. ИОМДИН: А я уже против.

ГОЛОС ИЗ ЗАЛА: На самом деле, ничего плохого не было. Просто мы так быстро проскочили. И если вернуться к тому, где был «jacket», то выстраивалась хорошая иерархия.

Н.В. ЛУКАШЕВИЧ: Мне кажется, тут еще проблема, что мы не нашли саму иерархию.

О.Н. ЛЯШЕВСКАЯ: У меня вопрос. Каков статус этих пояснений, которые можно считать толкованиями, в самом WordNet’е, насколько эта информация критична для самого WordNet’а? Или это просто некоторое пояснение, чтобы сами разработчики поняли?

Н.В. ЛУКАШЕВИЧ: Мне кажется, что это не совсем так, потому что если они сейчас эти определения, эти глоссы стали размечать значениями и переводить их в логические выражения, то это означает, что статус у них достаточно высокий. Другой вопрос, что, возможно, какие-то неудачные вещи, которые есть, они остались… Ведь если ресурс очень большой, то разработчик может редко попадать на какие-то плохие места и иметь время быстро что-то исправить. Можно, когда было массовое производство, привлекали студентов, и такая ситуация получилась — что-то недоделано.

В.П. СЕЛЕГЕЙ: Нет, отдельные вещи хорошие. Речь о том, что это не делалось в системе. Хотя казалось, что лексико-семантическое поле требует описания в рамках некоторой концепции и представления о его устройстве.

Б.Л. ИОМДИН: Я нашел кусочек презентации, который не показал. Скажу сейчас в защиту WordNet’а. К моему докладу были возражения, что, может быть, у нас все так плохо, поскольку сейчас меняется какая-то система, легкая промышленность плохо работает, много заимствований, а, наверно, в англоязычной среде все

хорошо разработано, все четко знают, где свитер, где джемпер. Там все так же. Я изучал разные патологии, в частности, в отношении посуды — и там полная мешанина. Вроде бы слова «glass» «goblet» разные, тем не менее вот это «glass», а это «goblet», а есть такая штука — «goblet glass», и это все в одном каталоге. И вот названия некоторых конкретных напитков, которых очень много, они тоже очень разные, и это все термины: «martini glass», «cosmo glass», «sherry glass». И по поводу бумажников и кошельков — есть сайт «Word Refering», где обсуждаются тонкие различия между словами — там, как и на наших форумах, бурные обсуждения, чем отличаются джемперы и пуловеры. И вот для примера, вопрос: чем отличаются «wallet» и «billfold», которые переводятся словарями как ‘бумажник’ — и диаметрально противоположные ответы, и все считают, что они знают. Так что WordNet’у тоже было сложно.

Д.О. ДОБРОВОЛЬСКИЙ: Это все часть синсетов? Как устроен синсет? Попадают все «свитера», «кардиганы» и другие в один синсет?

ШУМ В ЗАЛЕ: В разные.

И.Б. ЛЕВОНТИНА (ИРЯ РАН): А чем они объединены, отношениями?

Н.В. ЛУКАШЕВИЧ: В принципе, есть отношения, но о них не удалось сказать.

Л.Л. ИОМДИН (ИППИ РАН): Услышав эту часть дискуссии о WordNet’е, я не собирался выступать, считая, что уже все сказано. Есть ресурсы разного уровня, одни получше, другие похуже. Мне хотелось сказать пару слов в защиту WordNet’а: худо или бедно, этот словарь работает и используется в частности при составлении, скажем, таких вещей, как UNL. Это не такая вещь, которую вообще невозможно применить. Даже несмотря на то, что некоторые лексические классы не так хорошо проработаны, как другие. Но еще в 1970-е годы Апресян писал: «Как трудно, если вообще возможно истолковать конкретную лексику»: какие-нибудь птицы типа иволги вообще не допускают толкования, если не считать толкования остенсивного. Но это не является недостатком словарей вообще и WordNet’а в частности. Этот словарь хорош и ценен тем, что он задает стандарт, пусть даже стандарт низкого уровня.

Приведу другой пример. Может быть, кто-то видел синтаксически размеченный корпус русского языка, который мы сделали и который располагается на сайте корпуса. Ясно, что он обладает огромным количеством недостатков. Тем не менее, он задает, может быть, не самый лучший, но среднего уровня стандарт. И, как любой стандарт, он имеет право на существование.

А.Л. ВОСКРЕСЕНСКИЙ (Специальная (коррекционная) общеобразовательная школа-интернат № 101): Оговорю, что я все-таки дилетант. Мне кажется, что компьютерная лингвистика должна служить не только на пользу лингвистам, но и компьютерам. И сейчас ведутся работы по созданию SemanticWeb, в котором компьютеры должны обмениваться друг с другом информацией и не задумываться о том, как ее истолковывать. В 2007 году международный стандарт ISO, я не помню номера, common logic, который определяет порядок выдачи информации для онтологий, порядок обмена, common logic interchange формат — там логика первого порядка, и еще концептуальные графы Джона Совы — там используется второй вариант обмена информацией, концептуальными графами. Вот как это направление связано, и связано ли каким-то образом с работами по

компьютерной лингвистике в нашей стране? Как я понимаю, цель стоит какая — добраться до смысла текста и преобразовать его в такую вещь, которая была бы понятна абсолютно всем, не была бы неоднозначной, как UNL. Но UNL — это международный проект, а тут есть международный стандарт. Как работы, связанные с ресурсами, с выдачей из этих ресурсов связаны с этим стандартом, который принят в 2007 году?

Н.В. ЛУКАШЕВИЧ: Я не много знаю про стандарты в области SemanticWeb, но мне известно, что многие работы по разработке онтологий часто заканчиваются тем, что: «А давайте теперь припишем текстовое выражение к этой онтологии за счет того, что мы установим связи с той онтологией, которую мы создали, и WordNet'ом». То есть оказывается, что некоторые такого рода онтологические размышления часто заканчивается тем, что лексическое наполнение берется из WordNet'а.

Л.М. ПИВОВАРОВА (СПбГУ): С онтологиями те же проблемы, что с WordNet'ами, потому что все делают свои, все хотят один общий. Но есть общеевропейский проект онтологий SMO, который старается быть объединителем всех онтологий, и они действительно подгрузили к этой онтологии WordNet, так что WordNet часто бывает полезен. Я хотела еще сделать два замечания с точки зрения человека, который занимается прикладными вещами и должен эти ресурсы использовать. Первое кассается ситуации с тремя русскими WordNet'ами. Можно сколько угодно критиковать перевод с английского на русский, но мне кажется, что в этой конкуренции WordNet'ов выиграет не тот, кто будет более правилен и качествен, а тот, кто первый себя предложит. Я как человек прикладной буду очень ругаться и возмущаться, что они перевели с английского, но жизнь слишком коротка, и ждать, пока М.В. Яворская разберется, чем «жухлый» отличается от «хаки», не буду. Если ресурс будет в доступе, использовать будут его. И второе — тоже с прикладной точки зрения. Если мы говорим о том, что ресурс используется в какой-то системе как одно из звеньев, то качество этой системы определяется качеством самого слабого звена. И самым слабым звеном будет не WordNet, не лексические ресурсы, а какой-то другой. Потому что если мне предложат, назовем это словарь, где будет выделено 15 смыслов для слова «зеленый», то моя система не будет от этого работать лучше. Наоборот, она будет встречать слово «зеленый» в тексте и запинаться, заикаться, ходить по всем его связям. Лингвист, который делает такой ресурс, будет иметь чистую совесть относительно того, что он все срого сделал, но не факт, что он сделает лучше. И мне кажется, что надо перевести фокус внимания с создания ресурсов на их использование, то есть приставить телеге второе колесо и заставить ее ехать на двух колесах. А обсуждать качество ресурса без вариантов его использования, мне кажется, не совсем верно.

Н.В. ЛУКАШЕВИЧ: А видели ли вы второй питерский WordNet? Может быть, там столько шума, что его нельзя будет использовать в прикладных исследованиях?

Л.М. ПИВОВАРОВА: Мы его не видели, но мы о нем слышали, что он очень качественный. Пока они не выложили его, я не могу ничего сказать. И пока его не выложили, они для меня равноправны.

А.С. НАРИНЬЯНИ: Мне кажется, это порочный способ обсуждения. Во-первых, мы знаем, какие сейчас учебники для школы выходят. Так вот, если выходит очень плохой англо-русский словарь. Но уже есть, и все переводят. И что получается?

Надо ли это вообще? Или надо бить этим словарем автора по голове? Или раз он уже есть, то это хорошо?

Л.М. ПИВОВАРОВА: Я не говорю, что это хорошо, я говорю, что он выступит конкурентом в борьбе с очень хорошим, но бумажным.

А.Л. ВОСКРЕСЕНСКИЙ: Я беседовал с Сергеем Яблонским два года назад в Петербурге. И тогда он мне говорил: «Мы почему его не выкладываем?» Потому что у него были какие-то соглашения с теми, кто заказывал, он должен был ехать в Чехию на конференцию, представлять, что-то там должно было решаться. Может быть, ему просто не разрешают выложить в открытый доступ.

В.Г. ДИКОНОВ (ИППИ РАН): Здесь поднимался вопрос о UNL, готов ли UNL стать стандартом. А второй вопрос, правда, он несколько провокационным может оказаться, это общественная значимость лингвистических ресурсов. Провокационность заключается в утверждении, которое я полностью поддерживаю, что ресурс, который не выложен в Интернете, который нельзя в любой момент исследовать, попробовать и применить, все равно что не существует. Труд, который в него вложен, скорее всего, пропал даром и не будет доступен тем, кто будет его далее продолжать и улучшать. WordNet ценен тем, что он изначально очень давно появился в Интернете и к нему не было привязано никаких юридических ограничений, запрещающих его использовать так, как бы захотелось использующему. Вместе с тем это сформировало сообщество WordNet'ов и команд, которые делают WordNet'ы других языков, FrameNet'ы, OntoNet'ы и все возможные варианты, которые существуют. Как мне кажется, это сообщество можно критиковать как минимум за две вещи. Во-первых, это сообщество разобщено. Это провоцирует расхождения между разными национальными WordNet'ами и трудности их сопряжения в будущем. Одновременно английский WordNet, который является организующим центром и средоточием связей всех ресурсов друг с другом, он закрыт от сторонних изменений. Люди сталкиваются с ошибками в WordNet'е, не могут их исправить…

Н.В. ЛУКАШЕВИЧ: На самом деле, если вы считаете, что есть ошибка… Я сама лет 5 назад писала письма в WordNet, по адресу, который там был дан, мне отвечал Джордж Миллер и писал следующее: «Да, мы эту ошибку не видели, мы исправляем». Где-то 15 ошибок я нашла, про 7 они сказали, что нашли сами, с половиной они согласились, и потом я могла в следующих версиях видеть, как те ошибки, которые я им указала, изменили это описание. Не всегда они изменили так, как мне казалось, они должны были изменить, но тем не менее никто вас не останавливает от того, чтобы написать им. По крайней мере, 5 лет назад отвечали и меняли реально.

В.Г. ДИКОНОВ: Хорошо, что такая интерактивность существует. Я бы мог посоветовать заимствовать такой стандартный инструмент, как Bugzilla, это инструмент, ведущий учет ошибок, где можно было бы все подобные обращения проследить. То есть если вы нашли ошибку, вы можете посмотреть, найдена ли она ранее, или нет, будет ли она исправлена, или нет, и если вас не устраивает решение, вы можете оставить свое мнение и обосновать необходимость определенных действий. Это стандартный инструмент любых открытых программ типа Линукса и тому подобного.

Что касается WordNet'ов, как я говорил, происходит возникновение конкурирующих ресурсов, которые обслуживают одну и ту же задачу и команды которых не общаются. Для многих WordNet'ов, кроме Принстонского, наблюдается следующий цикл: получен грант, работа проведена, после чего работа кладется под сукно, и ее результатов больше никто не видит. Вот, например, EuroWordNet, который был сделан, потом закрыт, и организация под нехорошим названием ELDA продает его за деньги. Деньги могут быть и небольшими, но это обозначает, что проект не развивается, потому что денег больше нет, а если они и будут, то очень трудно получить согласие и разрешение продолжить ту работу, чтобы осовременить этот ресурс. В результате его можно считать мертвым.

Например, проект BalkaNet, тоже сходного масштаба, в настоящее время его домашняя страница мертва. В Интернете много публикаций, его критикующих или поддерживающих, но самого BalkaNet'а больше нет. Он лежит у кого-то на CD-Rom'е или даже на дискетах в сейфе.

Польские WordNet'ы существуют, 2 или 3, их в Интернете так же невозможно найти. Большинство ссылок на них с главного сообщества тоже оказываются мертвыми или исчезнувшими страницами. Это крупнейшая претензия к сообществу WordNet.

И.А. БОЛЬШАКОВ (Национальный политехнический институт, Мехико): Об испанском WordNet'е, который я купил 8 лет назад за 600 долларов. Деньги были не мои, но все равно мне их было жалко. Я не смог его использовать, а потом узнал, кто это делал, и еще больше застеснялся, что я так неразумно поступил. И пусть за счет родного мексиканского завода, но все равно было нехорошо сделано. А теперь я знаю, что им и испанцы не пользуются, и уж совсем загрустил. Это был такой тупой перевод. Я его предъявляю: «Ну, вот этот что?» — «Нет, мы этого слова не знаем. Ну, это слово испанское. А мы мексиканцы, мы по этому поводу ничего сказать не можем». Какие же это общие слова? Это же высокие, это топ-понятия. В общем, этот проект был, на мой взгляд, не очень удачным. Я не могу сказать за всю Одессу, за другие языки, но этот опыт был неудачен. Я не смог это применить. Ну вот мой коллега Гельбух что-то пытается применить эту рухлядь к чему-то, сажает на это студентов, но уровень студентов настолько низок, что им и этого хватает.

В.Г. ДИКОНОВ: Вот это как раз и подтверждает мое утверждение, что ресурс, который считается закрытым и законченным и не развивается, мертв. А развивать его, хоть, может быть, и есть желающие, но проблема в том, что им запрещено это делать. И есть конфликт между навязываемыми нашими правительствами и международными торговыми организациями законами о копирайте, которые существуют для того, чтобы защищать монополизм в информационной сфере, и потребностями творческого духа, который хочет исправить, улучшить и развить.

В.П. СЕЛЕГЕЙ: Я хотел бы провести какую-то связь между ресурсами, которые обсуждались. Если посмотреть на UNL как модель, по существу это объединение и FrameNet'а, и WordNet'а, и других ресурсов. Ясно, есть интерлингва, но, в сущности, локальных словари, о которых мы говорили, выполняют функции национальных WordNet'ов. И если все так сложно обстоит с WordNet'ом, может быть, UNL является альтернативой? Никакие государства не препятствуют свободным альтернативам свободных художников, которые занимаются UNL. Как у вас-то дела обстоят?

В.Г. ДИКОНОВ: Что касается UNL, проект достаточно давно существует, в нем тоже дала свои плоды разобщенность, поэтому в частности и я, и Игорь Богуславский очень надеемся на то, что удастся преодолеть эту разобщенность и добиться того, чтобы на словарь можно было опираться. Кроме того, существуют традиционно сложившиеся в разных UNL-центрах различные способы толкования собственно именных структур, но это несколько меньшая по масштабам проблема, потому что объем труда, необходимый для того, чтобы синхронизировать и исправить все словари, больше, чем необходимо для того, чтобы исправить правила, которые преобразуют конкретную структуру нашего языка, в котором мы работаем.

В.П. СЕЛЕГЕЙ: Вопрос в другом. UNL — это международный проект с какими-то стандартами, с каким-то центром? Или это просто брошенная некая модель, которой было сказано: «Плодитесь и размножайтесь»?

В.Г. ДИКОНОВ: Это две крайности, а истина посередине. На самом деле есть проекты, изначально существовал центр, но центр плохо руководил этим. Кроме того, спецификации, которые существуют, не настолько подробно описывают все, чтобы не оставить лазеек для того, чтобы принять собственные решения и образовать уникальный диалект. Кроме того, в них, очевидно, есть вещи, которые следует исправлять. Например, атрибуты модальности в UNL сделаны плохо.

В.П. СЕЛЕГЕЙ: То есть есть шансы заменить WordNet?

В.Г. ДИКОНОВ: Да, конечно. Во-первых, цель проекта — сформировать стандарт. Стандарт пусть долго и трудно, но все равно вырабатывается, если работу не прервать. Соответственно, и доступен он тоже будет.

Е.Н. ЗАРЕЦКАЯ: (АНХ при Правительстве РФ): Что, монополизм ввести, против которого вы выступаете?

В.Г. ДИКОНОВ: Монополизм — одно, а стандарт — другое. Стандартов может быть много.

ВОПРОС ИЗ ЗАЛА: Когда вы сможете сказать, что стандарт выработан, достигнут? Какова конечная точка?

В.Г. ДИКОНОВ: Что касается точки, когда выработан стандарт, то такие точки две. Фактическая, когда большинство групп, работающих в этой сфере, дозреет до такого состояния, чтобы UNL-представление, сделанное инструментами одной группы, стопроцентно верно понималось инструментами другой группы. Прошу не путать с корректностью прогенерированного текста, поскольку каждый лингвистический процессор может иметь свои ошибки. Не будем лишать себя права на ошибку. Но истолкование стандартов языка должно быть для всех едино. Это фактическая точка, а формальная точка — это принятие стандарта, например, стандарта ISO, номер такой-то. Пока эта точка не достигнута, можно говорить только о будущем времени.