М.В.Копотев

НЕОДНОЗНАЧНОСТЬ И ПУТИ ЕЕ РАЗРЕШЕНИЯ В ХЕЛЬСИНКСКОМ АННОТИРОВАННОМ КОРПУСЕ «ХАНКО»

Настоящая статья посвящена обсуждению проблемы языковой неоднозначности, возникающей при создании любого лемматизированного или аннотированного корпуса текстов. В статье обсуждаются типы неоднозначности и способы ее разрешения, принятые авторами аннотированного корпуса русских текстов «ХАНКО», который создается на Отделении славянских и балтийских языков и литератур Хельсинкского университета. В настоящий момент пользователям доступна только морфологически аннотированная часть корпуса, поэтому все положения иллюстрируются примерами морфологической неоднозначности[1].

1. Языковая неоднозначность

Проблема языковой неоднозначности имеет и теоретические, и технические аспекты, что предполагает и собственно лингвистические подходы к ее решению, и обсуждение ограничений, накладываемых машинным представлением языкового материала.

Словосочетание «лингвистическая неоднозначность» не входит в терминологическую систему в рамках русской лингвистической традиции: обычно при автоматической обработке текстов говорят о снятии омонимии. Однако часто имеют в виду не совсем то явление, которое называется омонимией в «большой» лингвистике. В то же время в европейской лингвистической литературе дифференциация между неоднозначностью и омонимии-

---

[1] Подробнее о корпусе см. сайт **www.slav.helsinki.fi/hanco,** а также *Копотвв М., Мустайоки А.* Принципы создания Хельсинкского анноти-рованного корпуса русских текстов (ХАНКО) в сети Интернет // Науч-но-техническая информация. Сер. 2. 2003. №6. Р. 33-37.

ей проводится более последовательно, причем последняя является частным случаем первой[1]. Для подтверждения правомерности такого различения рассмотрим примеры (1-2).

(1) *...рядовой беженский быт поначалу кажется в* ***общем*** *сносным.*

(2) *Завтракали и обедали мы* ***в общем ресторане*** *нашей части гостиницы.*

В предложении (1) «в общем» является идиомой, включаясь в класс наречий. Очевидно, что сочетание «в общем» в предложении (2)- случайное совпадение предлога и прилагательного, входящих в общую предложную группу «в общем ресторане». И если наречие «В ОБЩЕМ» справедливо фиксируется в словарях в качестве составной лексемы («эквивалента слова», по В.В. Виноградову), то было бы странным считать «В.PREP ОБЩЕМ.ADJ» омонимом к «В ОБЩЕМ.ADV» и включать эту пару в словарь омонимов.

Пример (3) демонстрирует неоднозначность при выделении форм предикатива и аналитического будущего.

(3) ***... мы должны будем идти*** *на непозволительные политические уступки.*

Как кажется, признание сочетания текстоформ «будем» и «идти» омонимичным форме будущего времени глагола ИДТИ (как в предложении *«Мы* ***будем идти,*** *пока не дойдем.»)* расходится с общепринятым. Представляется, что традиционный взгляд, согласно которому формы аналитического будущего времени в русском языке не имеют грамматических омонимов, является более последовательным.

См., например: *Lyons J.* Semantics. Vol. 2. London, 1 977. P . 396-409; *Constraint Grammar: A Language-Independent Framework, for Parsing Unrestricted Text l* F. Karlsson, A. Voutilainen, J. Heikkilä, A. Anttila (eds.). Berlin- New York, 1995; *BussmannH.* Lex ikon der Sprachwissenschaft. Stuttgart, 2002. P. 73,283.

В общем случае, принципиальная разница между омонимией и неоднозначностью состоит в том, что первая имеет собственно языковую мотивировку (процессы опрощения, заимствования и т.д.), тогда как неоднозначность возникает из-за случайного совпадения знаков (букв при обработке письменного текста). Неоднозначность не предполагает параллелизма языковых единиц, поэтому при лемматизации совпавшие знаки могут сводиться к разному набору лемм. Так, две текстоформы «в» и «общем» из примеров (1-2) сводятся к одной лемме (В ОБЩЕМ) или к двум леммам (В и ОБЩИЙ) в отличие от омонимии, которая не предполагает наложения текстоформ' (так, текстоформа «стекло» в любом случае сводится только к одной лемме: СТЕКЛО или СТЕЧЬ). Таким образом круг явлений языковой неоднозначности шире, чем область омонимии. И при машинной обработке текстов корректнее говорить все же не о снятии омонимии, а о снятии неоднозначности. Вовсе не настаивая на термине «языковая неоднозначность» (англ. *ambiguity),* хотел бы указать на необходимость определенной терминологической строгости. В противном случае существует опасность ввести в заблуждение лингвистов (и шире - преподавателей русского языка) в силу того, что общелингвистические термины употребляются в смещенном значении.

2. Типы неоднозначности

'2.1. Снимаемая неоднозначность

По причине широко распространенного синкретизма русской системы словоизменения, существует большое количество форм, имеющих более чем одну интерпретацию. Например, форма «книги» может иметь три альтернативных чтения: КНИГА.GEN.SG, КНИГА.NOM.PL, and КНИГА.ACC.PL; текстоформа «сети» с учетом формы второго предложного имеет шесть чтений. В большинстве случаев носитель языка легко разрешает

такого рода неоднозначность и, исходя из контекста, выбирает правильное чтение. Назовем этот тип) неоднозначности *снимаемая*. Так называемые грамматические омонимы составляют большинство случаев снимаемой неоднозначности. С другой стороны, примеры (1-3) иллюстрируют этот же тип неоднозначности, не являясь, однако, примерами грамматической омонимии.

Снимаемая неоднозначность, конечно, представляет серьезные сложности для автоматической обработки текста, и, кажется, в ближайшем будущем надежные результаты по снятию неоднозначности этого типа можно получить, используя программы-дизамбигуаторы с обязательной ручной постобработкой.

2.2. Контекстуальная неоднозначность

Второй тип неоднозначности связан с тем, что не все случаи множественной интерпретации могут быть разрешены однозначным образом. Например, в предложении

(4) *Двери не открывать ни в коем случае*

контекст не позволяет однозначно определить, какое из чтений текстоформы «двери» является верным: ДВЕРЬ.ACC, ДВЕРЬ.GEN или ДВЕРИ.ACC. Все три чтения имеют одинаковое право на существование. Это тип неоднозначности можно назвать *контекстуальной неоднозначностью*.

Пример (4) представляет случай внутрипарадигмальной неоднозначности, однако возможна и межкатегориальная неоднозначность, при которой нельзя определить категориальную, или частеречную принадлежность текстоформы. Пример такого рода представлен в предложении (5):

(5)*...предприниматель будет обращаться в одно ведомство, которое само проведет консультации со всеми заинтересованными инстанциями...*

В этом предложении контекст не позволяет однозначно определить, какое из двух чтений текстоформы «одно» является верным: ОДИН.PRON или ОДИН.NUM.

2.3. Теоретическая неоднозначность

Третий тип грамматической неоднозначности связан с теоретическими установками, на которые ориентируется лингвист. Так, существуют случаи, в которых контекст дает ясное представление о значении, но критерии лингвистической классификации настолько нечеткие, что отнести единицу к определенному типу очень сложно. Примером такого рода может служить группа слов «тысяча», «миллион», «миллиард» и т.д., которые по морфологическому признаку относятся к существительным, но по семантическим - к «счетным словам»[1]. Выбор одного из вариантов - если он и будет сделан - зависит от теоретических предпочтений лингвиста, а не от контекста. Такой тип неоднозначности можно назвать *теоретической* неоднозначностью (или неоднозначностью исходных теоретических положений). Конечно, при компьютерной обработке текстов теоретическая неоднозначность часто связана скорее с практическими задачами, чем со строгой теоретической мотивацией. Тем не менее, назовем этот тип теоретическим, потому что он так или иначе отражает некоторую классификацию материала, принятую лингвистом при создании корпуса.

Все три типа неоднозначности (снимаемая, контекстная и теоретическая), безусловно, должны быть учтены при автоматической обработке текста. И создатели корпусов, конечно, принимают определенные решения. Ниже будет показано, какие подходы к этой проблеме предлагают разработчики «ХАНКО».

[1] *Русская грамматика* / Под ред. Н.Ю. Шведовой. Т. 1. М., 1982.

3. Дизамбигуация в «ХАНКО»

3.1. Обработка снимаемой неоднозначности

Неоднозначность (прежде всего грамматическая омонимия) становится объектом внимания для любого аннотированного или просто лемматизированного корпуса. При создании «ХАНКО» снимаемая неоднозначность разрешалась автоматически с последующей ручной обработкой. Автоматический анализатор был создан в рамках теории «двухуровневой грамматики» *(Two-level Grammar*[1]*)* и «грамматики ограничений» *(Constraint Grammar*[2]*)* как составная часть системы TWOL[3]. Автоматическая дизамбигуация выполнена с помощью отдельного «постморфологического» модуля, при этом использовались правила, дающие точность в 99,9% при устранении грамматических омонимов. Часть таких правил связана с предложным управлением. Например, в предложении *Без друга пропадешь* программа автоматического анализа выдает для формы существительного два падежных варианта: GEN и ACC. Однако падежное управление предлога *без* позволяет убрать омонимию, устранив ACC из списка интерпретаций. Именно это и делает модуль «постморфологии». Фрагмент такого разбора приведен в примере (6):

(6) *Без друга пропадешь*
"<Без>"
"без" PREP
"<друга>"
"друг" N MASC SG GEN
* -"друг" N MASC SG ACC (0)

[1] *Koskenniemi K.* Two-level Morphology: a General Computational Model for Word-form Recognition and Production. Helsinki, 1983.

[2] *Constraint Grammar*... Op. cit.

[3] Сведения об анализаторе см. в статье: *Viikki L.* RUSTWOL: A System for Automatic Recognition of Russian Words // **www.lingsoft.fi/doc/rustwol/rustwol.txt**

3.2. Обработка контекстной неоднозначности

Проблема контекстной неоднозначности разрешается в «ХАНКО» путем сохранения множественной интерпретации (как это частично реализовано в Национальном корпусе русского языка (НКРЯ)). Следует, однако, заметить, что контекстная неоднозначность охватывает несопоставимо более широкий круг случаев, чем омонимия. В силу этого число единиц, получивших множественную интерпретацию, оказалось значительно больше ожидаемого. Так, например, текстоформа «см.» (СМОТРЕТЬ.IMP.SING и СМОТРЕТЬ.IMP.PL) выходит за рамки собственно лингвистических интересов, но представляет определенную сложность для компьютерного аннотирования реального текста. Рассмотрим следующий пример (7):

(7) *Какие семейные проблемы возникают у преуспевших в бизнесе* ***супругов.***

На иллюстрации, приведенной на рис. 1, можно видеть, как выглядит представление неоднозначных текстоформ в «ХАНКО».

3.3. Обработка теоретической неоднозначности

Вообще говоря, теоретическая неоднозначность не является обязательной составляющей при создании корпуса. Так если корпус создается в рамках строго определенных классификационных схем, позволяющих точно определить место единицы, а сам объем материала ограничен, то теоретической неоднозначности можно избежать.

Другой возможный путь - принятие определенных конвенций, которые, однако, должны быть максимально четко эксплицированы, чтобы не вводить в заблуждение возможных пользователей (так, например, это сделано в НКРЯ для компаратива и для других морфологических единиц, имеющих неоднозначную квалификацию).

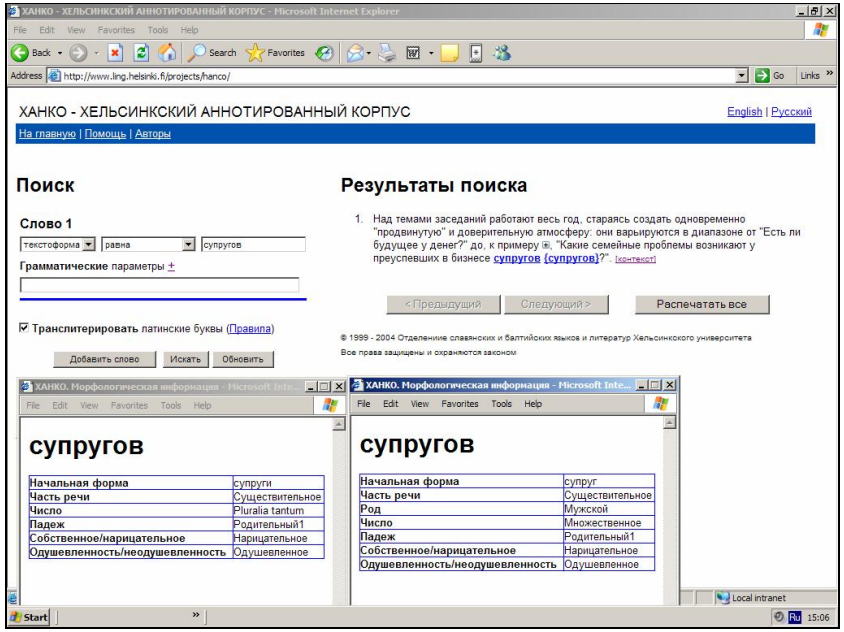

Рис. 1. Представление неоднозначных текстоформ в «Ханко»: пример (7)

В этом смысле перед создателями «ХАНКО» стояла непростая задача в силу того, что корпус, предназначенный для самых широкого круга пользователей (включая преподавателей с неродным русским и без серьезной лингвистической подготовки), должен был опираться на некоторые общепринятые представления о языке Таковым мы посчитали описание русского языка, приятое Русской грамматикой-80, Это описание, если и не считается единственно правильным, то во всяком случае знакомо большинству языковедов. Однако, как известно, оно не лишено внутренних противоречий, которые и приводят к теоретической неоднозначности. В такой ситуации создатели «ХАНКО» приняли решение представлять пользователю все возможные интерпретации.

Так, леммы МИЛЛИОН, ТЫСЯЧА и т.п. описаны и как существительные, и как числительные; леммы ГДЕ, КОГДА и т.п. и как наречия, и как местоимения (см. пример (8) на рис. 2):

(8) ГДЕ.ADV. и ГДЕ.PRON

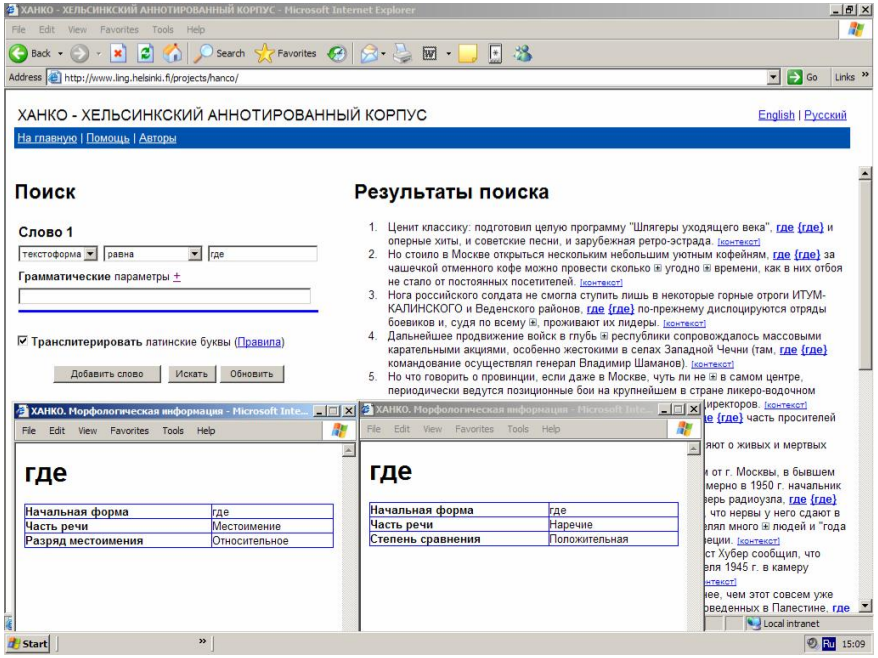

Рис. 2. Представление неоднозначных текстоформ в «Ханко»: пример (8)

Особым типом единиц, представленным в «ХАНКО», стали многокомпонентные служебные фраземы типа «потому что», «в продолжение» и др. Как показывает наш опыт, эти единицы представляют определенную проблему и с теоретической, и с

практической точек зрения[1]. В силу этого значительная часть таких единиц совмещает признаки и теоретической, и контекстной, и снимаемой неоднозначности. В корпусе ХАНКО такие единицы представлены двумя чтениями (пример (9) на рис. 3):

(9) ОТ ДУШИ.ADV и ОТ. PREP и ДУША.SUB.GEN

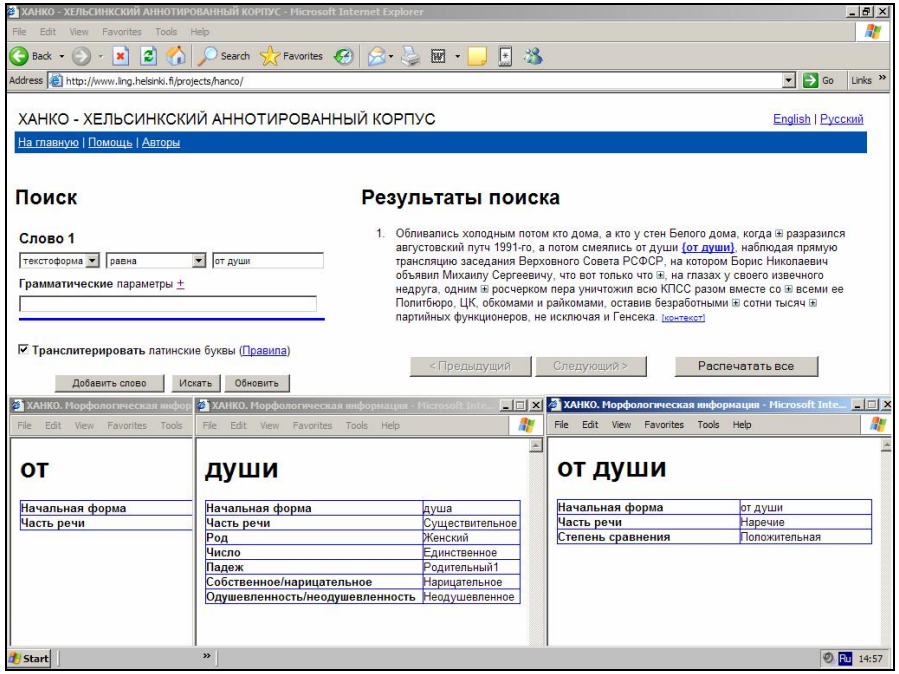

Рис. 3. Представление неоднозначных текстоформ в «Ханко»: Пример (9)

[1] (.) критериях выделения этих единиц см.: *Мустайоки А., Копо-тев М.* К вопросу о статусе эквивалентов слова типа *потому что, в зависимости от, к сожалению II* Вопросы языкознания. 20(М. №3. С.'. 88-107

Выводы

Разработка любого корпуса даже такого сравнительно небольшого, как «ХАНКО», ставит перед исследователями ряд практических и теоретических задач. Одна из них - обработка языковой неоднозначности, которая не совпадает полностью с проблемой разрешения языковой омонимии. Обсуждение проблем неоднозначности при компьютерной обработке и представлении языкового материала выявило определенную специфику этого феномена, несводимую к омонимии. Как показывает опыт разработчиков «ХАНКО», все три типа неоднозначности (снимаемую, контекстную и теоретическую) можно успешно разрешать, полностью снимая возникающую неоднозначность или оставляя в тексте корпуса единицы, имеющие множественную интерпретацию. Очевидно, что и другие объекты компьютерной лингвистики такие, как «текстоформа» (не словоформа), «лемма» (не лексема») обозначают круг явлений, не совпадающих с объектами «большой» лингвистики. Уточнение содержания этих единиц приведет к более четкому определению стоящих перед компьютерными лингвистами задач и к устранению возможных недоразумений со стороны неискушенного в компьютерной лингвистике пользователя корпуса.

ST. PETERSBURG STATE UNIVERSITY
PHILOLOGICAL FACULTY

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
«CORPUS LINGUISTICS - 2004»

October 11 – 14, 2004, St. Petersburg

St. Petersburg University Press
2004

САНКТ-ПЕТЕРБУРКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА - 2004»

11 – 14 октября 2004 г., Санкт-Петербург

Издательство С.-Петербургского университета
2004