Ассоциативные словари и WordNet

Красимира Петрова
Софийский университет им. Св. Климента Охридского,
Болгария

Первые попытки полуавтоматического создания ядра **болгарского WordNet** привели к результату из **9 585 существительных** [Николов, Петрова 2000, 2001]. Целостная база болгарского WordNet создается в рамках проекта **BalkaNet** [http://www.ceid.upatras.gr/Balkanet/].
Первоначальные данные были предоставлены лингвисту для экспертной оценки их точности и адекватности, были сделаны рекомендации по использованию толковых, синонимических, антонимических, словообразовательных словарей для пополнения и модификации полученных синонимических множеств, дефиниций, синтагматических и парадигматических связей [Николов, Петрова 2001].
Поскольку организация понятий в онтологии-тезаурусе WordNet соответствует **психолингвистическим принципам** организации понятий в лексической памяти как категоризированного опыта [Miller et all 1990], нам кажется логичным и естественным "имплементировать" данные ассоциативных словарей русского и болгарского языков, созданных в последнее время [БАС; РАС].
Интегрирование данных **ассоциативных словарей** может проходить через несколько этапов.Во-первых, могут быть раскрыты две новые рубрики в описании системных связей заглавного слова, где целиком поместить словарные статьи ассоциативного поля:

- из прямого словаря, где заглавное слово является стимулом;
- из обратного словаря, где перечисляются все стимулы, вызвавшие даннное слово в качестве реакции.

WordNet, как база данных, комбинирует несколько видов лексикографических источников, и может еще полнее быть использованной для (психо)лингвистических исследований, перевода, множества информационных исследований и в качестве лингводидактического средства и источника сопоставительных исследований, для когнитивного моделирования, построения концептуальных графов и других областей компьютерной лингвистики.

Одна из наиболее бурно развивающихся современных интердисциплин – технологии для обработки естественного языка **(HLT - human language technologies)** - ставит перед нами задачи создания и усовершенствования электронных лингвистических ресурсов славянских языков.
Среди этих лингвистических ресурсов особую популярность и широкое применение находит **WordNet** - электронная база лексических данных, сочетание онтологии и тезауруса, которая разработана и свободно доступна с начала 1990-х [версия 1.6 http://www.cogsci.princeton.edu/~wn/]. Эта база для английского языка расширена и

разработана для 21 европейского языка в **EuroWordNet -** проекте о создании многоязычной лексической базы данных, "связанных" между собой межъязыковым индексом (1996-1999) [http://www.hum.uva.nl/~ewn/]. Дополнение, усовершенствование и поиск возможностей для интегрирования этого ресурса как модуля других компьютерных программ и средств обработки естественного языка продолжается (см., например, обзор деятельности и форумов на сайте Глобальной (Всемирной) ассоциации по WordNet-у (GWA) – http://www.globalwordnet.org; см. также обзор направлений в этой области Пиека Восена – Vossen 2003).

Первые экспериментальные попытки полуавтоматического создания ядра для существительных **болгарского WordNet**, а также и оценка полученных данных лингвистом привели к результату из 9 585 существительных [Nikolov, Petrova 2000, 2001].

В настоящий момент задача создания целостной базы болгарского WordNet, наряду с аналогичными ресурсами для румынского, турецкого, чешского языков решается в рамках проекта BalkaNet [http://www.ceid.upatras.gr/Balkanet/].

Полученные примерно 10 000 синсетов для существительных были предоставлены мне как лингвисту для экспертной оценки их верности, точности и адекватности, были также сделаны рекомендации по использованию толковых, синонимических, антонимических, словообразовательных словарей для пополнения и модификации полученных синонимических множеств, глосс (дефиниций) и других синтагматических и парадигматических данных [Nikolov, Petrova 2001].

Работа рабочей группы по созданию болгарского WordNet-а включает попытки автоматического использования данных англо-болгарско-английского переводного словаря, а также данные синонимического и словообразовательного словарей болгарского языка [см. Totkov, Ivanova, Riskov 2003].

В данном сообщении мы останавливаемся коротко на возможностях применения данных **ассоциативных словарей** в целях расширения и усовершенствования существующих WordNet-ов, в частности, русского и болгарского языков.

Поскольку организация понятий в онтологии-тезаурусе WordNet соответствует **психолингвистическим принципам** организации понятий в лексической памяти как категоризированного опыта [Miller et all 1990], нам кажется логичным и естественным "имплементировать" данные ассоциативных словарей для русского и болгарского языков, созданных в последнее время [например, РАС, болгарского томика "Общеславянского ассоциативного теазуруса", изданного пока самостоятельно БАС 2003].

Интегрирование данных ассоциативных словарей может проходить через несколько этапов.Во-первых, могут быть раскрыты две новые рубрики в описании системных связей заглавного слова, где целиком помещаются словарные статьи ассоциативного поля:

- из прямого словаря, где заглавное слово является стимулом;
- из обратного словаря, где перечисляются все стимулы, вызвавшие даннное слово в качестве реакции.

Далее словарная статья из ассоциативного словаря может быть "расщеплена" на разные виды семантической и концептуальной информации: гипонимы, гипернимы, синонимы, антонимы, и т.д. Эти данные ассоциативных словарей, отражающие языковое сознание анкетированных лиц, могут дополнять или модифицировать установленную онтологию, заложенную в WordNet. Эти экспериментальные данные могут послужить также для некоторых прикладных аспектов семантики, как когнитивное моделирование, концептуальные графы и др. областей компьютерной лингвистики.

Одна из первых попыток использования материалов ассоциативного словаря для проведения автоматизированной классификации частотных глаголов на основании их синтагматических реакций основывается на гипотезу о существовании корреляции между

синтагматическими свойствами глагола и количеством его синтагматических реакций [Синопальникова 2002, 38]
Новизна в создании RussNet - это учет семантико-грамматических и семантико-деривационных отношений, использование методов дефиниционного, контекстного и деривационного анализа, применение теоретических разработок и детального различения абсолютных, фонетических и морфологических дублетов, стилистических, деривационных синонимов, а также деривационных гипонимов, прецизирование словарных дефиниций [Азарова, Митрофанова, Синопальникова 2003].
Факт, что "RussNet производится вручную, позволяет получить качественный тезаурус, учитывающий специфические особенности русского языка", с использованием данных синонимических словарей русского языка с целью ускорения и расширения процесса приобщения русского ресурса к EuroNordNet [см. Гельфенбейн и др. 2003].
Использование ассоциативного словаря для расширения WordNet-а, поскольку небольшая по объему словарная статья – ассоциативное поле – "это не только фрагмент вербальной памяти (знаний) человека, фрагмент семантических и грамматических отношений, но и фрагмент образов сознания, мотивов и оценок русских" [РАС, т.1, 6], только соответствующий специалист должен извлечь информацию из нее. Грамматический строй и словарный состав имеют сетевое представление в РАС [там же, 206]; для болгарского языка пока подобный анализ не проводился, а предстоит сделать на новом собранном материале, который наличен и в электорнном виде [БАС 2003].
Особую ценность представляет собой ассоциативный словарь еще и тем, что эта относительно небольшая по объему информация - ассоциативно-вербальная сеть - аналогична с устной речью говорящего, что оказалось неожиданностью для самих ученых, проводивших частотный анализ данных [см. Караулов 1993, 177]. Кроме этого, степень грамматикализованности подобна характеристикам морфологии текста [цит.соч., 188].
Эти характеристики ассоциативного поля дают возможность прибавить к разнообразной информации отдельного национального WordNet-а еще один источник, в сильно "свернутом", концентрированном виде аналогичный и сопоставимый с данными корпусной лингвистики и устной речи, и в таком смысле более экономный.
Ассоциативно-вербальная сеть содержит и лингвокультурологическую информацию, например, в воспроизведенных прецедентных текстах – пара *человек – собака* (стимул – реакция) развертывается как *Собака – друг человека* [Караулов 1993, 240].
Из наличных данных ассоциативных словарей может быть извлечена семантическая информация, модифицирующая или дополняющая семантические отношения между концептами, отраженными в словах-стимулах и реакциях, а также и грамматическая информация о сочетаемости, словоизменительных и деривационных свойствах слов. На данном этапе все эти рассуждения имеют общий теоретический характер из-за ожидания появления общедоступного болгарского и русского WordNet-а.
Таким образом, WordNet, как база данных, является комбинацией нескольких видов лексикографических источников, и может еще полнее быть использованной для (психо)лингвистических исследований, для перевода, множества информационных исследований, в качестве лингводидактического средства и источника сопоставительных исследований [см. Petrova 2002].

Литература

1. **Азарова, Митрофанова, Синопальникова 2003: Азарова И.В., Митрофанова О.А., Синопальникова А.А.** Компьютерный тезаурус русского языка типа WordNet. http://www.dialog-21.ru/Archive/2003/Azarova.htm

2. **БАС 2003:** *Български асоциативен речник.* Балтова П., Ефтимова А., Липовска А., Петрова К. София: Изд. СУ “Св. Кл. Охридски”. 2003.

3. Гельфенбейн и др. 2003: Гельфенбейн И.Г., А.В. Гончарук, В.П. Лехельт, А.А. Липатов, В.В. Шило. Автоматический перевод семантической сети WORDNET на русский язык. http://www.dialog-21.ru/Archive/2003/Goncharuk.htm

4. **РАС:** Русский ассоциативный словарь. Караулов Ю.Н., Ю.А. Сорокин, Е.Ф.Тарасов, Н.В.Уфимцева, Г.А.Черкасова, Кн.1-4, М., 1994-1998.

5. Синопальникова 2002: Синопальникова А.А. Использование материалов ассоциативного словаря для проведения автоматизированной классификации частотных глаголов на основании их синтагматических реакций. // Материалы к комьютерному тезурусу лексики русского языка / Сост. Азарова И.В., О.А.Митрофанова, СПб.: Изд-во С.-Петерб. Ун-та, 2002. С. 37-41.

6. **Fellbaum 1998: Fellbaum Cristiane** (Ed.), *WordNet: An Electronic Lexical Database.* The MIT Press, Cambridge, London, England, 1998.

7. **Miller et all 1990: Miller G.A., R.Beckwith, C.Fellbaum, D.Gross and K.J.Miller**. *Introduction to WordNet: an on-line lexical database*. In: International Journal of Lexicography 3 (4), 1990, Revised August 1993 - accessible at ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

8. **Nikolov, Petrova 2000: Nikolov T., K.Petrova.** *Building and evaluating a core of Bulgarian WordNet for nouns*. OntoLex'2000. Workshop on Ontologies and Lexical Knowledge Bases. Supported by Bulgarian Academy of Sciences. Sept. 8-10, 2000: Sozopol, Bulgaria. (under print).

9. **Nikolov, Petrova 2001: Nikolov T., K.Petrova.** *Towards Building Bulgarian WordNet.* Euroconference Recent Advances in Natural Language Processing. Proceedings. Ed. G.Angelova, K.Bontcheva, R.Mitkov, N.Nicolov, N.Nikolov, Tzigov Chark, Bulgaria, 5-7 September, 2001, pp.199-203.

10. **Petrova 2002: Petrova K.** *Bulgarian WordNet as a source for (psycho)linguistic studies*. Litora psycholinguistica, Sofia 2002, 339-344.

11. Totkov G., Ivanova P., Riskov Iv. Automated Improving and Forming WordNet Synsets on Conventional (non computer based) Synonym and Bilingual Dictionaries. - accessible at http://www.dialog-21.ru/Archive/2003/Totkov.htm

12. **Vossen Piek.** *WordNet, EuroWordNet, Global WordNet.* In: International conference RANLP: Recent Advances in Natural Language Processing. Tutorials. Borovetz, Bulgaria, 7-9 September, 2003.