*О.В. Мудрая, Б.В. Бабич, С.С. Пьяо, П. Рейсон, Э. Уилсон*

# РАЗРАБОТКА ИНСТРУМЕНТАРИЯ ДЛЯ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ ТЕКСТА[1]

## 1. Ведение.

Лексическо-семантические ресурсы играют важную роль как в корпусной лингвистике, так и в автоматической обработке естественного языка. Семантическая аннотация, и в частности анализ по семантическим полям, все чаще используются (и дают интересные результаты) в автоматическом анализе текста в качестве дополнительной процедуры снятия лексической омонимии и многозначности для разграничения различных значений слова. В течение последних 20-ти лет в университете г. Ланкастер, Великобритания, ведется работа над большим лексико-семантическим ресурсом – базой знаний для многоязыковой системы семантической разметки текста USAS (*UCREL semantic analysis system*) [2]. В отличие от лексиконов WordNet[3], EuroWordNet[4] и HowNet[5], где лексемы объединяются и формируют связи через отношения между значениями слов или их толкованиями, лексикон, разрабатываемый в Ланкастере, использует номенклатуру семантических

[1] Эта работа поддержана грантами фонда UK-EPSRC для проекта ASSIST (EP/C004574/1 для Ланкастерского Университета и EP/C005902 для Лидского Университета).

[2]Система семантической разметки текста USAS и английский лексикон доступны для научных исследований в качестве составной части инструментария *Wmatrix* **http://www.comp.lancs.ac.uk/ucrel/wmatrix/.**

[3] *Fellbaum C. (ed),* WordNet: an electronic lexical database. Cambridge, Mass., 1998.

[4] *Vossen P.,* Introduction to EuroWordNet // N. Ide, D. Greenstein, P. Vossen (eds.) Special issue on EuroWordNet. Computers and the humanities. 1998. № 32(2-3). P. 73-89.

[5] http://www.keenage.com

полей и отображает слова и шаблоны многословных выражений (МСВ) на их потенциальные семантические категории. Семантический теггер USAS снимает лексическую омонимию в соответствии с употреблением слов и МСВ в контексте. Набор широко определенных категорий семантичексих полей, организованных в структуру, подобную тезаурусу, классифицирует лексикон USAS.

Семантический теггер USAS первоначально разрабатывался для анализа расшифровок интервью (на английском языке)[1], после чего он подвергся дальнейшей разработке и усовершенствованию. В частности, для задач проектов Benedict[2] и ASSIST[3], усовершенствования коснулись двух измерений: увеличения лексических ресурсов и включения новых языков, в результате чего Английский семантический теггер (АСТ) был перенесен на финский и русский языки. В последующих разделах статьи мы опишем семантический теггер USAS, и в частности, работающий в режиме онлайн Русский семантический теггер (РСТ) с дружественным пользовательским интерфейсом[4], который является частью многоязыковой системы семантической разметки текста USAS. Особое внимание будет уделено таким характеристикам, как набор семантических тегов, лексические ресурсы, лексическое покрытие и применение.

---

[1] *Wilson A. and Rayson P.,* Automatic content analysis of spoken discourse // C. Souter and E. Atwell (eds.) Corpus based computational linguistics. Amsterdam, 1993. P. 215-226.

[2] *Löfberg L., Piao S., Rayson P., Juntunen J.-P., Nykänen A., and Varantola K.,* A semantic tagger for the Finnish language // Proceedings of the Corpus Linguistics 2005 conference. Birmingham, 2005.
**http://www.corpus.bham.ac.uk/PCLC/cl2005_fst_fullpaper_final.doc**

[3] *Sharoff S., Babych B., Rayson P., Mudraya P. and Piao S.*, ASSIST: Automated Semantic Assistance for Translators // Proceedings of the EACL 2006. Posters & Demonstrations. Trento, Italy. P. 139-142.

[4] **http://148.88.224.86:8080/nlp_tools/rus_sem_tagger**

## 2. Семантический теггер USAS

### *2.1. Семантическая разметка*

В основу семантической разметки текста USAS первоначально была положена система классификации лексики в близком соответствии с *Лонгманским лексиконом современного английского языка* Тома Мак-Артура[1], который включает в себя приблизительно 15 тыс. слов, относящихся к основной лексике английского языка и сгруппированных по 14 семантическим полям / темам, которые, в свою очередь, подразделены на 127 групп и 2441 подгруппу. Со временем, эта структура семантических классов была существенно усовершенствована[2] для лучшего удовлетворения нужд пользователей USAS. В настоящее время она охватывает 21 семантическую категорию, обозначенную заглавными буквами латинского алфавита, и 232 пронумерованные (до трех уровней) подкатегории[3].

Для сравнения отметим *Англо-русский идеографический словарь* Т. И. Шаталовой[4], в котором применен похожий подход к классификации английской лексики, правда, в значительно мень-шем объеме: этот словарь содержит около 3500 слов современно-го английского языка, сгруппированных вокруг 9 тем, которые, в свою очередь, разделены на более мелкие подтемы, варьирую-щиеся от 9 до 30. Интересно, что *Лонгманский*

---

[1] *McArthur T.,* Longman Lexicon of Contemporary English. London, 1981.

[2] *Archer D., Rayson P., Piao S., and McEnery T.,* Comparing the UCREL semantic annotation scheme with lexicographical taxonomies // G. Williams and S. Vessier (eds.) Proceedings of the EURALEX 2004. Lorient, France. P. 817–827.

[3] Полную семантическую разметку USAS можно найти на сайте **http://www.comp.lancs.ac.uk/ucrel/usas/**.

[4] *Шаталова Т. И.,* Англо-русский идеографический словарь. Москва. 1993.

*лексикон совре-менного английского языка* МакАртура включен в список лекси-кографических источников словаря Шаталовой.

Ядром программного инструментария для семантической разметки текста USAS является семантическая база знаний, в которой отдельные слова и МСВ находят свое отображение в семантических категориях. В дополнение к основной системе семантической разметки, в USAS используются особые метки для обозначения ряда отличительных признаков. Например, знак +/- употребляется для обозначения положительных / отрицательных аспектов значений, а такой набор меток как *m*, *f* и *n* обозначает мужской, женский и неопределенный пол соответственно.

Нередко многозначные лексические единицы отображаются во множественных семантических категориях. В этих случаях семантические метки расставляются в порядке употребительности, т.е. метка, соответствующая наиболее употребительному значению, проставляется первой в списке значений [1] . Всем лексичес-ким единицам также приписывается грамматическая категория части речи с целью уменьшения неоднозначности. Многие лексе-мы в лексиконе USAS одновременно принадлежат к двум и более семантическим категориям, образуя гибридную категорию[2], что обозначается с помощью косой черты:

| | | |
|---|---|---|
| *rebel* | *VV0* | *G1.2/A6.1- S8- A6.1-* |
| *waiter* | *NN1* | *I3.1/F1/S2.2m* |
| *адмирал* | *S* | *G3/S7.1+/S2mf L2mf* |
| *больничный* | *A* | *B3/H1 Q1.2/B2-* |

[1] На основе *Collins COBUILD on CD-ROM 2001 Lingea Lexicon, ver. 3.1*, а также *Encarta World English Dictionary 1999 Microsoft Corporation* для английского языка. Для русского языка использовались *ABBYY Lingvo 10 English-Russian Electronic Dictionary 2004* и *ГРАМОТА.РУ* **http://www.gramota.ru/**.

[2] *Leech G., Garside R., and Bryant M.,* CLAWS4: The tagging of the British National Corpus // Proceedings of the COLING 1994. Kyoto, Japan. P. 622-628.

### *2.2. Лексическо-семантические ресурсы*

Как было отмечено выше, система семантической разметки текста USAS была первоначально задействована в ACT, а затем успешно перенесена на финский и русский языки путем использования структуры английского инструментария с необходимыми корректировками. В ходе работы над проектом ASSIST мы занимаемся разработкой параллельного инструментария для русского языка – РСТ. В основу РСТ положены семантические категории USAS, которые совместимы с семантической категоризацией объектов и явлений в русском языке, например:

*poor JJ I1.1- A5.1- N5- E4.1- X9.1-*
*бедный A I1.1- A6.3- N5- O4.2- E4.1-* [1]

Однако, в отличие от аналитического английского языка, русский является синтетическим флективным языком с богатой морфологией: как правило, то что выражается синтаксическими структурами в английском языке, в русском находит свое выражение посредством аффиксации – окончаний и формообразующих суффиксов и приставок. Для анализа сложной морфосинтаксической структуры русских слов была выбрана программа морфологического анализа И.В. Сегаловича *mystem*[2], которая служит эквивалентом автоматического анализатора частей речи (POS tagger) *CLAWS*[3] в структуре USAS. Чтобы сделать вывод *mystem* в кодировке Cp1251, которая обычно используется для ки-

[1] I1.1- = Деньги: недостаток; A5.1- = Оценка: плохо; N5- = Количество: мало; E4.1- = Несчастный; X9.1- = Способность, интеллект: плохие; A6.3- = Сравнение: мало разнообразия; O4.2- = Суждение о внешности: плохо

[2] *Segalovich I.,* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the MLMTA 2003. USA. P. 273-280.

[3] *Garside R. and Smith N.,* A hybrid grammatical tagger: CLAWS4 // R. Garside, G. Leech, and A. McEnery, (eds.) Corpus annotation: linguistic information from computer text corpora. London, 1997. P. 102-121.

риллицы, совместимым с компонентами USAS, было необходимо преобразовать его в UTF8. Несмотря на эти модификации, архитектура программного обеспечения РСТ в целом воспроизводит архитектуру компонентов АСТ.

Подобно АСТ, лексические ресурсы РСТ включают в себя словник однословных лексических единиц и словник МСВ. Однако, вследствие высокоразвитого флективного словоизменения русских слов, только леммы включены в словник однословных лексических единиц, в отличие от словоформ английского лексикона. В некоторых случаях, когда несколько слов имеют одинаковую лемму, это может привести к их неправильному отображению в семантических категориях, что требует дальнейшей работы по снятию лексической и морфологической омонимии[1].

Еще одной модификацией в русском словнике однословных лексических единиц является выделение имен собственных (например, личных имен и географических названий) в особый подлексикон из-за того, что *mystem* не различает имена собственные и нарицательные. В том случае, когда имя собственное и имя нарицательное имеют одну и ту же форму, имени собственному дается преимущество. Так, полный рабочий процесс РСТ может быть представлен следующим образом: необработанный русский текст → морфологизатор *mystem* → русский семантический компонент (однословные лексические единицы / имена собственные + МСВ) → семантическая аннотация.

Лексические ресурсы для РСТ создаются путем эксплуатации словарей и корпусов. В первую очередь, используются легкодоступные материалы, например, списки имен собственных, которые затем классифицируются в соответствии с системой семан-

[1] Снятие омонимии еще не было осуществлено в РСТ. Наш подход к снятию семантической омонимии в АСТ описан в *Rayson P., Archer D., Piao S. L., McEnery T.,* The UCREL semantic analysis system // Proceedings of the workshop on Beyond named entity recognition semantic labelling for NLP tasks in association with the LREC 2004. Lisbon, Portugal. P. 7-12.

тической разметки USAS. Разработка словника однословных лексических единиц началась с включения в него 3000 наиболее частотных лемм из Национального корпуса русского языка[1]. На текущем этапе пополнение словника осуществляется тематическими списками с помощью онлайновых ресурсов[2]. В дальнейшем словник будет расширяться также посредством загрузки в РСТ текстов из различных источников и последующей семантической классификации ненайденных в нем слов из этих текстов.

На данный момент русский словник включает в себя 16 103 леммы, из которых 11 671 – имена нарицательные и 4432 – имена собственные, а также 713 МСВ. Ко времени завершения проекта ASSIST в конце марта 2007 года состав словника должен достигнуть 30 тыс. лемм и около 9 тыс. МСВ (ср. с 54 953 словоформами и 18 921 МСВ в лексиконе АСТ). Заметим, что многие МСВ являются шаблонами (с возможными словами-вставками), способными распознавать варианты лексем в составе МСВ:

*follow*_* {Np/P*/R*} through_RP* *A1.1.1 M1/K5 X2.4*
*без_* видим*_* {на/то} причин*_** *X2.5- A2.2-*

### *2.3. Оценка лексического покрытия РСТ*

В проекте ASSIST была произведена оценка лексического покрытия РСТ на специально созданном для проекта *Русском новостном корпусе* в 70 млн. слов, который включает выпуски трех основных российский газет – *Труд*, *Известия* и *Страна.Ru*, опубликованных в 2002-2004 гг., с результатом 79% (ср. с покрытием АСТ в 96%). Для оценки лексического покрытия Русский новостной корпус был пролемматизирован с помощью морфологизатора *mystem.* Приблизительное снятие лексической омонимии проводилось через выбор наиболее частотной леммы для данной словоформы, представленной в размеченной вручную части Национального корпуса русского языка в 1,6 млн. слов. Покрытие РСТ

[1] **http://ruscorpora.ru/**, а также **http://corpus.leeds.ac.uk/list.html**
[2] К примеру, **http://www.terms.ru/**.

оценивалось на лемматизированном корпусе, включающем пунктуацию. Частотные слова, не представленные в словнике РСТ, принадлежат к области современных политических и общественных событий; в будущем словник РСТ будет расширен за счет таких слов. Наша цель – достичь 90% лексического покрытия корпуса.

### 3. Применения РСТ

Самым очевидным применением РСТ является компьютерный семантический анализ русского текста. Другое, связанное с ним применение – это компьютерный контент-анализ, касающийся статистического анализа семантических признаков текстов посредством группировки слов и словосочетаний по категориям семантических полей и определения частотности слов и семантических тегов в текстах. РСТ также используется для разработки автоматизированных средств для переводчиков. Так, в проекте ASSIST РСТ выполняет семантическую аннотацию русского текста с целью нахождения в сравнимых корпусах непрямых переводных эквивалентов фраз, составляющих трудность при переводе[1]. Поиск переводных эквивалентов осуществляется через установление соответствий между похожими описаниями ситуаций, описанных в терминах семантических тегов. Система ASSIST находится еще в стадии разработки, состоящей в расширении АСТ, дальнейшей работе над РСТ с целью достижения покрытия 90% на корпусе русских текстов, усовершенствовании процедуры извлечения семантически похожих ситуаций и завершении работы над пользовательским интерфейсом системы ASSIST[2].

---

[1] *Sharoff S., Babych B. and Hartley A.,* Using comparable corpora to solve problems difficult for human translators // Proceedings of the COLING/ ACL 2006 Main Conference. Poster Sessions. Sydney. P. 739-746.
**http://www.aclweb.org/anthology/P/P06/P06-2095**

[2] **http://corpus1.leeds.ac.uk/assist/v05/**