A. V. Kashkina (Voronezh) Comparative Markeme Analysis: Challenges and Perspectives

The paper discusses the application of markeme analysis in comparing literary texts. The method was developed by Alexey Kretov who also coined two key notions: *markeme* and *Index of Thematic Markedness*. Markeme analysis is a quantitative approach which allows analyzing word frequency in the texts.

In the first section of the paper the author discusses criteria for markeme allocation in the text. In her previous works she applied markeme analysis to Russian poetry. The present research questions whether it is possible to use the same method for comparing text from two literary traditions – Russian and English.

The attempt to compare poetry with quantitative methods reveals several challenges for the researcher. The second section of the paper discusses problems which arise due to typological differences of Russian and English. The first problem crucial for automatic markeme allocation is correlation between pronunciation and orthographic forms of words in Russian and English. Grammatical and lexical homonymy poses the next problem. Yet another difference concerns the category of case: while in Russian the nominative case is one of the most reliable criteria for allocating markemes, it is totally irrelevant for English texts. Finally comes the question of semantic correlation between English and Russian lexemes.

Comparative application of the markeme analysis suggests that more reliable criteria are needed for markeme allocation in different languages.

А. В. Кашкина (Воронеж)

Сравнительный маркемный анализ: проблемы и перспективы

Ключевые слова: *маркема, индекс тематической маркированности, количественный анализ, язык поэзии, сопоставительное исследование*

Key words: *markeme, Index of Thematic Markedness, quantitative analysis, poetic language, comparative study*

Настоящая статья посвящена трудностям сопоставления данных маркемного анализа литературы на разных языках.

Маркемный анализ – метод исследования текстов, опирающийся на понятия *маркемы* и *индекса тематической маркированности*, сформулированные А. А. Кретовым (Кретов 2007; 2008; Кретов, Воронина, Попова, Дудкина 2009). **Индекс тематической маркированности** (ИнТеМа) – это квантитативная характеристика встречаемости словоформы в том или ином тексте, вычисляемая по следующей формуле:

ИнТеМа = Ч-вес – Д-вес,

где ***Ч-вес*** – относительная частота словоформ, а ***Д-вес*** – их функциональный вес (зависящий от длины словоформы). Эти величины определяются по формуле, предложенной В. Т. Титовым (Титов 2002; 2004):

$$P_{r_i} = \frac{\sum r - R_{i}}{\sum r},$$

где $\sum r$ – сумма единиц всех рангов (то есть, общее количество словоформ в данном тексте), $R_{1\text{-}i}$ – сумма единиц от первого до данного ранга. Ранги присваиваются словоформам в порядке убывания их частоты/длины. Длина лексем определяется в звуках, а не в буквах. При несоблюдении этого условия значение функционального веса окажется завышенным для одних слов и заниженным для других.

Для автоматизации подсчета ИнТеМа была разработана специальная программа тематического анализа лексики «ТемАЛ»;

Язык, коммуникация и социальная среда. Выпуск 8. Воронеж: ВГУ, 2010. С. 228-235.

идея – д. ф. н, проф., зав. кафедрой теоретической и прикладной лингвистики ВГУ А. А. Кретов, руководство – к. т. н., доц. каф. программного обеспечения и администрирования информационных систем ВГУ И. Е. Воронина, исполнение – студентка 4 курса ф-та прикладной математики и механики ВГУ Ирина Попова (Кретов, Воронина, Попова 2010). Эта программа на основе введённого в неё файла в формате *.txt* вычисляет относительную частоту, функциональный вес и индекс тематической маркированности словоформ в русскоязычных текстах.

Маркемы – первые 50 словоформ с наибольшим положительным индексом тематической маркированности, удовлетворяющие ряду критериев, предложенных А. А. Кретовым (Кретов, Катов 2009; Кретов, Катов, Фаустов 2009):

- *частеречный* (к маркемам относят только существительные);
- *грамматический* (существительные должны быть в форме именительного падежа единственного числа);
- *грамматико-семантический* (исключаются онимы, одушевлённые существительные, кроме слова *человек*);
- *тематико-семантический* (исключаются названия месяцев, дней недели, литературных жанров и т. д., названия артефактов, не являющихся символами, а также лексемы, специфичные для данного конкретного автора/жанра/направления);
- *стилистический* (отсеиваются стилистически окрашенные слова);
- *диалогический* (исключаются обращения);
- *классификационный* (отсеиваются слова-классификаторы, участвующие в конструкциях «СУЩЕСТВИТЕЛЬНОЕ В ИМЕНИТЕЛЬНОМ ПАДЕЖЕ + СУЩЕСТВИТЕЛЬНОЕ В РОДИТЕЛЬНОМ ПАДЕЖЕ»).

Приведённые выше понятия маркемы и ИнТеМа были использованы для анализа русских поэтических текстов 135 авторов с начала XVIII по начало XXI века. Одна из возможных перспектив этого исследования – сопоставление результатов, полученных для

русской поэзии, с маркемными характеристиками корпусов поэтических текстов на других языках, например, английском.

Одна из попыток сопоставительного анализа русской и английской поэзии с использованием квантитативных методов была отражена в нашем докладе «Динамика словаря английской и русской поэзии XVIII – XX веков» на пленарном заседании студенческой научной сессии (апрель 2009). Это исследование проводилось с опорой на работу Жозефины Майлз (Miles 1965), при помощи количественных методик изучившей лексический состав английских поэтических произведений XVI – XX веков и выявившей особенности эволюции языка английской поэзии. Заметим, что в нашем выступлении при сопоставлении англо- и русскоязычных текстов учитывались исключительно относительные частоты слов. Сравнительный анализ английской и русской поэзии с использованием понятия маркемы на настоящий момент не проводился.

Несомненно, что подобное исследование позволит сделать выводы о сходствах и различиях в характере и темпах эволюции поэтической лексики для английского и русского языков. Однако выделение маркем в английских поэтических текстах и их последующее сопоставление с русскими может столкнуться с рядом следующих проблем.

1. Автоматизация подсчета количества звуков в слове

Естественно, что взаимное соответствие букв и звуков для английского языка отлично от их соотношения в русском. Следовательно, программа «ТемАЛ», которая для вычисления ИнТеМа определяет, в частности, длину словоформ в звуках в русскоязычных текстах, для анализа необработанного корпуса текстов на английском языке использована быть не может. Применение данной программы становится возможным только в случае, если количество букв приведено в соответствие числу звуков в словах. Таким образом, предназначенные для анализа текстовые файлы необходимо предварительно отредактировать средствами Microsoft Word. Так, например, буквосочетание *sh* заменяется на *ш*, *wh* – на *в* и т. п. (разумеется, это исключительно функциональные обозначения, не передающие реальные звуки). Две одинаковые буквы, обозначающие согласный, преобразуются в одну заглавную (для удобства обратного преобразования). Сочетания *or*, *ar* и

подобные могут быть заменены, например, на русские буквы *о*, *а* и т. д. (компьютерная программа воспринимает их как отличные от английских символов *o*, *a* и т. п., что позволяет избежать путаницы). Долгие гласные допускают переобозначение через соответствующие этим звукам заглавные буквы, (например, *sheep* → *shIp*, *pool* → *pUl* и т.п.). Возможно, потребуются и другие преобразования текста. Таким образом, для обработки значительного по объему материала данный метод представляется слишком громоздким. Анализ крупного корпуса произведений английских писателей станет практически осуществимым лишь при условии создания программы, аналогичной «ТемАЛ», для работы с англоязычными текстами.

2. Грамматическая и лексическая омонимия

Многие слова в английском языке (*look, love, hate* и т. д.) могут выступать в качестве разных частей речи (например, *love*-существительное и глагол *to love*). Существующие в настоящее время компьютерные программы, и, в частности, «ТемАЛ», оказываются неспособны в данном случае определить частеречную принадлежность лексем такого рода. Конечно, эта неопределённость легко разрешается при обращении к контексту, но в случае анализа обширного текстового материала исследование всех контекстов оказывается слишком трудоёмким. Возможный способ разрешения этой проблемы аналогичен описанному в статье «Использование статистических данных употребительности падежей для уточнения результатов маркемного анализа» (Кашкина 2010). В данной работе рассматривается идея об устранении погрешности подсчета ИнТеМа, связанной с невозможностью автоматического определения падежа омонимичных форм русских существительных, при помощи статистического метода. Этот подход может быть применён и к некоторым случаям грамматической омонимии в английском. Так, наиболее типичный случай – совпадение существительного и глагола. Для реализации описанного в (Кашкина 2010) метода необходимо выбрать одну или несколько наиболее характерных лексем данного типа (например, *love*) и на основании анализа произведений небольшого числа писателей (3-4) некоторого хронологического среза выявить процент случаев употребления этой лексемы в качестве существительного от общего количества употреблений. Затем вычисляется

среднее арифметическое данных, полученных по исследованным авторам, которое и принимается за общий коэффициент для рассматриваемого периода. Чтобы определить действительную относительную частоту (а, следовательно, и ИнТеМа) употребления в роли существительного любой лексемы такого типа, встречающейся в исследуемом периоде, необходимо умножить относительную частоту этой лексемы, выданную программой ТемАЛ, на полученный коэффициент. Данный метод уже был апробирован в (Кашкина 2010) для снижения погрешности в вычислении ИнТеМа омонимичных грамматических форм в русских поэтических текстах. Однако существуют и более сложные ситуации – наложение грамматической и лексической омонимии, например, *lie*: «lie 1 v.i. & n. *(make a) statement that one knows to be untrue…* lie 2 v.i. 1. *be, put oneself, flat on a horizontal surface or in a resting position; be at rest* …. n. *the way sth lies*, (fig.) *the state of affairs*» (Hornby 1962). Правомерно ли применение статистического подхода к случаям такого рода, покажут дальнейшие исследования. Но в целом на данном этапе омонимия представляет собой серьезное препятствие для маркемного анализа.

3. Нерелевантность падежного критерия

Один из фильтров выделения маркем для русскоязычных текстов – грамматический (лексемы должны быть в единственном числе и именительном падеже). К сожалению, для английского языка особенности грамматики – отсутствие формальных признаков падежа – делают применение этого критерия невозможным. В этой ситуации, на наш взгляд, следует опираться на роль слова в предложении. Таким образом, мы можем видоизменить грамматический фильтр для англоязычных текстов: маркемами являются существительные в единственном числе, выступающие в предложении в качестве подлежащего. Но подобный подход делает выделение маркем невозможным без анализа контекста, что затрудняет исследование обширного текстового материала.

4. Отсутствие взаимно-однозначного соответствия между значением английских и русских лексем

Сопоставительный маркемный анализ русских и английских текстов предполагает сравнение ИнТеМа лексем. Для получения

адекватных результатов необходимо, чтобы сопоставляемые лексемы имели одинаковую семантику. Однако на практике значение слова либо в русском, либо в английском часто оказывается более широким. Так, например, лексеме *земля* соответствуют *earth, land, soil*, лексеме *dream – сон* и *мечта* (АРС 1971, ОРАС 2004). Теоретически возможно вести автономный подсчёт ИнТеМа для каждого из значений или даже оттенков значения слова. Но такой метод требует непрерывного обращения к контексту. В этом случае исследование значительных по объему корпусов русско- и англоязычных текстов практически неосуществимо из-за чрезмерной трудоёмкости.

Ввиду всех вышеперечисленных причин сравнительный маркемный анализ английских и русских поэтических текстов в настоящий момент представляется малоосуществимым. Возможно, на данном этапе нашего исследования имеет смысл провести сопоставление маркемных характеристик русской поэзии с данными для близкородственных языков, то есть, украинского и белорусского. Но, безусловно, при дальнейшем развитии и совершенствовании методик откроются широкие возможности для сравнительного маркемного анализа литературы на разных языках мира.

Литература

1. Кашкина, А. В. Использование статистических данных употребительности падежей для уточнения результатов маркемного анализа / А. В. Кашкина //Грамматика III тысячелетия в контексте современного научного знания: XXVIII Распоповские чтения: материалы Международной конференции, посвящённой 50-летию со дня основания кафедры русского языка филологического факультета ВГУ, 85-летию со дня рождения проф. И. П. Распопова, 75-летию со дня рождения проф. А. М. Ломова (Воронеж, 12-14 марта 2010 г.): в 2 ч. – Воронеж: ВГПУ, 2010. – Ч. I. – С. 84-90.

2. Кретов, А. А. Архаисты и новаторы в русской литературе XVIII – начала XX вв. / А. А. Кретов // Универсалии русской литературы: сб-к статей. – Воронеж: Воронежский государственный университет; Издательский дом Алейниковых, 2009, С.29-48.

3. Кретов, А. А. Метод формального выделения тематически нейтральной лексики (на примере старославянских текстов) /

А. А. Кретов // Вестник ВГУ. Серия Системный анализ и информационные технологии. – 2007. – № 1. – С. 81-90.

4. Кретов, А. А. Опыт выявления архетипов поэзии А. В. Кольцова / А. А. Кретов // Лінгвистичні студії. Зб. наук. праць. В. 16. / Укл.: Анатолій Загнітко (наук. ред.) та ін. – Донецьк: ДонНУ, 2008. – С. 353-366.

5. Кретов, А. А. «Программа выделения тематически маркированной лексики» / А. А. Кретов, И. Е. Воронина, И. В. Попова, зарегистрировано в Государственном информационном фонде неопубликованных документов ФГНУ «Центр информационных технологий и систем органов исполнительной власти» (№ 50201000004 от 11.01.2010).

6. Кретов, А. А. Функциональный подход к выделению ключевых слов: методика и реализация / А. А. Кретов, И. Е. Воронина, И. В. Попова, Л. В. Дудкина // Вестник ВГУ. Серия Системный анализ и информационные технологии. – 2009. – № 1. – С. 68-72.

7. Кретов, А. А. Поэзия А. В. Кольцова в контексте русской литературы XVIII – начала XX вв. / А. А. Кретов, М. В. Катов // А. В. Кольцов вчера, сегодня, завтра: Материалы Межвузовской научной конференции / Воронежский государственный университет. – Воронеж: Издатель О. Ю. Алейников, 2009. – С. 137-150.

8. Кретов, А. А. Сквозь призму маркем: Н.В. Гоголь в ближайшем контексте русской литературы / А. А. Кретов, М. В. Катов // Вестник ВГУ. Серия лингвистика и межкультурная коммуникация. – 2009, – № 2. – С. 12-21.

9. Кретов, А. А. Опыт лингвистической генеалогии (на примере Н. В. Гоголя) / А. А. Кретов, М. В. Катов, А. А. Фаустов // Проблемы изучения живого русского слова на рубеже тысячелетий: Материалы V Всероссийской научно-практической конференции / [А.Д. Черенкова (науч. ред.); ред. кол.]. – Воронеж: ВГПУ, 2009. – С. 76-83.

10. Титов, В. Т. Общая квантитативная лексикология романских языков / В. Т. Титов. – Воронеж: Изд-во Воронеж. ун-та, 2002. – 238 с.

11. Титов, В. Т. Частная квантитативная лексикология романских языков: Монография / В. Т. Титов. – Воронеж: Изд-во Воронеж. гос. ун-та, 2004. – 552 с.

12. Фаустов, А. А. Литературные универсалии: на пути к терминологической демаркации / А. А. Фаустов // Универсалии русской

литературы: сб-к статей. – Воронеж: Воронежский государственный университет; Издательский дом Алейниковых, 2009. – С. 8-28.

13. Фаустов, А. А. От ключевых слов к литературным универсалиям: несколько методологических соображений / А. А. Фаустов // Вестник ВГУ. Серия лингвистика и межкультурная коммуникация. – 2009. – № 2. – С. 7-11.

14. Miles, J. The continuity of poetic language; the primary language of poetry, 1540's-1940's / J. Miles. – New York: Octagon Books, 1965. – 542 c.

Источники

1. АРС: Англо-русский словарь / Сост. В. К. Мюллер. – М.: Советская Энциклопедия, 1971. – 912 с.

2. ОРАС: Оксфордский русско-английский словарь/ Сост. Маркус Уилер. – М.: Локид-Пресс, 2004. – 920 с.

3. Hornby, A. S. A learner's dictionary of current English / A. S. Hornby. – Oxford University Press, 1962. – 1200 p.

Получено 05.10.2010