УДК 81’32

М. С. Матыцина, д. филол. н.
Липецкий государственный технический университет, Липецк, Российская Федерация
M. S. Matytcina, Doctor of Sc. (Philology)
Lipetsk State Technical University, Lipetsk, Russian Federation

К ВОПРОСУ ИССЛЕДОВАНИЯ СОЧЕТАЕМОСТИ ЛЕКСИЧЕСКИХ ЕДИНИЦ С ПОМОЩЬЮ СТАТИСТИЧЕСКИХ МЕТОДОВ

TO THE ISSUE OF RESEARCHING THE COMBINABILITY OF LEXICAL UNITS BY MEANS OF STATISTICAL METHODS

Статья посвящена рассмотрению вопросов, связанных с определением сочетаемости слов в речи на примере английского корпуса English Web Corpus (enTenTen) и его подкорпусов. Автор рассматривает словосочетание глагола take с соседним словом, то есть биграммы двухсловного сочетания. Особое внимание уделяется сравнению коллокаций в подкорпусах, содержащих тексты разных жанров и тематик. Для решения поставленной задачи было проанализировано более 100 биграмм, полученных посредством мер ассоциации t-score, MI-score и Log Dice.

Ключевые слова: лингвистический корпус; подкорпус; коллокации; меры ассоциации; t-score, MI-score, Log Dice.

The article is devoted to the issues related to the definition of word combinability in speech. The research is based on the English Web Corpus (enTenTen) and its sub-corpuses. The author considers the word combination of the verb ‘take’ with a neighbouring word, i.e. bigrams of a two-word combination. For this purpose, four sub-corpora of the English Web Corpus (enTenTen), which constitute the largest percentage of the total corpus, were selected. For the task, more than 100 bigrams were analysed by means of the association measures t-score, MI-score and Log Dice.

Key words: linguistic corpus; sub-building; collocations; measures of association; t-score, MI-score, Log Dice.

Использование статистических методов представляет собой попытку выявить коллокации на основе больших корпусов и статистических показателей, обычно называемых статистическими мерами или мерами ассоциации (далее – МА). В. П. Захаров и М. В. Хохлова в своей работе «Анализ эффективности статистических методов выявления

коллокаций в текстах на русском языке» отмечают, что меры ассоциации «учитывают как частоту совместной встречаемости, так и другие параметры, прежде всего частоту в данном корпусе каждого отдельного элемента» [Захаров, Хохлова 2010]. Правильное и эффективное использование МА при извлечении коллокаций показывает не только частоту совместного употребления слов, но и направление изменения их лексических значений при устойчивости словосочетания. Это позволяет исследователям анализировать полученные результаты, как с качественной, так и с количественной точки зрения, дать оценку силе притяжения слов [Hunston 2002; McEnery, Hardie 2011].

Предметом рассмотрения являются результаты автоматического выделения глагольных словосочетаний, в которых поисковым запросом является семантически главный компонент – глагол *take*. Исследование глагольных словосочетаний проводилось с помощью мер ассоциации *t-score*, *MI-score* и *Log Dice*, позволяющих оценить силу связности компонентов словосочетания, с целью определения их эффективности в решении поставленной задачи. Материалом для проведения настоящего исследования послужили данные корпуса *English Web Corpus* (*enTenTen*), корпуса английского языка, состоящего из интернет-текстов различного объема и содержания.

Несмотря на большое количество корпусных исследований сочетаемости, по-прежнему существует потребность в более глубоком понимании факторов, играющих важную роль в установлении того, что можно считать коллокациями. Бесспорно, невозможно выделить все факторы, играющие важную роль в идентификации словосочетаний и дать подробное описание всех МА. Однако их критическое рассмотрение, определение сочетаемости слов, а также влияние жанров, регистров и модальности текстов на характер отношений между ними представляется весьма важным в решении профессиональных задач. Сравнительный анализ мер ассоциации *t-score*, *MI-score* и *Log Dice* корпуса *English Web Corpus* (*enTenTen*) и его подкорпусов, проведенный на основе полученных нами данных для лексемы *take*, дает возможность сделать выводы о том, что функционал имеющихся статистических мер ориентирован на выявление разных характеристик биграмм, а их адекватное применение диктуется конкретной исследовательской задачей. Исследование коллокаций требует математического и лингвистического обоснования каждой МА с тем, чтобы осмысленно применять данные меры и правильно интерпретировать полученные результаты.

Библиографические ссылки

Захаров В. П., Хохлова М. В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции «Диалог–2010», Бекасово, 26–30 мая 2010 г. М.: Издательство РГГУ, 2010. Вып. 9 (16). С. 137–143.

Hunston S. Corpora in applied linguistics. Cambridge, UK: Cambridge University Press, 2002. 241 p.

McEnery T., Hardie A. Corpus linguistics: method, theory and practice. Cambridge, UK: Cambridge University Press, 2011. 294 p.