ЧАСТОТНЫЙ СЛОВАРЬ СОЧЕТАЕМОСТИ ГЛАГОЛОВ В СИСТЕМЕ КЛИОС

Д. И. Фирстов, Д. Э. Терёхин, Е. А. Наумова, А. П. Савинов, Е. В. Михалёва, Т. С. Петровская
Национальный исследовательский Томский политехнический университет,
Россия, г. Томск, пр. Ленина, 30, 634050
E-mail: firstov@tpu.ru

FREQUENCY DICTIONARY OF VERBS COLLOCATIONS IN THE KLIOS SYSTEM

D. I. Firstov, D. E. Teryokhin, E. A. Naumova, A. P. Savinov, E. V. Mikhalyova, T. S. Petrovskaya
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050
E-mail: firstov@tpu.ru

Annotation. *The abstract deals with a frequency dictionary of collocations which was created for the intelligent tutoring system KLIOS. This dictionary contains collocations of words with frequency value of using in Russian speach. The author starts by telling about functionality and structure of the KLIOS system and also indicates dictionary importance in the system. Then abstract describes in detail dictionary structure and the method of dictionary creating. According to the author this dictionary helps to solve many problems connected with texts parsing, students answer checking and generation of training exercises. In conclusion author adduces dictionary disadvantages that were identified in the process of expluatation and proposes ways of their solution.*

Система КЛИОС — это интеллектуальный тренажёр для обучения русскому языку как иностранному. В системе реализованы как упражнения, имитирующие традиционные учебные задания, так и специальные тренировочные задания, использующиеся для отработки полученных на уроках навыков. Ключевой особенностью системы КЛИОС является встроенный в неё лингвистический процессор, который используется для разносторонней проверки упражнений и генерации тренировочных заданий. Важной частью лингвистического процессора системы КЛИОС является частотный словарь сочетаемости, о котором пойдёт речь в настоящей статье. Подробнее о системе КЛИОС и об использовании лингвистического процессора для разносторонней проверки упражнений можно прочитать в статье [1].

Частотный словарь сочетаемости системы КЛИОС выполнен в виде базы данных, состоящей из четырёх таблиц: «Главное слово», «Предлог», «Зависимое слово» и «Словосочетание». Первые три таблицы организованы по одинаковой схеме и состоят из двух столбцов: уникального идентификатора слова и самого слова в текстовом виде. Каждая строка таблицы «Словосочетание» содержит идентификаторы слов, входящих в данное словосочетание, и частотность данного словосочетания.

Словарь был создан автоматически, путём обработки синтаксическим анализатором Cognitive Dwarf большого (порядка 30 ГБ) массива текстов различной тематики: тексты, используемые на подготовительных курсах, свободно распространяемая часть Национального корпуса русского языка [2], художественная литература из библиотеки Мошкова [3], новостные статьи. Из каждого предложения выделялись все глаголы и их зависимые слова: существительные, прилагательные, наречия и глаголы в инфинитиве, — а также предлоги, связывающие главные и зависимые слова. Все словосочетания записывались в базу данных. Если какое-то словосочетание уже находилось в базе, то увеличивался счётчик её частотности. При этом словосочетания из текстов, используемых на подготовительных курсах, имели наибольший вес, а сочетания из текстов библиотеки Мошкова и новостных статей — наименьший.

Это было сделано в связи с тем, что словосочетания из текстов, используемых на подготовительных курсах, являются наиболее характерными для студентов, изучающих русский язык. Но использовать одни только тексты подготовительного отделения нельзя, так как их слишком мало для создания достаточно полного словаря. В то же время, тексты из библиотеки Мошкова и из новостных статей могут содержать ошибки, опечатки, просторечные или устаревшие выражения, поэтому их приоритет должен быть ниже.

Словарь сочетаемости в системе КЛИОС используется для синтаксического анализа (при возникновении сложных ситуаций, которые нельзя разрешить с помощью синтаксических правил — подробнее о подходе к решению задач синтаксического анализа с использованием словарей сочетаемости см. [4]), при генерации ответов на вопросы преподавателей в упражнениях (вставка наиболее частотных слов в пропуски), проверке упражнений (оценка корректности использования вставленного слова) и генерации тренировочных заданий.

В ходе эксплуатации системы КЛИОС выявились недостатки словаря. Первый недостаток связан с тем, что невозможно выделить все допустимые сочетания путём обработки текстов. Каким бы большим ни был объём массива текстов, всегда можно будет найти сочетания, не вошедшие в эту выборку. Для решения этого недостатка предлагается записывать в словарь данные о семантических классах и грамматической форме сочетающихся слов. Это позволит сделать словарь гибче.

Второй недостаток связан с тем, что в словаре нет данных о сочетаемости отдельных словосочетаний. В некоторых задачах это может привести к тому, что абсолютно бессмысленные предложения будут считаться системой правильными. Например, словосочетания «время летит» и «летит в Москву» сами по себе являются вполне допустимыми, но, соединив их, мы получим бессмысленное «Время летит в Москву». Этот недостаток планируется решить путём хранения в словаре полных структур предложений, как, например, в ресурсе Framebank [5].

СПИСОК ЛИТЕРАТУРЫ

1. Горисев С. А. и др. Интеллектуальный лингвопроцессорный комплекс «КЛИОС» для обучения РКИ // Современные проблемы науки и образования. — 2013. — № 6.
2. Использование корпуса. Национальный корпус русского языка [Электронный ресурс]. — Режим доступа: http://ruscorpora.ru/corpora-usage.html — 25.02.2014.
3. Lib.Ru: Библиотека Максима Мошкова [Электронный ресурс]. — Режим доступа: http://lib.ru/ — 25.02.2014.
4. Арефьев Н. В. Методы построения и использования компьютерных словарей сочетаемости для синтаксических анализаторов русскоязычных текстов: дис. на соиск. учён. степ. канд. физ.-мат. наук. — Москва, 2012. — 188 с.
5. Цели и задачи системы Framebank [Электронный ресурс]. — Режим доступа: http://framebank.ru/ — 25.02.2014.