Формализация Словаря Академии Российской

Е. Н. Подтележникова, email: podtelezhnikova@yandex.ru

Воронежский государственный университет

Аннотация. *В данной работе описано создание базы данных Словаря Академии Российской в рамках проведения исследования по сопоставительной лексикографии русского языка. Особенности словаря – старорусская орфография, ориентация на словопроизводство – требуют применения специальных средств формализации.*

Ключевые слова: *сопоставительная лексикография, параметрический анализ, Словарь Академии Российской, формализация словаря, Python.*

Введение

В лингвистических исследованиях словарь играет огромную роль, являясь важным компонентом изучения как отдельных элементов языка, так и всей лексико-семантической системы в целом. Однако нельзя отрицать, что словарь представляет собой субъективное отражение объективной лингвистической реальности [1-2]. Степень адекватности словарного описания зависит от таких факторов, как время, пространство, размер словаря, его тип и концепция авторов.

Значимость параметрического анализа лексики для сопоставительной лексикографии состоит в учёте временного фактора, так называемого мнимого «устаревания» словаря. Методы глоттохронологии и лексикостатистики, разработанные М. Сводешом и уточненные С.А. Старостиным, свидетельствуют о том, что скорость изменения лексического ядра языка составляет 5-6 слов за тысячу лет и, соответственно, одно слово в двести лет [3-4]. Это говорит о высокой стабильности и устойчивости ядра лексической системы и подтверждает целесообразность использования результатов параметрического анализа в сопоставительной лексикографии.

1. Принципы параметрического анализа

Квантитативная лексикология Титова В.Т. [5] послужила основой для создания нового направления - лексической нуклеологии, которая «разрешает противоречие между потенциальной безграничностью лексики и реальной необходимостью её ограничения при типолого-сопоставительном исследовании лексики языков обращением к параметрическому ядру лексики (большому или малому)» [6]. Выделение ядра предполагает определение параметрического веса лексемы словаря по следующим параметрам: «употребительность,

косвенно оцениваемая по длине слова; синтагматика - богатство сочетаемости; парадигматика - богатство синонимии; эпидигматика - многозначность. Слова попадают в ядро не случайным образом, а в соответствии с высокими показателями представленности того или иного параметра на фоне всех остальных единиц словаря» [6].

Совокупность показателей формируют вес лексемы и высчитываются по формуле на всем объеме словаря. Параметрические характеристики позволяют распределить слова в словаре-источнике в интервале от 1 до 0. Дальнейшая сортировка слов по суммарному параметрическому весу в порядке его убывания позволяет получить ядро лексико-семантической системы данного языка. При этом определяется доминанта словаря - слово, набравшее максимальный параметрический вес (по сумме четырёх параметров).

2. Словарь Академии Российской

Сопоставительное лексикографическое исследование русского языка началось с учета временного фактора через сравнение современных словарей русского языка и Словаря Академии Российской (САР). Параметрический анализ "Словаря русского языка в 4-х томах" под редакцией А.П. Евгеньевой уже проведён [7]. В данной статье описан первый, самый трудоёмкий этап параметрического анализа САР – создание базы данных.

Словарь Академии Российской (1789 г.) представляет собой первый академический толковый словарь русского языка, содержит 43357 слов. Именно с него началось формирование литературных норм современного русского языка. Инновационный характер словаря проявляется в формировании словника, структуре словаря в целом и отдельной словарной статьи, основной концепции – ориентирование на словопроизводство. Значимость САР состоит в том, что дальнейшее развитие русской академической лексикографии шло с учетом опыта его создания.

3. Особенности создания базы данных САР

Основная проблема при распознавании САР состоит в достаточно плохом качестве печати и результатов сканирования. Традиционно лучший результат показывает программа FineReader, с помощью которой все документы сканируются и форматируются в электронный формат максимально быстро.

После распознавания программа выдала читаемый для человека, но не для других программ текст. Огромное количество цифр вместо букв, буквы разного регистра и начертания. Было принято решение использовать утилиту Afterscan, которая подходит для работы со

старорусской орфографией. Преимущества AfterScan состоят в возможности загрузки дополнительных словарей для более точного распознавания текста, возможности форматирования текста и работе с текстом через внешний редактор. Применительно к САР с помощью AfterScan были устранены практически все ошибки, выровнен текст, правильно расставлен регистр.

Для создания базы данных САР были использованы язык программирования Python и Microsoft Excel - стандартные инструменты, которые наряду со многими другими [8] активно используются в лингвистических исследованиях. Однако ценность описанного подхода состоит в том, что комбинация этих инструментом была впервые применена для параметрического анализа лексики и существенно упростила работу. Ниже представлен код в Python (техническая поддержка – Лихобаба Д.).

```
#!/usr/bin/python
import smtplib
import base64
import os
import sys
import xlswriter
import xlwt
import datetime
import MySQLdb
from pyh import *
from email.MIMEMultipart import MIMEMultipart
from email.MIMEText import MIMEText
db = MySQLdb.connect("192.168.1.118","stp","stp","STP")
cursor = db.cursor()
query = ("""select * from stp_automation_output""")
cursor.execute(query)
myresults = cursor.fetchall()
workbook = xlwt.Workbook()
worksheet = workbook.add_sheet("My Sheet")
#date_format = workbook.add_format({'num_format': 'd mmmm yyyy'})
bold = workbook.add_format({'bold': 1})
worksheet.write('A1','Sno',bold)
worksheet.write('B1','function_name',bold)
worksheet.write('C1','input1',bold)
worksheet.write('D1','input2',bold)
```

```
worksheet.write('E1','input3',bold)
worksheet.write('F1','Expected_output',bold)
worksheet.write('G1','Actual_output',bold)
row = 1
col = 0
workbook.save()
```

Результатом выполнения кода является база данных в Microsoft Excel, структура которой соответствует задачам исследования. В отдельных столбцах размещены лемма старорусская и современный вариант леммы, необходимый для последующего соотнесения словаря со словарями современного русского языка. Также вынесены в отдельный столбец пометы, которые в дальнейшем позволят выделить номинативную лексику, что является одним из требований параметрического анализа. Номер значения и толкование необходимы для проведения эпидигматической стратификации, речения - для синтагматической.

Заключение

Проведённая обработка Словаря Академии Российской, с одной стороны, позволила выявить его особенности и трудности формализации, с другой – показала эффективность использованных инструментов и их перспективность для параметрического анализа. Ориентирование САР на словопроизводство, что проявляется в приведении каждого деривата в отдельной словарной статье, потребует применения других средств формализации для использования материала САР в сопоставительной лексикографии.

Список литературы

1. Кретов А. А. Общая лексикология : учебное пособие / А. А. Кретов, Е. Н. Подтележникова. – Воронеж : Воронежский государственный университет, 2010. – 88 с.
2. Воевудская О. М. Об использовании лексикографических источников в лингвистических исследованиях / О. М. Воевудская // Филологические науки. Вопросы теории и практики. – № 10 (40). – Тамбов: Грамота, 2014. – С. 73-76.
3. Starostin S. Methodology of Long-Range Comparison [Электронный ресурс] : Режим доступа : http://starling.rinet.ru/Texts/method.pdf
4. Swadesh M. Towards Greater Accuracy in Lexicostatistic Dating // International Journal of American Linguistics. – № 21. – 1955. – P. 121-137.

5. Титов В. Т. Общая квантитативная лексикология романских языков / В. Т. Титов. – Воронеж : Изд-во Воронежского гос. ун-та, 2002. – 240 с.

6. Меркулова И. А. Лексическая нуклеология славянских языков / И. А. Меркулова. – автореферат дисс. доктора филол. н. – Тверь, 2018. – 35 с.

7. Стародубцева Ю. А. Параметрическое ядро лексики русского языка по данным "Словаря русского языка в 4-х томах" под редакцией А.П. Евгеньевой (2-ое изд.) / Ю. А. Стародубцева. – автореферат дисс. к. филол. н. – Воронеж, 2018. – 20 с.

8. Подтележникова Е.Н. Анализ лексики произведения Даниила Андреева «Роза Мира» на основе онтологического подхода / Е.Н. Подтележникова, Р.В. Шмальц // Вестн. Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. - Воронеж, 2017. - № 4. - С. 154-158.