УДК 81’33’373
Ольга Валерьевна Донина, к. филол. н.
Воронежский государственный университет, Воронеж, Россия
эл. почта: olga-donina@mail.ru

Olga Valerievna Donina, Cand. of Sc. (Philology)
Voronezh State Linguistic University, Voronezh, Russia
e-mail: olga-donina@mail.ru

КОНТРАСТИВНОЕ ИССЛЕДОВАНИЕ ИДИОМОВ ПРИ ПОМОЩИ ОНЛАЙН-ТЕХНОЛОГИЙ

В статье проводится сравнительный анализ методик, применяемых для определения близости языков и диалектов. Проведенное сопоставление позволило выявить, что при помощи веб-приложения GabMap можно эффективно сравнивать идиомы даже на небольшом массиве данных.
Ключевые слова: идиом, язык, диалект, диалектология, диалектометрия, разграничение языка и диалекта, дистанция Левенштейна

COMPARING LANGUAGES WITH THE HELP OF ONLINE TECHNOLOGIES

The article provides a comparative analysis of the methods used to determine the proximity of languages and dialects. The comparison made it possible to reveal that using the GabMap web application, languages and dialects can be effectively compared even on a small data set.
Key words: language, dialect, dialectology, dialectometry, language variation, Levenshtein distance

Вопрос, является ли некоторая языковая разновидность языком или диалектом, относится к одной из наиболее сложных проблем лингвистики, причем последствия разграничения могут выходить далеко за её пределы. Если строгого выбора в обозначении конкретной разновидности языка лучше избе-жать, лингвисты обычно используют термин идиом [1].

С развитием компьютерной лингвистики появляются новые методы, которые показывают большую чувствительность к определению разграничения языков и диалектов и могут значительно упростить работу лингвистов. Одной из проблем в разграничении идиомов является то, что многие варианты мало изучены. Поэтому цель данной работы — выяснить, возможна ли и насколько эффективна работа нового веб-инструмента GabMap, применяемого лингвистами для разграничения языка и диалекта, на основе малого количества данных. Для оценки эффективности работы этого инструмента произведено его сравнение с давно существующим и доказавшим свою результативность методом – с дистанцией Левенштейна [2].

В качестве исследуемых языков были выбраны романские языки – французский, испанский, итальянский, португальский и румынский. Выбор пал именно на эти языки, так как цель данной работы – проверить эффективность работы веб-приложения, а не разграничить идиомы, что будет проще сделать на основе ранее изученных идиомов.

В качестве анализируемого текста был использован набор из 23 слов: названий семи дней недели, двенадцати месяцев и четырех времен года. Эти слова были использованы, так как они есть в каждом из анализируемых

языков и являются одними из базовых слов романских языков, что делает их сравнение и анализ возможным. Выбранные слова также не имеют синонимов, что позволяет получить более точные результаты исследования.

Сопоставление результатов расчета дистанции Левенштейна вручную и вычисления дистанции Левенштейна при помощи программы GabMap показало, что коэффициент корреляции между этими данными очень высок и составляет 0,99. Небольшие различия могут быть связаны с тем, что при ручном подсчете в отличие от инструмента GabMap не учитывались диакритические знаки.

Для оценки эффективности применения рассматриваемых методов при небольшом объеме данных сравним полученные результаты с данными сервиса eLinguistics.net (Quantitative comparative linguistics. URL: http://www.elinguistics.net) – полностью компьютеризированной модели для сопоставительной лингвистики, количественная оценка языковых отношений в которой основана на базовом словарном запасе и генерирует автоматическую классификацию языков по семействам и подсемействам. Коэффициент корреляции между полученными нами данными на основе анализа 23 слов и данными из сервиса eLinguistics.net составляет 0,95, что говорит о высокой схожести этих показателей и, соответственно, об эффективности использования рассматриваемых методов для анализа близости идиомов даже на небольшом объеме данных.

Веб-приложение GabMap также позволяет проводить кластеризацию, визуализацию и построение карт на основе вводимых данных, что будет продемонстрировано в рамках доклада.

На основании изложенного выше можно сделать вывод, что веб-приложение GabMap предоставляет лингвистам широкий круг возможностей по изучению идиомов и позволяет облегчить работу лингвиста, автоматизируя процесс анализа данных. Этот ресурс позволяет изучать идиомы в том числе при отсутствии представительных массивов данных, что дает возможность ученым-лингвистам исследовать малоизученные идиомы и сравнивать малоресурсные языки.

ЛИТЕРАТУРА

1. Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. М. : БРЭ, 2002. 709 с.

2. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. 1965. Т. 163, № 4. С. 845–848.