Возможность сравнения идиомов на малом объеме данных при помощи приложения GabMap

О.В. Донина, email: olga-donina@mail.ru[1]
Т.О. Сигаева, email: tanyasigaeva12@gmail.com

Воронежский Государственный Университет

Аннотация. *В настоящее время лингвистика стремительно развивается, особенно в области изучения вариантов языка, т.е. идиомов. 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 маленьком массиве данных. Это позволит исследовать малые языке, используя методику компьютерной лингвистики.*

Ключевые слова: *идиом, язык, диалект, диалектология, диалектометрия, разграничение языка и диалекта, дистанция Левенштейна.*

Введение

Вопрос, является ли некоторая языковая разновидность языком или диалектом, относится к одной из наиболее сложных проблем лингвистики, причем последствия разграничения могут выходить и далеко за её пределы. Если строгого выбора в обозначении конкретной разновидности языка лучше избежать, лингвисты обычно используют термин идиом [Лингвистический энциклопедический словарь 2002].

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 нашей работы — выяснить, возможна ли и насколько эффективна работа нового инструмента, применяемого лингвистами для разграничения языка и диалекта (веб-инструмент GabMap), 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.

Под дистанцией Левенштейна понимается минимальное количество операций удаления, вставки и замены символов, необходимое для преобразования одной строки в другую. Метод назван в честь советского математика В.И. Левенштейна, который рассмотрел показатель расстояния в 1965 году [Левенштейн 1965].

Веб-приложение GabMap было создано П. Клейвегом под руководством Дж. Нербонна. Gabmap разрабатывался с конца 2010 года и впервые опубликован на Github 4 июня 2011 года. Gabmap предлагает широкий спектр возможностей обработки языковых данных [Nerbonne 2011]. О некоторых из них будет рассказано ниже.

1. Сбор данных

В качестве исследуемых нами языков были выбраны романские языки – французский, испанский, итальянский, португальский, румынский. Выбор пал именно на эти языки, так как наша цель – проверить эффективность работы веб-приложения, а не разграничить идиомы, что будет проще сделать на основе ранее изученных идиомов, чтобы сравнить полученные нами результаты с уже существующими данными.

В качестве анализируемого текста был использован набор слов, по аналогии со списком Сводеша. Наш список состоит из 23 слов: названий семи дней недели, двенадцати месяцев и четырех времен года. Мы использовали именно эти слова, так как они есть в каждом из анализируемых языков и являются одними из основных и базовых слов романских языков, что делает их сравнение и анализ возможным. Если брать за основу 100 словный список Сводеша и, например, слово «мужчина», то во французском языке его можно выразить двумя значениями – «la mâle» и «l'homme». Выбранные нами слова не имеют синонимов, что позволяет получить более точные результаты исследования.

Транскрипция списка слов производилась в соответствии с Международным Фонетическим алфавитом, который разработан и поддерживается Международной фонетической ассоциацией. Были использованы: Толковый словарь Румынского языка (Dicţ ionarul explicativ al limbii române) [DexOnline] и онлайн-словарь Wiktionary [Wiktionary] (для транскрипции португальского, французского, испанского и итальянского языков).

Изучив данные словари, мы приступили к транскрипции слов, согласно правилам произношения соответствующего языка. Таким образом, мы получили данные, которые показаны в таблицах ниже (таблицы 1-4). Первая колонка представляет собой название языка (Ф. –

французский, Ис. – испанский, П. – португальский, Ит. – итальянский, Р. – румынский), остальные – непосредственно транскрипцию слов.

Таблица 1

Транскрипция дней недели

	Пн.	Вт.	Ср.	Чт.	Пт.	Сб.	Вскр.
Ф.	lœ̃di	maʁ d i	mɛ ʁ kʁ ədi	ʒ ødi	v□̃dʁ ədi	samdi	dim□̃ ʃ
Ис.	ˈ lune s	ˈ ma ɾ tes	ˈ mjeɾ k oles	ˈ xweβ es	ˈ bjeɾ nes	sabado	doˈ miŋg o
П.	ˈ fej ɾ a	terca	ˈ kwaɾ t a	ˈ kĩnta	ˈ sestu	sabado	duˈ mĩng u
Ит.	luneˈ di	marte di	merkole ˈ di	ʤoveˈ di	vener ˈ di	ˈ sabat o	doˈ mɛ ni ka
Р.	luni	marţ i	miercuri	joi	vineri	sâmbăt a	dumi nică

Таблица 2

Транскрипция месяцев 1

	Январь	Февраль	Март	Апрель	Май	Июнь
Ф.	□□̃vje	fevʁ ije	maʁ s	avʁ il	me	ʒ yœ̃
Ис.	eˈ neɾ o	feβˈ ɾ eɾ o	ˈ maɾ θ o	aβˈ ɾ il	ˈ ma□□ ʝ o	ˈ xunjo
П.	ʒ ɐ ˈ nej ɾ u	feveˈ ɾ ej ɾ u	ˈ maɾ k u	aˈ bɾ i w	ˈ maju	ˈ ʒ ũɲ u
Ит.	ʤenˈ najo	febˈ brajo	ˈ marts o	apˈ rile	ˈ maddʒo	ˈ ʤuɲ ɲ o
Р.	ianuarie	februarie	martie	aprilie	mai	iunie

Таблица 3

Транскрипция месяцев 2

	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
Ф.	ʒ ɥ ij ɛ	au	sɛ pt□̃bʁ	ɔ ktɔ bʁ	nɔ v□̃b ʁ	des□̃bʁ
Ис	ˈ xulj o	aˈ ɣ os to	seβˈ tjemb ɾ e	okˈ tuβɾ e	noˈ βje mbɾ e	diˈ θjemb ɾ e

.						
П.	ˈ ʒ u ʎ u	aˈ ɡ os tu	seˈ tẽjbɾ u	owˈ tub ɾ u	noˈ vẽjb ɾ u	deˈ zẽjbɾ u
Ит.	ˈ luʎ ʎ o	aˈ ɡ os to	setˈ tɛ mbr e	otˈ tobre	noˈ vɛ mbre	diˈ ʧ ɛ m bre
Р.	iulie	august	septembrie	octombrie	noiembri e	decembrie

Таблица 4

Транскрипция времен года

	Весна	Лето	Осень	Зима
Ф.	p□ □̃t□̃	ete	otɔ n	ivɛ ʁ
Ис.	pɾ imaˈ βeɾ a	beˈ ɾ ano	oˈ toɲ o	imˈ bjeɾ no
П.	pɾ imaˈ vɛ ɾ a	veˈ □ □̃w	owˈ tonu	˜inˈ vɛ ɾ nu
Ит.	primaˈ vɛ ra	esˈ tate	awˈ tunno	inˈ vɛ rno
Р.	primăvară	vara	toamna	iarna

2. Дистанция Левенштейна

На следующем этапе мы приступили к расчёту дистанции Левенштейна. Расчет происходил, исходя из того, что каждая операция имеет вес равный 1. Операции представляли собой удаление и замену символов. Диакритические знаки при подсчёте не учитывались. Работа производилась вручную, без использования каких-либо автоматизированных сервисов подсчета. Мы делили слово одного языка на парное слово другого языка, как показано на примере (рис. 1) французского “me” и румынского “mai” для которых дистанция равна 2.

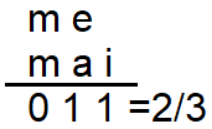

Рис. 1. Пример расчета дистанции Левенштейна

Для каждой пары языков была рассчитана дистанция Левенштейна. В результате мы получили следующие показатели (таблица 5). Цифры показывают количество замен, которые нужно произвести, чтобы получить две одинаковых строки.

Таблица 5

Дистанция Левенштейна для наших данных

	Франция	Испания	Португалия	Италия	Румыния
Франция	0	129	124	110	130
Испания	129	0	87	76	106
Португалия	124	87	0	90	111
Италия	110	76	90	0	85
Румыния	130	106	111	85	0

Таким образом, мы можем сказать, что самые похожие языки – это испанский и итальянский. Самые далекие – французский и румынский. Самым отдаленным от всех языков является французский.

3. GabMap

Далее мы приступили к работе с веб-инструментом для изучения идиомов GabMap. Для работы нам нужно было создать два файла с данными. Первый файл – карта, создать которую можно с помощью Google Earth. Нужно нарисовать границы исследуемой области и добавить метки для местоположения, где были собраны данные. Важно, что названия мест должны быть написаны точно так же, как в файле с лингвистическими данными. После создания карты в Google Earth необходимо сохранить ее в виде файла .xml или .kz, который можно загрузить непосредственно в Gabmap. С его помощью мы выделили языковые области и отметили язык, который распространен на каждой их них. Таким образом мы получили следующую карту (рис. 2):

- Зеленый – Франция, французский язык;
- Синий – Испания, испанский язык;
- Оранжевый – Португалия, португальский язык;
- Голубой – Румыния, румынский язык;
- Фиолетовый – Италия, итальянский язык.

Рис. 2. Карта расположения языков

Второй файл – данные о диалекте. Изначально таблица создается в Excel и затем переформатируется в формат txt. Пример такой таблицы показан в таблице (таблица 6) для двух слов из наших данных. Каждая строка в полученном файле – это строка в таблице. Ячейки таблицы разделяются символом табуляции. Содержимое ячейки не заключено в кавычки.

Таблица 6

Пример файла для GabMap

	Monday	Tuesday
France	lœ̃di	maʁ di
Spain	' lunes	' maɾ tes
Portugal	' fejɾ a	terca
Italy	lune' di	martedi
Romania	luni	marţ i

После загрузки созданных файлов мы получаем показатели, которые могут быть применены для дальнейшего анализа идиомов и которые будут описаны далее.

Первый показатель – общие сведения, о загруженных нами файлах. Мы получаем:

- Местоположения: 5
- Элементы (слова): 23
- Экземпляры (общее количество анализируемых слов): 115
- Символы (общее количество символов): 798
- Уникальные символы: 54

Также мы получаем сведения о каждом символе: о его количестве в загруженных данных и его распространении. Под количеством понимается то, сколько раз тот или иной символ встречается в наших файлах. Например, символ «c» в наших данных присутствует в португальском и румынском языках и встречается 5 раз. Карта для символа «c» показана на рисунке (рис. 3).

Рис. 3. Карта распространения символа «c»

Мы можем посмотреть распространение каждого элемента данных на карте, т.е. каждой транскрипции слова. Так, для элемента "lœdi" мы получаем следующую карту (рис. 4). Исходя из этой карты, мы понимаем, что данный элемент встречается только во французском языке.

Рис. 4. Карта распространения элемента "lœdi"

Нам также доступны сведения о расстояниях между каждым элементом данных. Так, например, расстояние для элемента «saturday» для Франции и Испании равно 5. То есть, чтобы из «samdi» получить «sabado» нужно произвести 5 операций по замене или удалению.

Локальная некогерентность [Nerbonne 2007] – это числовая оценка локального напряжения, присвоенная набору различий между элементами, связанных с географическими расстояниями между этими элементами. Чем меньше значение, тем меньше несогласованность и тем лучше результаты измерений. У нас этот показатель равен 0,11. Соответственно, можно сделать вывод, что измерения будут достаточно точны.

Коэффициент Альфа Кронбаха [Heeringa 2004] – коэффициент надежности измерений разницы. Альфа Кронбаха можно описать как отношение количества слов к средней корреляции между ними. Обычно значение варьируется между 0 и 1. Если значение Альфа Кронбаха очень низкое, рекомендуется добавить в анализ больше элементов, чтобы получить более надежные результаты. Для наших данных этот показатель равен 0,93. Значение близкое к 1 говорит нам о том, что географически варианты очень близки, что является истиной в нашем случае.

Расчет различий в числовом формате происходит с использованием алгоритма дистанции Левенштейна (таблица 7). Согласно полученным результатам, самые близкие языки – это итальянский и испанский. Самыми далекими языками являются испанский и французский.

Таблица 7

Различия в числовом формате

	France	Spain	Portugal	Italy	Romania
France	0	0.74921	0.691797	0.656611	0.686473
Spain	0.74921	0	0.478451	0.445	0.621758
Portugal	0.691797	0.478451	0	0.55127	0.70189
Italy	0.656611	0.445	0.55127	0	0.496161
Romania	0.686473	0.621758	0.70189	0.496161	0

Сотовые карты создаются путем рисования линий, соответствующих суммарному расстоянию между парами объектов. Линии карты обозначают степень различия идиомов. Родственные диалекты разделены светлыми линиями, а более отдаленные диалекты разделены темными линиями на карте [Goebl 1993] (рис. 5). Мы можем заметить, что, согласно этой карте, самые далекие языки – испанский и французский, а самые близкие – испанский, португальский и итальянский.

Рис. 5. Сотовая карта

4. Сравнение двух методов

Следующим этапом нашей работы является анализ полученных результатов в ходе исследования двух методов изучения идиомов и сопоставление их.

Мы сопоставили таблицу 5 и 7 между собой и выявили, что в общем показатели похожи, но есть и различия. Они могут быть связаны с тем, что, проводя расчет вручную мы не учитывали диакритические знаки, а в инструменте GabMap они учитываются. Коэффициент корреляции для этих таблиц равен 0,99. Можно сказать, что веб-приложение позволяет произвести расчет дистанции Левенштейна на базе маленького количества данных достаточно эффективно.

Заключение

На основании изложенного выше мы можем резюмировать, что современный метод изучения идиомов с помощью веб-приложения GabMap не уступает своим «старшим товарищам», а наоборот позволяет облегчить работу лингвиста, автоматизируя процесс анализа данных. Использование этого метода может позволить ученым-лингвистам намного быстрее и проще разграничивать языки и выявлять новые диалекты.

Этот инструмент позволяет изучить идиомы, не имея большого массива данных, что дает возможность ученым лингвистам исследовать малоизученные идиомы и разграничивать их.

Список литературы

1. eLinguistics.net [Электронный ресурс] URL: http://www.elinguistics.net/ (Дата обращения 15.06.22)
2. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов. // Доклады Академий Наук СССР, 1965. – С. 845–848.
3. Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. – 2-е изд., доп. – Москва: Большая Российская энциклопедия, 2002. – 709 с.
4. DexOnline Толковый словарь румынского языка [Электронный ресурс] URL: https://dexonline.ro/ (Дата обращения: 15.06.2022)
5. Gabmap — A Web Application for Dialectology. / J. Nerbonne R. Colen, Ch. Gooskens, P. Kleiweg, Th. Leinonen. // Dialectologia – 2011 – Special Issue II – P. 65-89.
6. Goebl H. Dialectometry: A short overview of the principles and practice of quantitative classification of linguistic atlas data // Contributions to Quantitative Linguistics – Salzburg, 1993 – P. 277-315.
7. Heeringa W. Measuring dialect pronunciation differences using Levenshtein distance. // Groningen Dissertations in Linguistics, 2004. -№ 46. – P. 170-177
8. Nerbonne J., Kleiweg P. Toward a Dialectological Yardstick. // Journal of Quantitative Linguistics, 2007. - №14. - P. 148-167.
9. Wiktionary Свободный словарь [Электронный ресурс] URL: https://www.wiktionary.org/ (Дата обращения: 15.06.2022).