Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2016»

Москва, 1–4 июня 2016

ПРИМЕНЕНИЕ РАСПОЗНАВАТЕЛЕЙ ФОНЕМ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ УРОВНЯ БЛИЗОСТИ ЯЗЫКОВ

Берзинь А. У. (ansis@latnet.lv)

Латвийский Университет, Рига, Латвия;
Режицкая Технологическая академия, Режица, Латвия

В докладе рассказывается о результатах применения распознавателей фонем к фонограммам спонтанной речи, с последующим применением к результатам распознавания текстовых методов задания расстояний. Эксперименты проводятся на звукозаписях латышских и латгальских говоров, но методы применимы и к любым другим идиомам.

Ключевые слова: речь, идиом, язык, диалект, распознаватель фонем, фонограмма, близость языков, расстояние между языками

USAGE OF PHONEME RECOGNIZERS FOR AUTOMATED DETERMINATION OF A SIMILARITY LEVEL BETWEEN LANGUAGES

Bērziņš A. A. (ansis@latnet.lv)

University of Latvia, Riga, Latvia;
Rēzekne Academy of Technology, Rēzekne, Latvia

The problem of automated determination of language similarity (or even defining of a distance on languages space) could be solved in different ways — working separately with phonetical transcriptions or with speech recordings. We propose a new method — a combined one: first we use a phoneme

recognizer, and for output of it we use a method, designed for phonetical transcriptions. In our case two recognizers have been tried: PhnRec (with models for Czech, Russian, Hungarian and English) and Sphinx 4 (with model for English). To calculate distance we use n-gramms' method (based on Zipf's law) — we even proved that it is metrics. Categorization was carried out using agglomerative hierarchical categorization of the smallest distance. The input data was a good-quality spontaneous speech corpus of five Latvian/Latgalian dialects.

The results were good — the calculated distances between languages corresponds to analytical understanding of similarity between them. Our conclusions: 1. Phoneme recognizers are applicable for automated language similarity determination or even distance calculation, based just on speech recordings' collections (we can call them non-annotated speech corpuses). 2. Model training language affects results, but just in a small level. It basically means that any model could be used for any pair of languages (however, it is not recommended to use a model with a training language significantly more relative to one of the pair's languages). 3. Architecture of phoneme recognizer does not affect the results of distance calculation significantly even if a "readability" level of created phonetical trancsriptions is different.

Key words: speech, language, dialect, phonemes recognizer, recording, proximity of languages, distance between languages

Введение

Мы уже длительное время занимаемся поиском различных методов оценки близости естественных идиомов. Под идиомом понимается общее название для разновидностей языка, безотносительно их точного статуса [Виноградов 1990, стр. 658]. В данной статье будут использованы как синонимы термины «идиом» и «язык» в широком понимании этого слова, т. е., включая говор, диалект, наречие, язык. Так как первоначальной и основной реализацией идиома является его устная форма, то её существование мы принимаем за обязательное условие. Наличие письменной формы несущественно.

Проблема определения близости или отдалённости идиомов имеет большое практическое значение для определения степени самостоятельности или несамостоятельности языка, при разграничении языков и диалектов, при уточнении места идиома в языковых семьях и группах, при совершенствовании информационного моделирования когнитивных процессов. В научном плане выявление близости идиомов является проблемой языковой систематики, которая пытается выработать объективные, сугубо лингвистические инструменты для определения того, являются ли два близких идиома диалектами или разными языками — а этот вопрос уже выходит за рамки лингвистики в область социального и политического. Например, в контексте языковых реалий Латвии важно выяснить, является ли латгальский самостоятельным языком или диалектом латышского языка.

Доселе мы работали отдельно — либо с фонетической транскрипцией, либо напрямую с фонограммами речи. Появление и развитие распознавателей

фонем, применимых для звукозаписей спонтанной речи, натолкнуло нас на мысль о комбинированном методе — сперва с помощью распознавателя фонем для фонограмм создаются соответствующие транскрипции, а потом к ним применяются уже имеющиеся методы определения уровня близости, проверенные на мануально транскрибированных фонотекстах.

При первом взгляде может показаться, что установка изначально неправильна, так как пока что неизвестно о существовании универсального (т. е., языконезависимого) распознавателя фонем — все нам известные распознаватели обучены (т. е., для них созданы модели) на конкретных языках и, соответственно, для тех же языков и предназначены. Но, как нам стало известно из устных бесед с коллегами, участвовавшими в разработке распознавателя PhnRec, они (ради интереса) провели эксперимент по распознаванию фонем звукозаписей языков, которые не являются для модели целевыми. Из их отзывов следовало, что они были удовлетворены результатами распознавания, при этом они отмечали, что в случае распознавания фонограмм языка обучения модели результаты были более точными.

Поэтому мы позволили себе предположить, что при распознавании фонограмм пары языков при помощи распознавателя фонем, обученного на другом — третьем — языке, ошибки и погрешность могут существенно не повлиять на фонотактические характеристики, которыми мы пользуемся для определения близости.

Гипотеза: распознаватели фонем применимы к речевым фонокорпусам для автоматической оценки степени близости языков, в том числе и нецелевых для моделей распознавателей.

Исходные данные

В нашем распоряжении были собранные (записанные) нами звукозаписи спонтанной речи пяти идиомов (латвийских говоров) — один из Курляндии: Дундажской волости, и четыре из Латгалии: Аулеи, Бальтинова, Вилека и Рудзатов. Курляндия исторически была под немецким игом, поэтому местные говоры подверглись влиянию (нижне)немецкого языка, а северокурляндские говоры, в том числе и дундажский, содержут большой субстрат ливонского языка (принадлежащего к прибалтийско-финской подгруппе финно-угорских языков). Латгалия, в свою очередь, была под поляками, поэтому в латгальских говорах присутствует влияние польского языка, также — в силу близкого соседства и наличия белорусских и старообрядческих деревень — белорусского и русского. Бальтиновский и вилекский являются говорами северолатгальскими, которые от западнолатгальского рудзатского и южнолатгальского аулейского отличаются существенно — и морфологически, и лексически.

Все звукозаписи собирались согласно заданным нами принципам сбора информации для автоматизированного анализа фонограмм [ბერზიში 2011, стр. 43–45], т. е., все записи были однородными, записанными однотипной аппаратурой (использовался динамический микрофон одностороннего

направления, фиксированый на голове информанта), в условиях уменьшенного влияния внешних шумов. Все записи были мануально вычищены, удалению подверглись все посторонние звуки и голоса, оставив только прямую речь информанта. Качество записи — 44,1 кГц / 16 битов. В зависимости от требований конкретной модели распознавателя фонем, для фонограмм выполнялось понижение частоты дискретизации либо до 8 кГц, либо до 16 кГц.

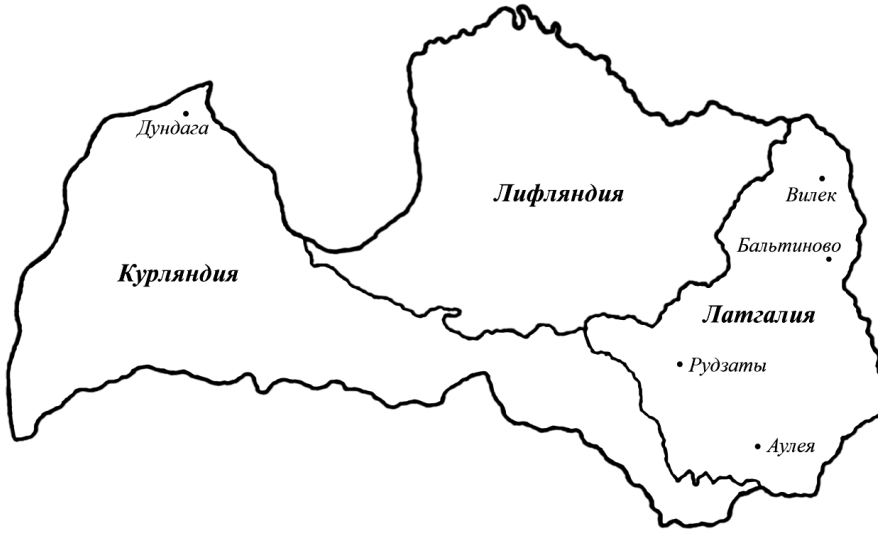

Рис. 1. Расположение центров распространения говоров на карте Латвии

Таблица 1. Характеристика набора фонограмм, используемого в эксперименте

Говор	Минут	Информантов	Мужчин	Женщин
Аулея	95	14	8	6
Бальтиново	140	23	9	14
Дундага	161	17	4	13
Рудзаты	246	28	11	17
Вилек	238	30	11	19

Распознаватель фонем PhnRec

Распознаватель фонем PhnRec разработан в Группе обработки речи Факультета иформационных технологий Брненского технического университета [BUT Speech@FIT 2016]. Главная его особенность заключается в учитывании

длинного временного контекста (до нескольких сот милисекунд)[1]. Выделение характеристик речи основано на разбиении временного контекста[2], в качестве классификатора используются искуственные нейронные сети[3], а декодировка строк фонем проводится при помощи алгоритма Витерби[4].

В пакет программы включены уже обученные модели чешского, английского, русского и венгерского языков. Мы решили воспользоваться ими всеми, дабы иметь возможность сравнения.

Распознаватель фонем Sphinx

В состав прораммного пакета Sphinx 4 включена программа pocketsphinx, работающая в том числе и как распознаватель фонем [CMU Sphinx 2015].

К сожалению, нам не удалось найти публикацию, описывающую алгоритм распознавания фонем в Sphinx 4, а разбирать код самим не было времени. Из общей документации удалось понять, что программа при обучении модели языка создаёт скрытую модель Маркова для каждой фонемы (т. е., длинный временной контекст не учитывается), а декодировка проводится не только при помощи «классического» алгоритма Витерби, но и по алгоритму Бушдерби[5].

По умолчанию в пакет включена только модель английского языка. Доступны модели других языков от внешних разработчиков, но мы решили воспользоваться только моделью от разработчиков самого распознавателя.

Метод n-грамм

В 1994 г. В. Канвар и Дж. Тренкл предложили пользоваться частотными списками n-грамм для категоризации текста [Canvar 1994, p. 161–165]. Его суть заключается в том, что по закону Ципфа множество слов (в нашем случае — знакосочетаний или звукосочетаний) можно упорядочить по частоте пользования ими.

---

[1] *The accuracy comes from modelling of long temporal contexts for phonemes (few hundreds of milliseconds).* [Schwarz 2008, p. 1]

[2] *If we are not able to classify long trajectories in the feature space because there are simply many of them and very big portion was not seen during training, let us to split the trajectores into more parts.* [Schwarz 2008, p. 35]

[3] *The artificial neural network is a discriminatively trained classifier that separates classes by hyperplanes.* [Schwarz 2008, p. 11]

[4] *Output posterior vectors are concatenated, transformed by logarithm and sent to another (merging) neural network trained again to deliver phoneme posteriors. Finally, the phoneme posteriors are decoded by a Viterbi decoder and strings of phonemes are produced.* [Schwarz 2008, p. 36]

[5] *Search in Sphinx-4 can be performed using the conventional Viterbi algorithm, or a more general algorithm called Bushderby, which performs classification based on free energy, rather than likelihoods.* [Lamere 2003, p. 1181]

В 2004 году мы применили данный метод для сравнения и определения степени близости балтийских говоров [Берзиньш 2004, стр. 65–71]. Суть идеи была в том, что в процессе нахождения наиболее близкого (вероятностного) соответствия, для каждого частотного списка-модели расчитывается некое число (сумма разницы местоположений фонемы в списках, если она присутствует в обоих списках, или размера списка иначе), характеризующее его степень близости с частотным списком текста на входе. Так как результаты применения метода были положительными, то мы вправе им воспользоваться и сейчас.

Размерность используемых n-грамм (т. е., n) выбирается эмпирически. Мы пользовались n = 1..5, т. е., от униграмм то квинтаграмм, но фактически в наших списках в основном присутствуют униграммы, биграммы и триграммы, так как квадриграммы и квинтаграммы встречаются слишком редко. Это, конечно, зависит и от размера (количества записей) списка: его надо задать таким, дабы он по возможности лучше представлял текст, по которому создан, но в тоже время во всех текстах «хватило» n-грамм, т. е., мы достигли заданного размера. В нашем случае мы задали N = 400.

Следует отметить, что фонотактические данные в себе содержат и фонетическую, и морфологическую, и лексическую и даже синтаксическую информацию. На пропорцию учитывания разных типов информации также влияют параметры n и N (например, если n = 1 или N существенно меньше количества фонем в идиоме, то будет учитываться только фонетическая информация).

$$d(a,b)=\sum_{i=1}^{N}\begin{matrix}|i-j|;\exists j: a_i = b_j\\ n;иначе\end{matrix}$$; где N — размер списков

Формула 1. Расстояние между списками n-грамм

Утверждение 1. *Заданное подобным образом расстояние является метрикой.*

Давайте в этом удостоверимся. Сперва проверим аксиому тождества. Очевидно, если a = b, то такое j всегда будет находится, причём оно будет равно i. Из чего следует, что слагаемые суммы всегда будут нолями, что и требовалось доказать.

Теперь проверим аксиому симметрии. Число несовпадающих n-грамм от порядка аргументов независит. В свою очередь, в случае нахождения n-граммы в обоих списках, симметричность обеспечивается тем, что берётся не сама разница, а её модуль.

Выполнение аксиомы треугольника не столь очевидно, но всё же усматриваемо. Для доказания исполнения неравенства треугольника для суммы, достаточно его доказать для всех слагаемых. Рассмотрим все возможные случаи слагаемых наших сумм:

1) n-грамма присутствует во всех трёх списках. В таком случае всегда $|i-j| \le |i-k| + |k-j|$.
2) n-грамма не присутствует в одном из двух начальных списков. В таком случае $N \le N + |k-j|$.

3) n-грамма не присутствует в третьем списке: $|i - j| \le 2N$.
4) n-грамма не присутствует в обоих начальных списках: $x \le 2N$, где x — любое слагаемое суммы.
5) n-грамма не присутствует в одном из двух начальных списков и в новом списке: $N \le 2N$.

Таким образом метричность расстояния доказана.

Эксперимент

Эксперимент проводился при помощи скриптов, написанных нами на языке PERL. Сначала ко всем имеющимся фонограммам интересующих нас идиомов мы применили все имеющиеся в нашем распоряжении распознаватели и их модели, общим числом 5: PhnRec для чешского, русского, английского и венгерского, а также Sphinx 4 для английского. Потом мы полученные файлы фонем привели в вид двухбайтового фонотекста, всех информантов одного идиома обьединив в один файл. Затем, с помощью исправленной и модифицированной нами PERL-программы TextCat, рассчитали расстояния между полученными фонотекстами.

Таблица 2. Расстояния между идиомами, распознаватель: PhnRec, язык модели: чешский

	Аулея (Auleja)	**Бальтиново (Baļtinova)**	**Дундага (Dundag)**	**Рудзаты (Rudzātys)**	**Вилек (Vileks)**
Аулея	0	35 104	50 996	31 718	41 810
Бальтиново	35 104	0	52 847	31 787	30 988
Дундага	50 996	52 847	0	49 741	48 986
Рудзаты	31 718	31 787	49 741	0	37 232
Вилек	41 810	30 988	48 986	37 232	0

Таблица 3. Расстояния между идиомами, распознаватель: PhnRec, язык модели: русский

	Аулея (Auleja)	**Бальтиново (Baļtinova)**	**Дундага (Dundag)**	**Рудзаты (Rudzātys)**	**Вилек (Vileks)**
Аулея	0	33 073	46 448	30 638	34 175
Бальтиново	33 073	0	47 752	30 989	29 439
Дундага	46 448	47 752	0	43 906	45 767
Рудзаты	30 638	30 989	43 906	0	34 626
Вилек	34 175	29 439	45 767	34 626	0

Таблица 4. Расстояния между идиомами, распознаватель: PhnRec, язык модели: венгерский

	Аулея (Auleja)	**Бальтиново (Baļtinova)**	**Дундага (Dundag)**	**Рудзаты (Rudzātys)**	**Вилек (Vileks)**
Аулея	0	36589	50677	39510	42803
Бальтиново	36589	0	47708	33217	30514
Дундага	50677	47708	0	47088	47711
Рудзаты	39510	33217	47088	0	33512
Вилек	42803	30514	47711	33512	0

Таблица 5. Расстояния между идиомами, распознаватель: PhnRec, язык модели: английский

	Аулея (Auleja)	**Бальтиново (Baļtinova)**	**Дундага (Dundag)**	**Рудзаты (Rudzātys)**	**Вилек (Vileks)**
Аулея	0	44098	52208	41165	46056
Бальтиново	44098	0	51315	37773	30283
Дундага	52208	51315	0	50077	55491
Рудзаты	41165	37773	50077	0	36577
Вилек	46056	30283	55491	36577	0

Таблица 6. Расстояния между идиомами, распознаватель: Sphinx, язык модели: английский

	Аулея (Auleja)	**Бальтиново (Baļtinova)**	**Дундага (Dundag)**	**Рудзаты (Rudzātys)**	**Вилек (Vileks)**
Аулея	0	36751	51130	38058	35900
Бальтиново	36751	0	47878	37603	30205
Дундага	51130	47878	0	49830	51087
Рудзаты	38058	37603	49830	0	34568
Вилек	35900	30205	51087	34568	0

Из таблиц видно, что предложенный подход достаточно хорошо справляется с поставленной задачей: более близкие в интуитивном и аналитическом понимании идиомы более близки и по рассчитаным нами расстояниям.

Для большей наглядности мы решили провести категоризацию идиомов по таблицам расстояний методом иерархической аггломеративной категоризации наименьшего расстояния (описание алгоритма см. в [Берзинь 2006, стр. 66]).

В результате мы получили два вида графов. Оба в отдельный подвид выделяли северолатгальские говоры (что неудивительно, ввиду их большой близости), и в отдельный вид — курляндский говор. Различия в графах проявились

в категоризации южно- и западнолатгальских идиомов (Аулея, Рудзаты): в первом случае (на моделях русского и чешского языков) они оказались выделенными в общую ветку, во втором случае (на модели венгерского и двух моделях английского языков) они остались в изолированных позициях. С точки зрения интуитивного и лингвоаналитического восприятия первый вариант более правилен — Аулейский и Рудзатский говоры являются похожими на говоры среднелатгальские, хотя и находятся по разные стороны ареала их распространения. Однако и второй граф достоин внимания, потому что зафиксированные в нём автономные позиции обоих идиомов, возможно, отражают различия между говорами южнолатгальскими (имеющими большее влияние говоров польских, белорусских, русских и верхнелитовских (аукштайтских)) и западнолатгальскими (имеющими большее влияние говоров селонских).

Рис. 2. Результаты категоризации: PhnRec, модели чешского и русского

Рис. 3. Результаты категоризации: PhnRec, модели венгерского и английского, Sphinx, модель английского

Всё же, так как русский и чешский (в отличии от венгерского и английского) являются языками славянскими, а по генеалогической классификации языков наиближайшей группой к балтийским языкам являются именно славянские языки, то волей-неволей напрашивается вывод о более качественных результатах при более родственном языке обучения модели.

Выводы

Во-первых, мы доказали, что наша основная гипотеза подтвердилась: распознаватели фонем применимы к речевым фонокорпусам для автоматической оценки уровня близости языков.

Во-вторых, мы удостоверились, что язык обучения модели распознавателя влияет на результаты, но не настолько, чтобы исключить применение. Т.е., в принципе можем пользоваться любым распознавателем для любых идиомов, но для чистоты опыта желательно, чтобы уровень родства между языком обучения модели и каждым из языков эксперимента различался не существенно (например, некорректно проводить эксперимент, если один из идиомов из той же группы, что и идиом обучения модели, а второй — из другой).

В-третьих, архитектура фонемного распознавателя на результаты определения расстояния существенно не влияет. (Хотя результаты распознавания в случае PhnRec были более «читабельными».) Предположительно, ошибки распознавателей последовательны (т. е., при похожих обстоятельствах совершаются похожие ошибки), поэтому на результатах сравнения фонотактической статистики существенно не отражаются.

В заключении отметим, что, решая вынесенную в заголовок проблему, попутно мы фактически задали метрику на пространстве языков, при чём вычисляемую автоматически, без предварительной подготовки данных, т. е., без ручного труда (что весьма перспективно с точки зрения практического применения).

Литература

1. *Schwarz, P.* (2008) Phoneme recognition based on long temporal context, Doctoral thesis, Brno, Brno University of Technology, Faculty of Information Technology.
2. *Lamere P., Kwok P., Walker W., Gouvea E., Singh R., Raj B., Wolf P.* (2003) Design of the CMU Sphinx-4 decoder. // Proceedings of the 8th European Conference on Speech Communication and Technology, Geneve, Switzerland, pp. 1181–1184.
3. ბერზინი ა. [*Berzini A.*] (2011) ინფორმაციის მოპოვების პრინციპები ფონოგრამების ავტომატური ანალიზისთვის [inp'ormats'iis mopovebis prints'ipebi p'onogramebis avtomaturi analizist'vis] = Принципы сбора информации для автоматизированного анализа фонограмм // ქართული ენა და თანამედროვე ტექნოლოგიები — 2011 [k'art'uli ena da t'anamedrove tek'nologiebi — 2011]. თბილისი [t'bilisi]: «მერიდიანი» [«meridiani»], стр. 39–46.
4. *Берзиньш А. У.* (2004) Сравнение балтийских языков методом n-грамм // Труды международной коференции «Корпусная лингвистика — 2004». СПб.: Издательство С.-Петербургского университета, стр. 65–71.
5. *Берзинь А. У.* (2006) Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера // Труды международной конференции «Диалог 2006». М.: Издательство РГГУ, стр. 65–72.
6. *Canvar, W. B., Trenkle, J. M.* (1994) N-Gram-Based Text Categorization // Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, Nevada: Information Science Research Institute, University of Nevada, pp. 161–175.
7. *CMU Sphinx* (2015) Phoneme Recognition (caveat emptor). Available @ http://cmusphinx.sourceforge.net/wiki/phonemerecognition
8. *BUT Speech@FIT* (2016) Phoneme recognizer based on long temporal context. Available @ http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context
9. *Виноградов В. А.* (1990) Идиом // Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцевой. М.: Советская энциклопедия, стр. 685.