А.У. Берзиньш

СРАВНЕНИЕ БАЛТИЙСКИХ ЯЗЫКОВ МЕТОДОМ N-ГРАММ

Вопросы близости языков[1] всегда были поводом для политических и общественных спекуляций. Конечно, речь не идёт об очевидно неродственных языках или языках, которые, например, имеют только минимальные лексические различия. Речь идёт о тех смежных областях, которые образуют не вполне родственные языки. В связи с этим нам уже давно не даёт покоя мысль о разработке некой процедуры или даже, скорее всего, набора неких процедур автоматизированной оценки близости языков.

Надо признать, что данная мысль не нова – начав интересоваться проблемой, мы натолкнулись на работы Ю. Тамбовцева. Так, в своей статье[2] автор рассчитывает фонотипологические расстояния между баскским и рядом других языков разных языковых групп по формуле евклидова расстояния в восьмимерном пространстве, измерениями которого являются частоты встречаемости разных видов согласных. Хотя результаты экспериментов Ю. Тамбовцева весьма интересны, к сожалению, они не могут нас удовлетворить: автор работает только с базовыми литературными языками, при чём не с устной речью или её фонетической транскрипцией, а с письменными текстами (книгами), отражающими определённую письменную традицию и переводимы-

[1] Здесь и далее, если специально не оговорено иначе, под языками понимаем любые людские языки или говоры, вне зависимости от их классификации, в том числе диалекты, наречия и т.п.

[2] *Tambovtsev, Y.* Phonological Similarity Between Basque and Other World Languages Based on the Frequency of Occurrence of Certain Typological Consonantal Features // The Prague Bulletin of Mathematical Linguistics. 2003. № 79-80. P. 121-126.

ми в предполагаемое звучание по некоторым правилам. Безусловно, как запись, так и последующий перевод достаточно условны: строго говоря, они создают некий искусственный, отличный от естественного язык. Сравнение нескольких полученных подобным образом языков вполне корректно, однако мы должны искать другие способы записи, так как нас интересуют в том числе и бесписьменные языки. Вполне естественной является мысль работать с фонетической транскрипцией интересующих нас языков. Конечно, в таком случае мы сразу сталкиваемся с вопросом, какой фонетической записью пользоваться, ведь существует значительное число разных стандартов, у каждого из которых имеются недостатки (к обсуждению этого вопроса мы вернёмся чуть позже). Вообще, выдвижение письменного языка на доминирующие позиции по отношению к устному языку является большой общественной проблемой современности, из которой вытекает множество неудобств (например, неумение правильно произносить редко встречающиеся слова и заимствованные имена), но речь сейчас идёт не об этом. Предложенный Ю. Тамбовцевом метод применим и к фонотекстам, поэтому мы надеемся, что подобные эксперименты будут когда-либо проведены. Однако на сей раз мы расскажем о другом методе, применённом нами ещё до ознакомления с работами Ю. Тамбовцева.

В 1994 г. В. Канвар и Дж. Тренкл предложили пользоваться частотными списками *n*-грамм[1] для категоризации текста[2]. Дело в том, что по закону Ципфа, множество слов (в нашем случае – знакосочетаний или звукосочетаний) можно упорядочить по частоте пользования ими. В. Канвар и Дж. Тренкл предлагают со-

[1] В данном случае понятием n-граммы обозначаются подпоследовательности (слов текста) длинной в n символов.

[2] *Canvar, W.B., Trenkle, J.M.* N-Gram-Based Text Categorization // Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval. Las Vegas (NV), 1994. P. 161–175.

ставлять частотные списки n-грамм для различных текстов и путём сравнения этих списков и частотного списка n-грамм входного текста определять категорию, к которой тот относится. Таким образом можно успешно автоматически определять язык, кодировку и даже тематику текста. Что в этом общего с нашими задачами? Дело в том, что при сравнении списков n-грамм рассчитывается число[1], характеризующее степень сходства текстов. Поэтому нас заинтересовал вопрос, можно ли применить этот метод для определения сходства языков? Очевидно – да, но опять-таки при условии пользования одностандартной фонетической записью, так как различные традиции записи, а так же различные кодировки, которые только способствуют решению задачи категоризации, для нас являются неприемлимыми. Кроме того, требования к категоризации как таковой существенно слабее требований к определению близости, поэтому можно предположить, что для получения результатов, устраивающих нас по точности, нам придётся работать с текстами гораздо большего объёма.

Существующие стандарты хранения фонетических данных на ЭВМ нас не устраивают, так как все нам известные (различные локальные, Unicode, SAMPA, IPA/ASCII) либо содержат не все звуки[2], либо описывают каждый звук разным количеством символов (т.е. либо байтов, либо пар байтов). Нам же было бы гораздо удобнее работать с текстом, состоящим из одноразмерных элементов, т.е. каждый звук должен описываться одинаковым количеством единиц информации. Так как одного байта (256 пермутаций) для описания всех звуков человеческой речи не достаточно, то наиболее удобно их описывать двумя байтами (65536 перму-

[1] Сумма разниц номеров местонахождений в списках (если n-грамма присутствует и там, и там) и длин списков (иначе).

[2] Здесь и далее за нежеланием лишний раз пользоваться заимствованными словами, называем звуками те единицы, которые терминологически принято называть фонемами.

тациями). Конечно, теоретически возможно и более мелкое звуковое деление, но нам оно представляется нецелесообразным, так как при слишком мелком делении при транскрибировании возрастает ошибочность и зависимость от индивидуального восприятия специалистов, производящих транскрипцию. Кроме того, насколько нам известно, более мелким делением не пользуются и при записи в уже существующих стандартах фонетической транскрипции. Очевидно, в перспективе будет целесообразно разработать специальные шрифты Unicode, а так же редактор и конверторы к ним. На данный момент для работы с фонотекстами балтийских языков мы ограничились использованием двухбайтового псевдокода: каждый встречаемый звук описывали двумя ASCII-символами по определённой схеме.

К сожалению, насколько нам известно, корпусы фонотекстов балтийских языков в электронном виде до нынешнего времени нигде не создавались. Поэтому нам пришлось самостоятельно вводить фонотексты в ЭВМ из существующих печатных сборников[1]. В силу ограниченности по времени мы ввели в ЭВМ «минимальный набор» текстов «минимального размера»: 4 образца разных говоров Латвии (3 латгальских и 1 курляндский) объёмом от 500 до 1000 звукознаков. Далее мы изменили программу TextCat на языке PERL (написанную голландцем Г. ван Нордом в 1994 г. под впечатлением от публикации В. Канвара и Дж. Тренкла) так, чтобы она работала с двухбайтными символами вместо однобайтных, т.е., n-граммы состояли из 2n, а не n байтов.

Результаты сравнения приведены в табл. 1.

[1] *Latviešu izlokšņu teksti*. / Sast. Marta Rudzīte. Rīga: P. Stučkas Latvijas Valsts universitāte, 1963; *Augšzemnieku dialekta teksti:* Latgaliskās izloksnes. / Sast. N. Jokubauska. Rīga: Zinātne, 1983. *Latviešu izlokšņu teksti*. / Sast. Benita Laumane. Liepāja: Liepājas Pedagoģiskā akadēmija, 2000.

Таблица 1. Расстояния
между входными и сравнительными текстами

Язык входного текста	**Язык сравнительного текста**	**Условное расстояние между входным и сравнительным текстами**
Шкылбаны (Škylbāni)	Бальтиново	118344
	Нерза	125831
	Джукстэ	128467
Бальтиново (Baļtinova)	Шкылбаны	118336
	Нерза	118369
	Джукстэ	130354
Нерза (Nierza)	Бальтиново	118353
	Шкылбаны	125811
	Джукстэ	133562
Джукстэ (Džūkste)	Шкылбаны	128501
	Бальтиново	130396
	Нерза	133584

Представленные в табл. 1 данные показывают, что несмотря на предельно малый объём введённых фонотекстов, результаты нашего эксперимента полностью оправдывают наши ожидания. Так, два севернолатгальских говора соседних волостей – бальтиновский и шкылбанский – наиболее близки друг к другу, следующим по близости к ним является латгальский нерзинский, и только потом следует курляндский джукстенский. Наиболее близким к нерзинскому оказывается бальтиновский, что обосновано и географически (см. рис. 1), и лингвистически, а наиболее отдалённым, конечно, джукстенский. Наиболее близким джукстенскому из латгальских говоров оказывается шкылбанский, что тоже не удивляет – севернолатгальские говоры имеют с ним

больше общего и с лексической точки зрения, и с точки зрения звукового оформления отдельных грамматических форм.

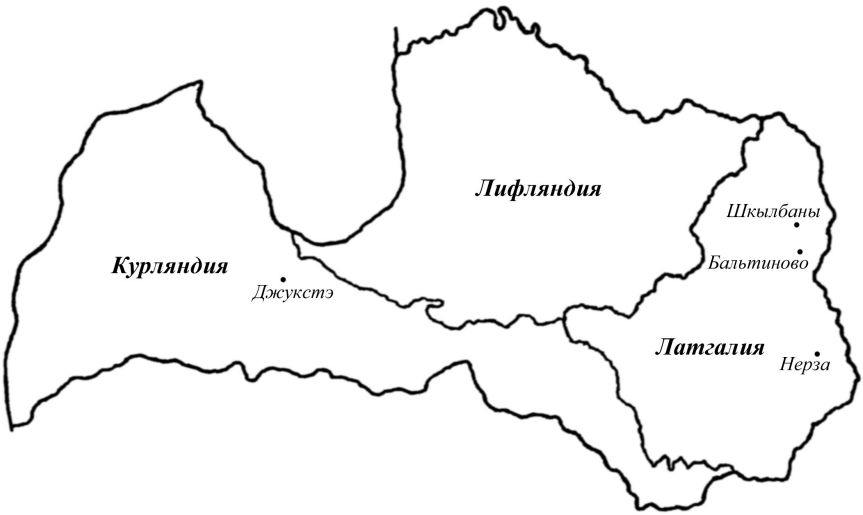

Рис. 1. Карта Латвии

Итак, при очень малых текстах и двоичной процедуре сравнения[1] мы имеем достаточно хорошие результаты. Это позволяет предполагать, что при использовании более объёмных и качественных корпусов фонотекстов, а также более скрупулёзном подходе к обоснованию и разработке процедуры сравнения результаты могут быть даже очень хорошими, и на их основании мы сможем делать вполне достоверные выводы. Поэтому можем констатировать, что разработки в данном направлении следует продолжить. Необходимо предпринять следующие шаги:

1) разработать двухбайтную кодовую таблицу обозначения всех возможных звуков человеческой речи;

[1] Строки частотных списков n-грамм мы сравниваем по двоичному принципу, т.е. совпадение – 0, несовпадение – 1.

2) создать к ней шрифты отображения, редактор ввода и конверторы для других стандартов фонетической записи;

3) создавать в ней корпусы всевозможных говоров мира, в частности для наших ближайших экспериментов – балтийских и славянских;

4) на пространстве всех звуков (т.е. в нашей таблице) ввести некую меру, определяющую фонетическое расстояние между любыми двумя звуками, обоснованную физическими свойствами звуков и физиологическими характеристиками их произношения;

5) разработать процедуру сравнения языков, зависящую от этой меры.