ИЗМЕРЕНИЕ ФОНОМОРФОЛЕКСИЧЕСКОГО РАССТОЯНИЯ МЕЖДУ ЛАТЫШСКИМИ НАРЕЧИЯМИ ПУТЁМ ПРИМЕНЕНИЯ РАССТОЯНИЯ ВАГНЕРА-ФИШЕРА

MEASUREMENT OF PHONOMORPHOLEXICAL DISTANCE BETWEEN LATVIAN DIALECTS USING WAGNER-FISCHER DISTANCE

А.У. Берзинь *(ataols@latnet.lv)*

Латвийский университет, Рига

В докладе рассказывается о попытке измерения фонетической, морфологической и лексической близости языков на материале диалектологического атласа латышских говоров. Сначала проводится эксперимент с применением расстояния Левенштейна, а потом – расстояния Вагнера-Фишера с фонетически обоснованной функцией цена замены, в конце – сравниваются результаты и делаются выводы.

Введение

Мы уже несколько лет интересуемся возможностями автоматизированного измерения близости языков. Ввиду доступности материалов начальные разработки ведём на материале балтийских наречий, хотя эти методы применимы и к другим языкам. В 2004 году мы представили доклад на конференции „Корпусная лингвистика“, в котором рассказывалось об измерении близости наречий по частотным спискам n-грамм [4]. Суть метода заключается в создании текстов (т.е., корпусов) разных языков в унифицированной фонетической транскрипции, рассчёта частотного списка n-грамм для каждого из них и определении степени близости пар языков как суммы разниц местоположений одинаковых n-грамм в частотных списках данных языков (если n-грамма присутствует только в одном списке, то разница местоположения определяется как длина списка). Данный метод показал хорошие результаты даже на небольших „пробных“ корпусах, а на корпусах побольше, очевидно, повёл бы себя ещё лучше, но создание таковых, т.е., цифровых корпусов текстов наречий в унифицированной фонетической транскрипции, является очень трудоёмкой задачей.

Это и было причиной, по которой мы занялись поиском других методов, трудовой вклад в подготовку данных для которых был бы существенно меньше. Вполне естественным способом уменьшения обьёма сравниваемых корпусов был бы переход от произвольных корпусов к параллельным. Конечно, сбор информации для параллельных корпусов тоже требовал бы больших трудозатрат, но оказалось, что данные уже имеются, правда в тетрадках, т.е., их необходимо только оцифровать. В 1954 году Институт языка и литературы Академии наук Латвийской ССР издал „Программу сбора материалов диалектологического атласа латышского языка“ [1], которая по сути является: 1) спецификацией системы фонетической записи; 2) списком вопросов, ответы на которые требуется получить по каждому наречию. Данный вопросник содержит 670 статей, из которых 103 составлены для отражения фонетических, 160 – морфологических, 107 – синтаксических, а 300 – лексических различий.

Далее осталось найти подходящий метод сравнения. Оказалось, что с похожими данными уже работали и ирландские [2], и голландские [3] коллеги, поэтому мы решили воспользоваться рекоммендованным ими методом – определением расстояний между языками как суммарного или среднего расстояния Левенштейна между одинаковыми словами (строками в фонетической записи), а далее – иерархической аггломеративной категоризацией метода среднего расстояния для анализа полученных результатов.

Расстояние Левенштейна

Расстоянием Левенштейна называют цену перередактирования одной текстовой строки в другую, где действия – удаление символа, добавление символа или замена символа на другой, а цена каждой операции равна единице. Т.е., в данном случае цена равна минимальному количеству действий редактирования. Например, чтобы из слова „диалог“ получить слово „одеяло“, нам придётся провести следующие действия: добавить „о“, заменить „и“ на „е“, заменить „а“ на „я“ и стереть „г“. Итого расстояние Левенштейна между данными словами равно 4.

Обычно для вычисления расстояния Левенштейна пользуются алгоритмом динамического программирования, описание которого в псевдокоде приводим ниже:

```
целое_число РасстояниеЛевенштейна(строка1, строка2)
  дс1 := длинна(строка1);
  дс2 := длинна(строка2);
```

```
  # d - матрица размером дс1*дс2
  целое_число d[0..дс1, 0..дс2];
  # вспомогательные переменные
  целое_число i, j, цена;

  для i от 0 до дс1
    d[i, 0] := i;
  для j от 1 до дс2
    d[0, j] := j;

  для i от 1 до дс1
    для j от 1 до дс2
      # строки есть массивы символов
      если строка1[i] = строка2[j]
        тогда цена := 0
        иначе цена := 1;
      d[i, j] := минимум(
                  # стирание
                  d[i-1, j] + 1,
                  # вставка
                  d[i, j-1] + 1,
                  # замена
                  d[i-1, j-1] + цена
                 );

  возвратить d[дс1, дс2];
конец РасстояниеЛевенштейна;
```

Алгоритм иерархической категоризации проще описать словами:

1. Пусть у нас имеются n обьектов, которые мы хотим распределить по категориям, а также задано расстояние d между обьектами.

2. Создадим n начальных одноэлементных категорий - для каждого обьекта по одной, и определим расстояния между ними как расстояния между состовляющими их обьектами.

3. Найдём две наиближайшие (с наименьшим расстоянием) категории и обьединим их в одну так, что новая будет содержать все элементы обеих. Таким образом количество категорий уменьшится на одну.

4. Рассчитаем расстояния между новой и остальными категориями по формуле:

а) в случае метода наименьшего расстояния:

$$\min\{d(x,y): x \in A, y \in B\}$$

б) в случае метода наибольшего расстояния:

$$\max\{d(x,y): x \in A, y \in B\}$$

в) в случае метода среднего расстояния:

$$\frac{1}{card(A)\cdot card(B)} \sum_{x \in A} \sum_{y \in B} d(x,y)$$

(Где A - новосозданная категория, а B - другая категория, расстояние до которой рассчитываем. Кардинальное число категории есть её мощность, т.е., количество обьектов в ней.)

5. Если количество категорий больше 1, вернёмся к шагу №**3**, иначе - завершим работу.

Мы ввели небольшие выборочные данные (13 слов, характеризующих фонетические, 10 – морфологические и 8 – лексические различия) 13 латышских наречий разных областей Латвии *(рис.1)*. Дабы облегчить ввод (с любой консоли) и обработку (существенна фиксированность количества байтов в одной фонеме), данные вводились в фонетической псевдозаписи, в которой каждая фонема обозначалась 6 байтами (1, 2 – сама фонема, 3 – долгота, 4 – интонация, 5 – ударность, 6 – слоговость). Например, слово в латышской фонетической транскрипции записывающееся „čārmyûkša“, у нас хранится в виде „č02010a04311r02010m02000y02D01u02201k02000š02000a02001“.

Рис. 1. Говоры на карте Латвии

На языке ПЕРЛ мы написали программу, измерящую средние расстояния Левенштейна между парами наречий (т.е., среднее по парам слов) и далее проводящую категоризацию.

После нескольких пробных экспериментов мы констатировали, что метод иерархической аггломеративной категоризации среднего расстояния иногда работает несоответствующе (т.е., в случаях, когда очевидна близость двух говоров – одного в группе, а другого - вне, по среднему арифметическому ближе оказывается другой говор), и что по свойствам рассматриваемых нами обьектов лучше пользоваться иерархической аггломеративной категоризацией наименьшего расстояния, т.е., когда две категории обьединяются в одну, расстояние между новой, обьединённой категорией и остальными категориями мы определяем как наименьшее, а не среднее арифметическое, расстояние между элементами новой и элементами остальных категорий.

	Шкылбаны (Škylbāni)	Цыбла (Cybla)	Зимер (Zīmeris)	Аташиня (Atašine)	Алшванга (Ališvang)	Уолайнэ (Olaine)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rite)	Деры (Ģeri)
Шкылбаны	0,000	1,230	1,769	1,769	3,307	3,076	1,153	3,230	2,769	2,769	1,846	2,076	2,923
Цыбла		0,000	2,153	2,153	3,384	3,384	1,076	3,384	2,846	3,076	2,230	2,461	3,153
Зимер			0,000	2,076	2,769	2,384	1,923	2,769	2,230	2,076	1,692	1,461	2,538
Аташиня				0,000	3,153	2,846	2,230	3,076	2,538	2,615	2,000	2,230	2,846
Алшванга					0,000	1,461	3,076	0,846	1,538	1,846	3,000	3,076	1,153
Уолайнэ						0,000	3,076	1,846	1,076	0,846	3,000	2,769	1,615
Аулея							0,000	3,076	2,538	2,846	1,923	2,000	3,153
Зуры								0,000	1,846	2,000	2,923	3,153	1,307
Руцово									0,000	1,384	2,538	2,538	1,846
Куоцены										0,000	2,692	2,615	1,307
Лейвон											0,000	1,076	2,846
Ритэ												0,000	3,000
Деры													0,000

Таб. 1. Расстояния между говорами: фонетические вопросы, расстояние Левенштейна

	Шкылбаны (Škylbāni)	Цыбла (Cybla)	Зимер (Zīmeris)	Аташиня (Atašine)	Алшванга (Ališvang)	Уолайнэ (Olaine)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rite)	Деры (Ģeri)
Шкылбаны	0,000	0,750	1,833	1,000	3,333	3,250	0,916	3,166	2,916	3,083	2,416	2,583	3,250
Цыбла		0,000	1,916	1,166	3,250	3,166	0,416	3,166	2,833	3,083	2,083	2,000	3,083
Зимер			0,000	1,833	2,416	2,333	1,833	2,500	2,750	2,333	2,083	1,916	2,750
Аташиня				0,000	3,250	3,250	1,083	3,166	2,916	2,916	2,000	2,250	3,250
Алшванга					0,000	1,500	3,250	0,750	2,333	1,916	3,333	3,083	1,916
Уолайнэ						0,000	3,166	1,833	1,583	1,333	3,166	3,000	1,750
Аулея							0,000	3,083	2,833	2,916	1,833	1,833	3,166
Зуры								0,000	2,583	2,083	3,250	3,083	2,000
Руцово									0,000	1,916	3,583	3,333	2,250
Куоцены										0,000	2,833	2,833	1,000
Лейвон											0,000	0,916	3,083
Ритэ												0,000	2,916
Деры													0,000

Таб. 2. Расстояния между говорами: морфологические вопросы, расстояние Левенштейна

	Шкылбаны (Škylbāni)	Цыбла (Cybla)	Зимер (Zīmeris)	Аташиня (Atašine)	Алшванга (Ališvang)	Уолайнэ (Olaine)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rite)	Деры (Ģeri)
Шкылбаны	0,000	1,500	4,375	2,500	5,625	6,250	1,625	5,500	5,125	6,000	4,000	5,125	6,250
Цыбла		0,000	4,875	2,750	5,750	6,500	0,750	5,625	5,500	6,250	3,250	5,250	6,500
Зимер			0,000	4,250	4,500	4,125	4,875	4,625	5,125	4,375	4,125	4,125	4,625
Аташиня				0,000	5,500	5,875	3,000	5,375	5,125	5,875	4,500	5,000	5,875
Алшванга					0,000	4,000	5,750	0,875	4,000	4,000	5,750	5,250	4,000
Уолайнэ						0,000	6,500	4,125	4,500	2,000	5,750	5,750	1,625
Аулея							0,000	5,625	5,375	6,125	3,500	5,375	6,500
Зуры								0,000	3,875	4,000	5,750	5,250	4,125
Руцово									0,000	4,375	6,375	6,000	4,375
Куоцены										0,000	5,750	5,750	1,000
Лейвон											0,000	3,250	5,875
Ритэ												0,000	5,875
Деры													0,000

Таб. 3. Расстояния между говорами: лексические вопросы, расстояние Левенштейна

Рис. 2. Результаты категоризации: фонетические вопросы, расстояние Левенштейна

Рис. 3. Результаты категоризации: морфологические вопросы, расстояние Левенштейна и лексические вопросы, расстояние Левенштейна

Расстояние Вагнера-Фишера

Результаты эксперимента оказались вполне хороши и соответствовали представлению об отдалённости или близости тех или иных наречий. Но мы всё-таки решили усложнить наш эксперимент переходом с расстояния Левенштейна на расстояние Вагнера-Фишера, т.е., вместо постоянной цены замены фонемы на другую пользоваться ценой, зависящей от близости данных двух фонем. (Например, цена замены узкого *e* на широкое должна быть меньше, чем замены на *o*, а та, в свою очередь – меньше, чем на *b*.)

Приведём алгоритм вычисления расстояния Вагнера-Фишера в псевдокоде:

```
# функция цены замены символа на символ
действительное_число цена(символ1, символ2)
  ...
конец цена;

# начальный (нулевой) символ
ε := ...;

действительное_число РасстояниеВагнераФишера(строка1, строка2)
  дс1 := длинна(строка1);
  дс2 := длинна(строка2);

  # d - матрица размером дс1*дс2
  действительное _число d[0..дс1, 0..дс2];
  # вспомогательные переменные
  целое_число i, j;

  d[0,0] = 0;
  для i от 1 до дс1
     d[i,0]:= Σ_{k=1}^{i} цена(строка1[k],ε);
  для j от 1 до дс2
     d[0,j]:= Σ_{k=1}^{j} цена(ε,строка2[k]);
  для i от 1 до дс1
    для j от 1 до дс2
      # строки есть массивы символов
      d[i, j] := минимум(
                # стирание
                d[i-1, j] + цена(строка1[i],ε),
                # вставка
```

```
                d[i, j-1] + цена(ε,строка2[j]),
                # замена
                d[i-1, j-1] +
                    цена(строка1[i],строка2[j])
                  );
  возвратить d[дс1, дс2];
конец РасстояниеВагнераФишера;
```

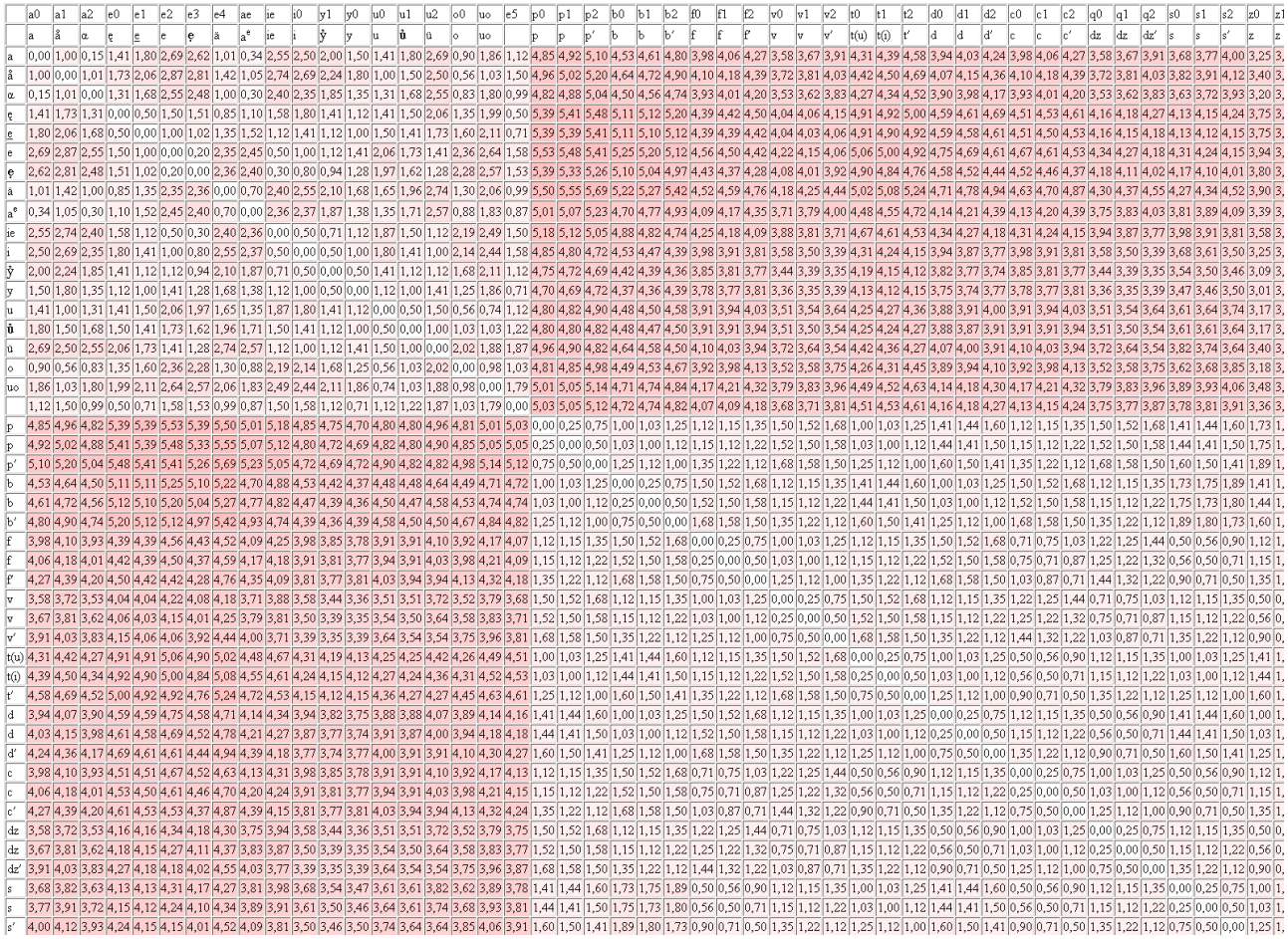

Рис. 4. Фрагмент таблицы расстояний между фонемами

Наиболее корректным способом, как задать расстояние между всеми парами звуков латышских и латгальских наречий, мы посчитали создание такой системы координат, в которой каждая фонема обладала бы уникальными координатами, более-менее соответствующими и акустическим, и физиологическим, и общественным представлениям о данном звуке, а также задание расстояния на ней. После долгих раздумий и консультаций с коллегами нами было создано шестимерное фонетическое пространство с осями звучности, мягкости, места, лабиализации, открытости и дрожания, на котором мы ввели евклидово расстояние. Мы ввели координаты всех фонем наречий и рассчитали таблицу расстояний *(рис. 4)* при разных вариантах веса осей, из которых выбрали наилучший (таблица которого, между прочим, вполне соответствовала нашим ожиданиям). Фактически, данная система координат является отдельным научным результатом, требующим более подробного описания, но не в этом докладе (основным новаторством является то, что мы поделили традиционное измерение места на два измерения: место (альвеолярность, дентальность и лабиальность) и мягкость (палатализированность и палатальность в одном направлении, а велярность, увулярность и фарингальность – в другом), так как иначе система координат получалась нелогичной).

Выбрать начальную (нулевую) точку было непросто. Мы рассматривали три варианта: просто нулевые координаты, наиболее нейтральное физиологическое состояние и средняя точка используемых нами осевых интервалов. Наиболее удачным оказался последний, поэтому мы остановились на нём.

Далее мы написали программу на языке ПЕРЛ, рассчитывающую среднее расстояние Вагнера-Фишера между словами пар наречий, цену замены фонемы на фонему определив как комбинированное евклидово расстояние между соответствующими точками в нашем шестимерном пространстве фонем и нескольких дополнительных измерениях: долготы, интонации, ударности и слоговости. На оси интонаций мы расстояние задали двоично (совпадает – 0, не совпадает – 1), так как интонаций пять, но трудно оценить, какая из них к какой ближе.

	Шкылбаны (Škylbāni)	Цыбла (Cybla)	Зимер (Zīmeris)	Аташина (Atašine)	Алшванга (Alšvang)	Уолайнэ (Olaine)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rīte)	Деры (Ģeri)
Шкылбаны	0,000	0,488	0,432	0,503	1,174	0,895	0,390	1,210	0,850	0,844	0,434	1,025	0,674
Цыбла		0,000	0,667	0,824	1,407	1,121	0,394	1,493	0,964	1,075	0,708	1,369	0,850
Зимер			0,000	0,518	1,161	0,734	0,518	1,232	0,714	0,688	0,397	1,010	0,443
Аташина				0,000	1,064	0,733	0,758	1,101	0,678	0,685	0,464	0,922	0,690
Алшванга					0,000	0,722	1,249	0,338	0,749	0,822	1,069	0,376	1,147
Уолайнэ						0,000	0,954	0,870	0,322	0,250	0,796	0,679	0,829
Аулея							0,000	1,278	0,797	0,862	0,567	1,191	0,635
Зуры								0,000	0,853	0,927	1,128	0,496	1,233
Руцово									0,000	0,462	0,701	0,789	0,794
Куоцены										0,000	0,768	0,566	0,759
Лейвон											0,000	0,980	0,382
Ритэ												0,000	1,056
Деры													0,000

Таб. 4. Расстояния между говорами:фонетические вопросы, расстояние Вагнера-Фишера

	Шкылбаны (Škylbāni)	Цыбла (Cybla)	Зимер (Zīmeris)	Аташина (Atašine)	Алшванга (Alšvang)	Уолайнэ (Olaine)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rīte)	Деры (Ģeri)
Шкылбаны	0,000	0,184	0,343	0,299	0,941	0,725	0,332	1,022	0,856	0,673	0,518	0,757	0,549
Цыбла		0,000	0,470	0,459	0,960	0,736	0,197	1,058	0,866	0,700	0,400	0,842	0,381
Зимер			0,000	0,218	0,792	0,521	0,331	0,965	0,669	0,481	0,390	0,747	0,367
Аташина				0,000	0,870	0,677	0,303	1,038	0,807	0,550	0,388	0,769	0,419
Алшванга					0,000	0,546	0,879	0,266	0,639	0,597	0,923	0,595	0,880
Уолайнэ						0,000	0,654	0,765	0,181	0,262	0,637	0,551	0,618
Аулея							0,000	1,014	0,784	0,552	0,218	0,832	0,216
Зуры								0,000	0,859	0,796	1,043	0,513	1,017
Руцово									0,000	0,311	0,789	0,577	0,760
Куоцены										0,000	0,531	0,450	0,545
Лейвон											0,000	0,849	0,126
Ритэ												0,000	0,829
Деры													0,000

Таб. 5. Расстояния между говорами: морфологические вопросы, расстояние Вагнера-Фишера

	Шкылбаны (Škylbāni)	Цыбла (Cybla)	Зимер (Zīmeris)	Аташина (Atašine)	Алшванга (Alšvang)	Уолайнэ (Olaine)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rīte)	Деры (Ģeri)
Шкылбаны	0,000	0,296	1,263	0,657	1,839	2,004	0,420	1,830	1,594	1,839	0,866	1,365	1,847
Цыбла		0,000	1,305	0,768	1,851	2,018	0,223	1,857	1,595	1,877	0,699	1,419	1,878
Зимер			0,000	1,038	1,406	1,471	1,341	1,434	1,468	1,438	1,026	1,084	1,489
Аташина				0,000	1,689	1,840	0,921	1,684	1,441	1,739	1,062	1,279	1,690
Алшванга					0,000	1,437	1,847	0,190	1,276	1,310	1,559	1,490	1,33*9
Уолайнэ						0,000	2,011	1,459	1,259	0,550	1,699	1,800	0,474
Аулея							0,000	1,852	1,656	1,872	0,799	1,423	1,857
Зуры								0,000	1,287	1,384	1,578	1,472	1,398
Руцово									0,000	1,298	1,654	1,670	1,168
Куоцены										0,000	1,606	1,740	0,238
Лейвон											0,000	0,923	1,616
Ритэ												0,000	1,758
Деры													0,000

Таб. 6. Расстояния между говорами: лексические вопросы, расстояние Вагнера-Фишера

Рис. 5. Результаты категоризации: фонетические вопросы, расстояние Вагнера-Фишера

Рис. 6. Результаты категоризации: морфологические вопросы, расстояние Вагнера-Фишера

Выводы

Из таблиц расстояний и схем результатов категоризации видно, что при применении расстояния Вагнера-Фишера не появились существенные различия с результатами по расстоянию Левенштейна, хотя кое-какие нюансы, интуитивно кажущиюся обоснованными, есть. Так, например, по Вагнеру-Фишеру два селонских говора разных губерний (Лейвон – Латгалия, Ритэ – Курляндия) образуют пару и по фонетическим, и по морфологическим вопросам, но по лексическим Лейвон оказывается чуть-чуть ближе к другим говорам Латгалии, а по Левенштейну они образуют пару во всех трёх случаях.

Рис. 7. Результаты категоризации: лексические вопросы, расстояние Вагнера-Фишера

Неожиданностью сначала может показаться морфологическая близость Аташини и Зимера, выявляющаяся только по Вагнеру-Фишеру – первое является глубокоселонским наречием Латгалии, а второе – глубоколатгальским наречием Лифляндии, причём размещены они весьма отдалённо; но если подумать, то это находит вполне логичное объяснение: оба наречия длительно находятся под влиянием среднелатышских наречий, и именно на морфологию оно наложило наибольший отпечаток. Есть и настолько яркие примеры разграничения групп вопросов, что они появляются и по Вагнеру-Фишеру, и по Левенштейну, например, Деры являются Лифляндским ливонизированным наречием, поэтому не удивляет, что фонетически оно оказывается ближе к Курляндским ливонизированным наречиям (Зуры, Алшванга), а лексически – к своим соседям – Лифляндскому среднелатышскому говору (Куоцены). О меньшей нюансированности при применении расстояния Левенштейна свидетельствует и факт совпадения результатов категоризации по морфологическим и лексическим вопросам.

Поэтому можем сделать вывод, что в подобных измерениях имеет смысл пользоваться расстоянием Вагнера-Фишера вместо расстояния Левенштейна в тех случаях, когда нас интересуют языковые нюансы. В тех

же случаях, когда нужна грубая, но быстродействующая категоризация, оправдано применение расстояния Левенштейна.

В будущем было бы интересно ввести полный набор материалов атласа и провести категоризацию всех 490 наречий. А в перспективе планируем подключить и наречия других языков: сначало более родственных – литовского, славянских, а потом, возможно, и более отдалённых – других индоевропейских, финно-угорских и т.д. Т.е., сама методика может быть полезна не только для диалектологических исследований какого-то отдельного языка, но и для изучения связей разных языков.

Список литературы

1. Latviešu valodas dialektoloģijas atlanta materiālu vākšanas programma. Rīga: Latvijas PSR ZA izdevniecība, 1954.

2. Kessler B. Computational dialectology in Irish Gaelic // Proc. of the European ACL. Dublin: 1995. P. 60-66.

3. Nerbonne J., Heeringa W., van den Hout E., van der Kooi P.,Otten S., van de Vis S. W. Phonetic Distance between Dutch Dialects // CLIN VI, Papers from the sixth CLIN meeting. Antwerp: University of Antwerp, Center for Dutch Language and Speech. P. 185-202.

4. Берзинь А.У. Сравнение балтийских языков методом n-грамм // Труды международной коференции „Корпусная лингвистика - 2004“. СПб: Издательство С.-Петербургского университета, 2004. С. 65-71.