

## 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, расчёта 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 объёма 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 рекомендованным 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 := длина(строка1);
дс2 := длина(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
    );

  вернуть d[дс1, дс2];
конец РасстояниеЛевенштейна;

```

Алгоритм иерархической категоризации проще описать словами:

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

2. Создадим  $n$  начальных одноэлементных категорий – для каждого объекта по одной, и определим расстояния между ними как расстояния между составляющими их объектами.

3. Найдём две ближайшие (с наименьшим расстоянием) категории и объединим 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:

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

(Где  $A$  – новосозданная категория, а  $B$  – другая категория, расстояние до которой рассчитываем. Кардинальное число категории есть её мощность, т.е., количество объектов 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 „čārtuûkša“, у нас хранится в виде „č02010a04311r02010m02000y02D01u02201k02000š02000a02001“.

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



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
Аташина			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

a0	a1	a2	e0	e1	e2	e3	e4	ae	ae	je	jo	y1	y0	uo	u1	u2	uo	uo	e5	p0	p1	p2	b0	b1	b2	o0	o1	o2	o3	o4	o5	o6	o7	o8	o9	c1	c2	q0	q1	q2	q3	q4	q5	q6	q7	q8	q9	s1	s2	s3	s4	s5	s6	s7	s8	s9	t1	t2	t3	t4	t5	t6	t7	t8	t9	z1	z2	z3	z4	z5	z6	z7	z8	z9																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										
0,00	1,00	1,00	0,15	1,41	1,80	2,69	2,62	1,01	0,34	2,55	2,50	2,00	1,50	1,41	1,80	2,69	0,90	1,86	1,12	4,85	4,92	5,10	4,53	4,61	4,80	3,98	4,06	4,27	3,58	3,67	3,91	4,31	4,39	4,58	3,94	4,03	4,24	3,98	4,06	4,27	3,58	3,67	3,91	3,68	3,77	4,00	3,25	3,3	3,5	3,6	3,8	3,9	4,1	4,2	4,4	4,5	4,7	4,8	5,0	5,1	5,3	5,4	5,6	5,7	5,9	6,0	6,2	6,3	6,5	6,6	6,8	6,9	7,1	7,2	7,4	7,5	7,7	7,8	8,0	8,1	8,3	8,4	8,6	8,7	8,9	9,1	9,2	9,4	9,5	9,7	9,8	10,0	10,1	10,3	10,4	10,6	10,7	10,9	11,0	11,2	11,4	11,5	11,7	11,8	12,0	12,1	12,3	12,4	12,6	12,7	12,9	13,0	13,2	13,3	13,5	13,6	13,8	13,9	14,1	14,2	14,4	14,5	14,7	14,8	15,0	15,1	15,3	15,4	15,6	15,7	15,9	16,0	16,2	16,3	16,5	16,6	16,8	16,9	17,1	17,2	17,4	17,5	17,7	17,8	18,0	18,1	18,3	18,4	18,6	18,7	18,9	19,0	19,2	19,3	19,5	19,6	19,8	19,9	20,1	20,2	20,4	20,5	20,7	20,8	21,0	21,1	21,3	21,4	21,6	21,7	21,9	22,0	22,2	22,3	22,5	22,6	22,8	22,9	23,1	23,2	23,4	23,5	23,7	23,8	24,0	24,1	24,3	24,4	24,6	24,7	24,9	25,0	25,2	25,3	25,5	25,6	25,8	25,9	26,1	26,2	26,4	26,5	26,7	26,8	27,0	27,1	27,3	27,4	27,6	27,7	27,9	28,0	28,2	28,3	28,5	28,6	28,8	28,9	29,1	29,2	29,4	29,5	29,7	29,8	30,0	30,1	30,3	30,4	30,6	30,7	30,9	31,0	31,2	31,3	31,5	31,6	31,8	31,9	32,1	32,2	32,4	32,5	32,7	32,8	33,0	33,1	33,3	33,4	33,6	33,7	33,9	34,0	34,2	34,3	34,5	34,6	34,8	34,9	35,1	35,2	35,4	35,5	35,7	35,8	36,0	36,1	36,3	36,4	36,6	36,7	36,9	37,0	37,2	37,3	37,5	37,6	37,8	37,9	38,1	38,2	38,4	38,5	38,7	38,8	39,0	39,1	39,3	39,4	39,6	39,7	39,9	40,0	40,2	40,3	40,5	40,6	40,8	40,9	41,1	41,2	41,4	41,5	41,7	41,8	42,0	42,1	42,3	42,4	42,6	42,7	42,9	43,0	43,2	43,3	43,5	43,6	43,8	43,9	44,1	44,2	44,4	44,5	44,7	44,8	45,0	45,1	45,3	45,4	45,6	45,7	45,9	46,0	46,2	46,3	46,5	46,6	46,8	46,9	47,1	47,2	47,4	47,5	47,7	47,8	48,0	48,1	48,3	48,4	48,6	48,7	48,9	49,0	49,2	49,3	49,5	49,6	49,8	49,9	50,1	50,2	50,4	50,5	50,7	50,8	51,0	51,1	51,3	51,4	51,6	51,7	51,9	52,0	52,2	52,3	52,5	52,6	52,8	52,9	53,1	53,2	53,4	53,5	53,7	53,8	54,0	54,1	54,3	54,4	54,6	54,7	54,9	55,0	55,2	55,3	55,5	55,6	55,8	55,9	56,1	56,2	56,4	56,5	56,7	56,8	57,0	57,1	57,3	57,4	57,6	57,7	57,9	58,0	58,2	58,3	58,5	58,6	58,8	58,9	59,1	59,2	59,4	59,5	59,7	59,8	60,0	60,1	60,3	60,4	60,6	60,7	60,9	61,0	61,2	61,3	61,5	61,6	61,8	61,9	62,1	62,2	62,4	62,5	62,7	62,8	63,0	63,1	63,3	63,4	63,6	63,7	63,9	64,0	64,2	64,3	64,5	64,6	64,8	64,9	65,1	65,2	65,4	65,5	65,7	65,8	66,0	66,1	66,3	66,4	66,6	66,7	66,9	67,0	67,2	67,3	67,5	67,6	67,8	67,9	68,1	68,2	68,4	68,5	68,7	68,8	69,0	69,1	69,3	69,4	69,6	69,7	69,9	70,0	70,2	70,3	70,5	70,6	70,8	70,9	71,1	71,2	71,4	71,5	71,7	71,8	72,0	72,1	72,3	72,4	72,6	72,7	72,9	73,0	73,2	73,3	73,5	73,6	73,8	73,9	74,1	74,2	74,4	74,5	74,7	74,8	75,0	75,1	75,3	75,4	75,6	75,7	75,9	76,0	76,2	76,3	76,5	76,6	76,8	76,9	77,1	77,2	77,4	77,5	77,7	77,8	78,0	78,1	78,3	78,4	78,6	78,7	78,9	79,0	79,2	79,3	79,5	79,6	79,8	79,9	80,1	80,2	80,4	80,5	80,7	80,8	81,0	81,1	81,3	81,4	81,6	81,7	81,9	82,0	82,2	82,3	82,5	82,6	82,8	82,9	83,1	83,2	83,4	83,5	83,7	83,8	84,0	84,1	84,3	84,4	84,6	84,7	84,9	85,0	85,2	85,3	85,5	85,6	85,8	85,9	86,1	86,2	86,4	86,5	86,7	86,8	87,0	87,1	87,3	87,4	87,6	87,7	87,9	88,0	88,2	88,3	88,5	88,6	88,8	88,9	89,1	89,2	89,4	89,5	89,7	89,8	90,0	90,1	90,3	90,4	90,6	90,7	90,9	91,0	91,2	91,3	91,5	91,6	91,8	91,9	92,1	92,2	92,4	92,5	92,7	92,8	93,0	93,1	93,3	93,4	93,6	93,7	93,9	94,0	94,2	94,3	94,5	94,6	94,8	94,9	95,1	95,2	95,4	95,5	95,7	95,8	96,0	96,1	96,3	96,4	96,6	96,7	96,9	97,0	97,2	97,3	97,5	97,6	97,8	97,9	98,1	98,2	98,4	98,5	98,7	98,8	99,0	99,1	99,3	99,4	99,6	99,7	99,9	100,0	100,2	100,3	100,5	100,6	100,8	100,9	101,1	101,2	101,4	101,5	101,7	101,8	102,0	102,1	102,3	102,4	102,6	102,7	102,9	103,0	103,2	103,3	103,5	103,6	103,8	103,9	104,1	104,2	104,4	104,5	104,7	104,8	105,0	105,1	105,3	105,4	105,6	105,7	105,9	106,0	106,2	106,3	106,5	106,6	106,8	106,9	107,1	107,2	107,4	107,5	107,7	107,8	108,0	108,1	108,3	108,4	108,6	108,7	108,9	109,0	109,2	109,3	109,5	109,6	109,8	109,9	110,1	110,2	110,4	110,5	110,7	110,8	111,0	111,1	111,3	111,4	111,6	111,7	111,9	112,0	112,2	112,3	112,5	112,6	112,8	112,9	113,1	113,2	113,4	113,5	113,7	113,8	114,0	114,1	114,3	114,4	114,6	114,7	114,9	115,0	115,2	115,3	115,5	115,6	115,8	115,9	116,1	116,2	116,4	116,5	116,7	116,8	117,0	117,1	117,3	117,4	117,6	117,7	117,9	118,0	118,2	118,3	118,5	118,6	118,8	118,9	119,1	119,2	119,4	119,5	119,7	119,8	120,0	120,1	120,3	120,4	120,6	120,7	120,9	121,0	121,2	121,3	121,5	121,6	121,8	121,9	122,1	122,2	122,4	122,5	122,7	122,8	123,0	123,1	123,3	123,4	123,6	123,7	123,9	124,0	124,2	124,3	124,5	124,6	124,8	124,9	125,1	125,2	125,4	125,5	125,7	125,8	126,0	126,1	126,3	126,4	126,6	126,7	126,9	127,0	127,2	127,3	127,5	127,6	127,8	127,9	128,1	128,2	128,4	128,5	128,7	128,8	129,0	129,1	129,3	129,4	129,6	129,7	129,9	130,0	130,2	130,3	130,5	130,6	130,8	130,9	131,1	131,2	131,4	131,5	131,7	131,8	132,0	132,1	132,3	132,4	132,6	132,7	132,9	133,0	133,2	133,3	133,5	133,6	133,8	133,9	134,1	134,2	134,4	134,5	134,7	134,8	135,0	135,1	135,3	135,4	135,6	135,7	135,9	136,0	136,2	136,3	136,5	136,6	136,8	136,9	137,1	137,2	137,4	137,5	137,7	137,8	138,0	138,1	138,3	138,4	138,6	138,7	138,9	139,0	139,2	139,3	139,5	139,6	139,8	139,9	140,1	140,2	140,4	140,5	140,7	140,8	141,0	141,1	141,3	141,4	141,6	141,7	141,9	142,0	142,2	142,3	142,5	142,6	142,8	142,9	143,1	143,2	143,4	143,5	143,7	143,8	144,0	144,1	144,3	144,4	144,6	144,7	144,9	145,0	145,2	145,3	145,5	145,6	145,8	145,9	146,1	146,2	146,4	146,5	146,7	146,8	147,0	147,1	147,3	147,4	147,6	147,7	147,9	148,0	148,2	148,3	148,5	148,6	148,8	148,9	149,1	149,2	149,4	149,5	149,7	149,8	150,0	150,1	150,3	150,4	150,6	150,7	150,9	151,0	151,2	151,3	151,5	151,6	151,8	151,9	152,1	152,2	152,4	152,5	152,7	152,8	153,0	153,1	153,3	153,4	153,6	153,7	153,9	154,0	154,2	154,3	154,5	154,6	154,8	154,9	155,1	155,2	155,4	155,5	155,7	155,8	156,0	156,1	156,3	156,4	156,6	156,7	156,9	157,0	157,2	157,3	157,5	157,6	157,8	157,9	158,1	158,2	158,4	158,5	158,7	158,8	159,0	159,1	159,3	159,4	159,6	159,7	159,9	160,0	160,2	160,3	160,5	160,6	160,8	160,9	161,1	161,2	161,4	161,5	161,7	161,8	162,0	162,1	162,3	162,4	162,6	162,7	162,9	163,0	163,2	163,3	163,5	163,6	163,8	163,9	164,1	164,2	164,4	164,5	164,7	164,8	165,0	165,1	165,3	165,4	165,6	165,7	165,9	166,0	166,2	166,3	166,5	166,6	166,8	166,9	167,1	167,2	167,4	167,5	167,7	167,8	168,0	168,1	168,3	168,4	168,6	168,7	168,9	169,0	169,2	169,3	169,5	169,6	169,8	169,9	170,1	170,2	170,4	170,5	170,7	170,8	171,0	171,1	171,3	171,4	171,6	171,7	171,9	172,0	172,2	172,3	172,5	172,6	172,8	172,9	173,1	173,2	173,4	173,5	173,7	173,8	174,0	174,1	174,3	174,4	174,6	174,7	174,9	175,0	175,2	175,3	175,5	175,6	175,8	175,9	176,1	176,2	176,4	176,5	176,7	176,8	177,0	177,1	177,3	177,4	177,6	177,7	177,9	178,0	178,2	178,3	178,5	178,6	178,8	178,9	179,1	179,2	179,4	179,5	179,7	179,8	180,0	180,1	180,3	180,4	180,6	180,7	180,9	181,0	181,2	181,3	181,5	181,6	181,8	181,9	182,1	182,2	182,4	182,5	182,7	182,8	183,0	183,1	183,3	183,4	183,6	183,7	183,9	184,0

```

# 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

```

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 заданное расстояние 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 (палатализованность 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 (Zimeris)	Аташина (Atašine)	Алшванга (Alšvang)	Уолайнэ (Olaīne)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rīte)	Деры (Deri)
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 (Zimeris)	Аташина (Atašine)	Алшванга (Alšvang)	Уолайнэ (Olaīne)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivõns)	Ритэ (Rīte)	Деры (Deri)
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 (Arašīne)	Алшванга (Alšvang)	Уолайнэ (Olaine)	Аулея (Auleja)	Зуры (Zūrs)	Руцово (Rucava)	Куоцены (Kocēni)	Лейвон (Leivōns)	Ритэ (Rīte)	Деры (Deri)
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 кажущиеся 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 Аташины и Зимера, выявляющаяся только по Вагнеру-Фишеру – первое является глубокоселонским наречием Латгалии, а второе – глубоколлатгальским 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: сначала 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 Kooij 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 конференции „Корпусная лингвистика - 2004“*. СПб: Издательство С.-Петербургского университета, 2004. С. 65-71.