Применение распознавателей фонем для автоматического определения уровня близости языков

Ansis Ataols Bērziņš, RA IF
ansis@latnet.lv

Москва, 4 июня 2016 г.

Причины

„Язык - это диалект + армия и флот“

טאַ שפּראַך איז אַ דיאַלעקט מיט אַן אַרמיי און פֿלאָט

[а шпрах из диалект мит арми ун флот]

Мейер Макс Вейнрейх, 1945 г.

Я разрабатываю методы противостояния
этой устоявшейся практике
путём численных оценок.

Возможности

Rakstīts vārds *Transcripciones de habla*:
- ortogrāfiskā pierakstā *transcripción ortográfica*
(izloksnēm neiespējami *para muchas lenguajes no es posible)*
- fonētiskā pierakstā *transcripción fonética*

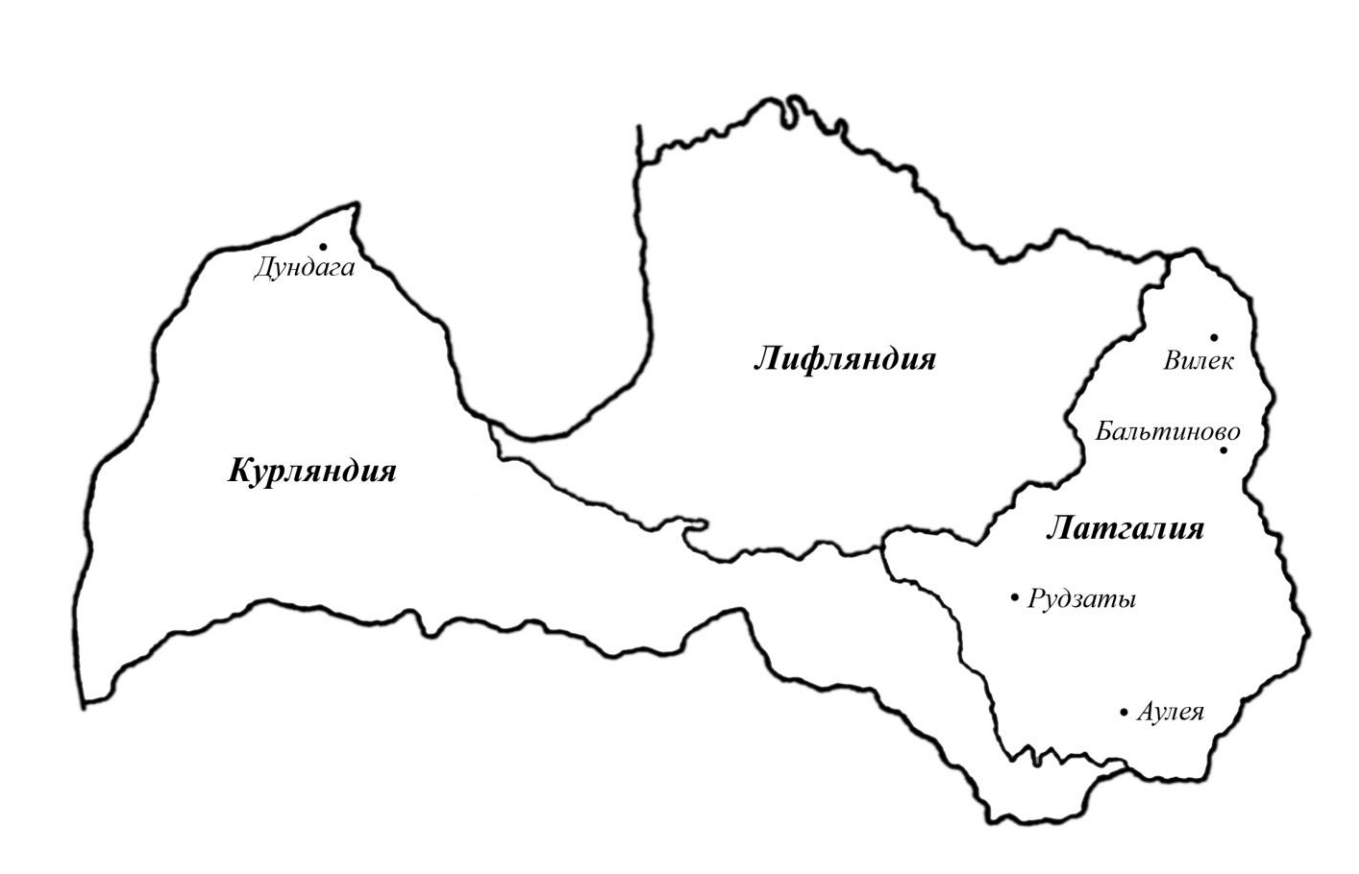
Skanošs vārds *Grabaciones de habla*:
- lasīta runa *habla leída*
*(*izloksnēm tikpat kā neiespējami *para mayoria de las lenguajes casi imposible)*
- spontāna runa *habla espontánea*

Darbs ar runas ierakstiem
Trabaja con grabaciones de habla

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

Таб.1. Характеристика набора фонограмм, используемого в эксперименте.

Darbs ar runas ierakstiem
Trabaja con grabaciones de habla



Таб.1. Характеристика набора фонограмм, используемого в эксперименте.

Fonēmu atpazinējs
Reconocedor de fonemas

Ir iespējams izmantot fonēmu atpazinēju, un pēc tam – fonētiskā pieraksta metodes, piemēram, n-gramu metodi.

Es posible utilizar un reconocedor de fonemas, y luego – métodos para transcripciones fonéticos (por ejemplo, el método de n-gramas).

PhnRec – модели чешского, английского, русского и венгерского языков.
Sphinx 4 – модель английского.

Šķirošana pēc n-grammām
Categorización por n-gramas

1994.g. Viljams Kanvars un Džonis Trenkls piedāvā izmantot tekstu šķirošanā pēc tēmas (piem., e-pasta vēstuļu automātiska pirmapstrāde)

Pēc Cipfa likuma, dabiskās valodas vārdu kārtas numurs to biežuma sarakstā ir apgriezti proporcionāls to sastapšanas valodā biežumam.

Por la ley de Zipf, la posición de las palabras de un lenguaje natural en la lista de frecuencias es inversamente proporcional a la frecuencia de estas palabras en el lenguaje.

Šķirošana pēc n-grammām
Categorización por n-gramas

Varam pielietot arī valodu salīdzināšanai.
Podemos usar lo tambien para la comparación de lenguas.

G. van Norda PERLa programmas TextCat modifikācija.

~1000 vārdus gari izlokšņu teksti 2-baitīgajā fonētiskajā transkripcijā.
400 fonētiskā pieraksta n-grammas (no 2 līdz 5).
textos con longitud de ~1000 palabras en transcripción fonética, donde cada fonema usa 2 bytes.
400 n-gramas en cada lista de frecuencias (de 2 a 5).

Fonēmu atpazinējs
Reconocedor de fonemas

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	35104	50996	31718	41810
Baļtinova	35104	0	52847	31787	30988
Dundag	50996	52847	0	49741	48986
Rudzātys	31718	31787	49741	0	37232
Vileks	41810	30988	48986	37232	0

Таб.1. Характеристика набора фонограмм, используемого в эксперименте.

Fonēmu atpazinējs
Reconocedor de fonemas

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	33073	46448	30638	34175
Baļtinova	33073	0	47752	30989	29439
Dundag	46448	47752	0	43906	45767
Rudzātys	30638	30989	43906	0	34626
Vileks	34175	29439	45767	34626	0

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

Fonēmu atpazinējs
Reconocedor de fonemas

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	36589	50677	39510	42803
Baļtinova	36589	0	47708	33217	30514
Dundag	50677	47708	0	47088	47711
Rudzātys	39510	33217	47088	0	33512
Vileks	42803	30514	47711	33512	0

Tabula 8.4. Attālums starp izloksnēm, atpazinējs: PhnRec, modeļa valoda: ungāru.

Fonēmu atpazinējs
Reconocedor de fonemas

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	44098	52208	41165	46056
Baļtinova	44098	0	51315	37773	30283
Dundag	52208	51315	0	50077	55491
Rudzātys	41165	37773	50077	0	36577
Vileks	46056	30283	55491	36577	0

Tabula 8.5. Attālums starp izloksnēm, atpazinējs: PhnRec, modeļa valoda: angļu.

Fonēmu atpazinējs
Reconocedor de fonemas

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	36751	51130	38058	35900
Baļtinova	36751	0	47878	37603	30205
Dundag	51130	47878	0	49830	51087
Rudzātys	38058	37603	49830	0	34568
Vileks	35900	30205	51087	34568	0

Tabula 8.6. Attālums starp izloksnēm, atpazinējs: Sphinx, modeļa valoda: angļu.

Fonēmu atpazinējs
Reconocedor de fonemas

Результаты категоризации: PhnRec, модели чешского и русского.

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

Выводы

1.Распознаватели фонем **применимы** к речевым фонокорпусам **для автоматической оценки уровня близости языков**.
2. Язык обучения модели распознавателя влияет на результаты, но не настолько, чтобы исключить применение.
3. Архитектура фонемного распознавателя на результаты определения расстояния существенно не влияет.

Попутно задана метрика, при чём не нуждающаяся в ручной подготовке данных.

Paldies par uzmanību!

Gracias por su atención!

Спасибо за внимание!