Usage of i-Vectors for Automated Determination of a Similarity Level Between Languages

Применение i-векторов для автоматизированного определения уровня близости языков

Ansis Ataols Bērziņš
ansis@latnet.lv

Moscow, December 6th, 2019

Introduction (1)

The aim: various statistical methods for assessing proximity of natural idioms.

We have 8 methods developed during writing of our thesis. This presentation is about one of them.

Introduction (2)

„A language is a dialect with an army and navy“

טאַ שפּראַך איז אַ דיאַלעקט מיט אַן אַרמיי און פֿלאָ

[a šprah iz a dialekt mit an armī un flot]

Meijer Max Weinreich, 1945

What to do against it?

Introduction (3)

Texts:

- orthographic transcription (most dialects do not have; characterizes tradition instead of language)
- **phonetic transcription**

Speech recordings:

- read speech (for most dialects data is not available)
- **spontaneous speech**

Introduction (4)

i-Vectors are state of the art speech identification technology:
- SID (speaker identification)
- LID (language identification)

“The main idea of the iVector model in acoustic … recognition is to represent each utterance dependent GMM with a low-dimensional latent variable and use the low-dimensional representation of the utterance as a feature vector to the following … classifier.” [Soufifar, 2014]

Introduction (5)

The main idea of our experiment(s):
- to use full length recordings of spontaneous speech of different speakers of the same language to create i-Vectors
- to calculate a kind of distance between these vectors created

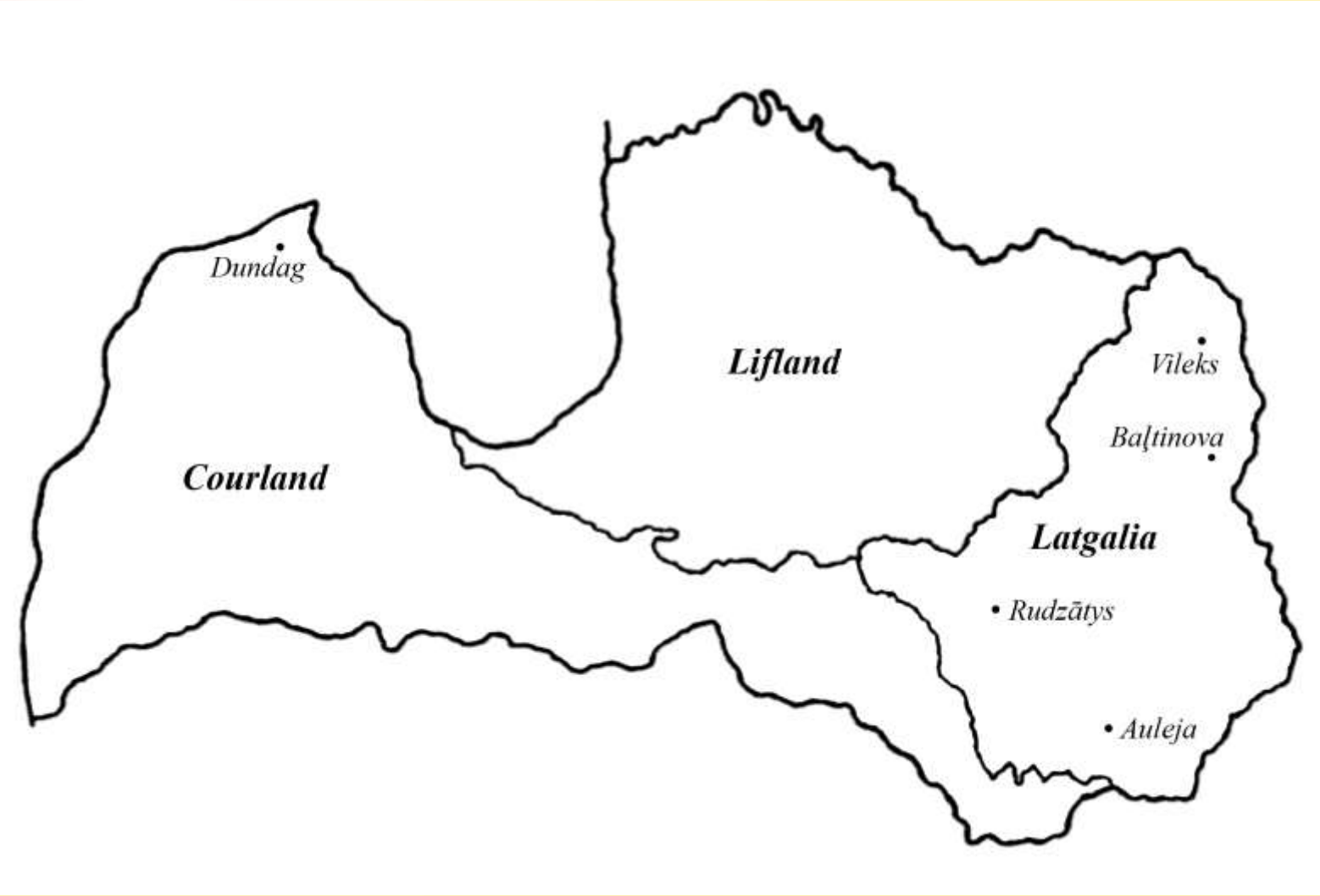
Will this distance characterise how far or close the languages are?

Data (1)

Latvian dialects were more accessible to us, we decided first to be based on them.

In year 2008 we have been collected spontaneous speech recordings of five Latvian dialects, four of them – Latgalian, and one – Couronian.

Data (2)


Dundag
Lifland
Vileks
Baļtinova
Courland
Latgalia
Rudzātys
Auleja

Data (3)

All recordings were uniformed, recorded with the same type of hardware (a dynamic one-way microphone fixed on heads of speakers was used), an external noise was minimized as far as possible. All entries were manually cleared – i.e., all other voices and sounds were cut out, leaving only the speech of the main speaker. Recordings' technical quality was 44.1 kHz / 16 bit.

Data (4)

Dialect	*Minutes collected*	*Number of informants*	*Including male*	*Including female*
Auleja	95	14	8	6
Baļtinova	140	23	9	14
Dundaga	161	17	4	13
Rudzātys	246	28	11	17
Vileks	238	30	11	19

Data (5)

All informants were asked to tell their **life stories**: about parents, grandparents, brothers, sisters, children, other family members, school, work, weddings, farm, military service, etc.

It means the **lexicon** used by the informants was traditional and homogeneous.

Experiment – SID (1)

1) preprocessing – VAD (*Voice Activity Detection*)
2) SID i-Vectors calculated by script, included in *Voice Biometry Standart* package

Experiment – SID (2)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	86	137	99	94
Baļtinova	86	0	129	109	47
Dundag	137	129	0	78	122
Rudzātys	99	109	78	0	116
Vileks	95	47	122	116	0

*Angles in degrees from **cosine similarity** on SID i-vectors of our recordings (values rounded).*

Experiment – SID (3)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	8,04	11,42	8,71	9,00
Baļtinova	8,04	0	9,78	7,56	6,56
Dundag	11,42	9,78	0	9,27	9,13
Rudzātys	8,71	7,56	9,27	0	7,24
Vileks	9,00	6,56	9,13	7,24	0

Euclidean *metric between SID i-vectors of our recordings (values rounded).*

Experiment – SID (4)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	0,480	0,663	0,515	0,542
Baļtinova	0,480	0	0,582	0,458	0,405
Dundag	0,663	0,582	0	0,547	0,549
Rudzātys	0,515	0,458	0,547	0	0,443
Vileks	0,542	0,505	0,549	0,443	0

Standardised Euclidean *metric between SID i-vectors of our recordings (values rounded).*

Experiment – SID (5)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	1,20	1,39	1,24	1,36
Baļtinova	1,20	0	1,35	0,95	0,87
Dundag	1,39	1,35	0	1,10	1,04
Rudzātys	1,24	0,95	1,10	0	0,94
Vileks	1,36	0,87	1,04	0,94	0

***Jordan** metric between SID i-vectors of our recordings (values rounded).*

Experiment – SID (6)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	155	223	170	173
Baļtinova	155	0	188	149	127
Dundag	223	188	0	182	178
Rudzātys	170	149	182	0	142
Vileks	173	127	178	142	0

City block *or* ***L1*** *metric between SID i-vectors of our recordings (values rounded).*

Experiment – LID (1)

1) preprocessing – VAD (*Voice Activity Detection*)
2) LID i-Vectors calculated by Brno Technical university Speech lab

Experiment – LID (2)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	61	93	55	148
Baļtinova	61	0	104	93	120
Dundag	93	104	0	87	92
Rudzātys	55	93	87	0	145
Vileks	148	120	92	145	0

*Angles in degrees from **cosine similarity** on LID i-vectors of our recordings (values rounded).*

Experiment – LID (3)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	6,07	9,59	6,22	6,76
Baļtinova	6,07	0	8,75	6,91	4,71
Dundag	9,59	8,75	0	8,39	8,45
Rudzātys	6,22	5,91	8,39	0	6,14
Vileks	6,76	4,71	8,45	6,14	0

Euclidean *metric between LID i-vectors of our recordings (values rounded).*

Experiment – LID (4)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	0,381	0,586	0,390	0,428
Baļtinova	0,381	0	0,543	0,377	0,303
Dundag	0,586	0,543	0	0,519	0,528
Rudzātys	0,390	0,377	0,519	0	0,394
Vileks	0,428	0,303	0,528	0,394	0

Standardised Euclidean *metric between LID i-vectors of our recordings (values rounded).*

Experiment – LID (5)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	0,75	1,13	0,79	0,87
Baļtinova	0,75	0	1,31	0,75	0,59
Dundag	1,13	1,31	0	1,11	1,06
Rudzātys	0,79	0,75	1,11	0	0,81
Vileks	0,87	0,59	1,06	0,81	0

***Jordan** metric between LID i-vectors of our recordings (values rounded).*

Experiment – LID (6)

	Auleja	Baļtinova	Dundag	Rudzātys	Vileks
Auleja	0	120	188	123	131
Baļtinova	120	0	171	115	91
Dundag	188	171	0	164	167
Rudzātys	123	115	164	0	119
Vileks	131	91	167	119	0

City block *or* ***L1*** *metric between LID i-vectors of our recordings (values rounded).*

Conclusions (1)

i-Vectors (both SID and LID) **are applicable** for tasks of automated determination
of a similarity level between languages.

Conclusions (2)

The cosine similarity and Gordian metrics on the space of i-Vectors **are partially characterizing** the relations between real objects the models are created for (in a level which is not enough to define them as a distance between languages).

Conclusions (3)

The Euclidean metrics and L_1 metrics on the space of i-Vectors **are characterizing** the relations between real objects the models are created for and could be defined as a distance between languages.

Conclusions (4)

The method – i-Vectors based automated determination of a similarity level between languages – **is usable** and could be **recommended for real** production **use** since it is technically implementable and the results are reliable.

Paldies par uzmanību!

ansis@latnet.lv