Perplexity as a measure of language distance and effects of genre on perplexity

Yu An, Dušica Božović, Elena China-Kolehmainen, Aleksiina Läykki

December 13, 2018

Linguistics in the Digital Age

Introduction

- How is perplexity used for measuring distances between languages?
- Relationship between language identification and language distance
- Gamallo et al. tested it on the Bibles
- Would it be different when tested on another domain?

Background literature

- Gamallo 2017:
    - Tested the Bibles on 44 European languages
    - Compared perplexity and ranking-based distance measuring
    - "Close to phonological" encoding: 24 consonants and 10 vowels (normalization)
    - Combining results with data in Ethnologue
- Östling and Tiedemann (2016) on disadvantages:
    - Limited size
    - One translation only for many languages
    - Only New Testament
    - Narrow domain

What were we interested in?

- Would it be different if we would change the genre?

Our hypothesis

- No or small effect of genre

Our hypotheses

- No or small effect of genre:
    - Perplexities on genre 1
    - Perplexities on genre 2
    - Two set of perplexites should correlate to each other.
- Visualize the distances to show the relationships between languages.

Data collection

- Bible corpus and Little Prince corpus
- 20 languages, 18 written in Latin script, 2 in Cyrillic (Russian and Bulgarian; both of them are transliterated by the same script) The languages are:

1. Bulgarian
2. Catalan
3. Croatian
4. Czech
5. English
6. Esperanto
7. Estonian
8. Finnish
9. French
10. German
11. Hungarian
12. Italian
13. Latvian
14. Lithuanian
15. Polish
16. Portuguese
17. Romanian
18. Russian
19. Spanish
20. Turkish

Data preparation

- The Bible corpus is a collection of parallel Bible by Mayer & Cysouw (2014), most of which are Jehovah's witnesses translations (New World versions). For Catalan and Latvian we used alternative versions based on availability.
- Most of the translations of the Little Prince found on the Czech Odaha's fanpage collection site. The Portuguese version was found on the website Livros Digitais (2010) and the Finnish translation was added by scanning the book.

Data preparation

- The Little Prince's French version counts about 82000 characters.
- We took roughly the same amount of characters for the Bible to have corpora of the same size

Data preparation

- Training and testing sets divided:
    - 70% training material
    - 30% test material
    - Bibles split manually based on characters of the French version and rounded to match a full verse. The training set for the Bible ranges from verse 01001001 to verse 01015015 included. The test set for the Bible ranges from verse 01015016 to verse 01022003 included.
    - Little Princes split partially manually: 20 chapters for training and 7 chapters for testing. Partially split automatically based on lines, 70-30%

Inconsistencies in the Bible data

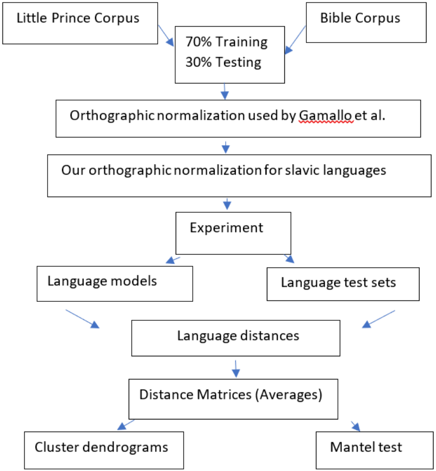
- Catalan and Latvian: since we had no New World versions available we first decided to drop them from the experiment. They were kept after all because their interesting relation with related languages. All the available verses were kept with the result that the data for these two languages is significantly larger (1M vs 80K characters)
- Spanish: different sizes in training and testing.
  - Spanish: 121/430 verses
  - others: 375/174 verses

Examples of perplexity changes for Spanish

Inconsistencies were acknowledged late, no time to start from the beginning. For Spanish the perplexity values would have slightly lowered:

SPA train>	Used	Actual perplexity
bul	209	197
cat	20	18
cz	92	84
deu	48	45
eng	50	45
epo	58	47
...		

Experiment setup



Little Prince Corpus

Bible Corpus

70% Training
30% Testing

Orthographic normalization used by Gamallo et al.

Our orthographic normalization for slavic languages

Experiment

Language models

Language test sets

Language distances

Distance Matrices (Averages)

Cluster dendrograms

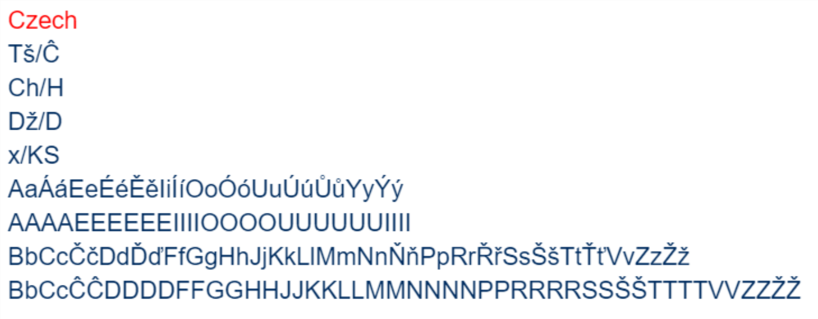
Mantel test

Analysis pipeline describing corpora preprocessing, model training, computation of distances.

Experiment setup

- ”Copy fast, refactor later” (Gardner et al., 2018)
- Part of scripts are from (Pablo Gamallo, 2018):
  - Collect corpora (too clean to cleanse)
  - Split the data set (some are done manually and some with Unix commands)
  - Normalize texts (transcript.perl, tokenizer_ch.perl)
  - Modelling (7grams.perl)
  - Test and calculate perplexity (perplexity_setegrams.perl; model_setegrams.perl)
- Issues with normalization: Slavic languages

Normalization



Czech

Tš/Ĉ

Ch/H

Dž/D

x/KS

AaÁáEeÉéĚěIiÍíOoÓóUuÚúŮůYyÝý

AAAAEEEEEEIIIIOOOOUUUUUUIIII

BbCcČčDdĎďFfGgHhJjKkLlMmNnŇňPpRrŘřSsŠšTtŤťVvZzŽž

BbCcĈĈDDDDFFGGHHJJKKLLMMNNNNPPRRRRSSŠŠTTTTVVZZŽŽ

Figure 1: Normalization for Czech

Example of tokenization

i

n

#

p

r

i

n

c

i

Example of tokenization

p

i

o

#

c

r

e

o

Example of tokenization

#

i

#

c

i

e

l

i

Example of 7-grams

# i n # p r

i n # p r i

i n # p r i n

n # p r i n c

# p r i n c i

p r i n c i p

r i n c i p i

Example of 7-grams

i n c i p i o

n c i p i o #

c i p i o # c

i p i o # c r

p i o # c r e

i o # c r e o

o # c r e o #

Example of 7-grams

# c r e o # i

c r e o # i #

r e o # i # c

e o # i # c i

o # i # c i e

# i # c i e l

i # c i e l i

# c i e l i #

Matrix 0

Perplexities with Gmallo's transcript																				
train–¿	bul	cat	cz	deu	eng	epo	est	fin	fra	hrv	hun	ita	lat	lit	pol	por	ron	rus	spa	tur
bul	5	221	132	210	230	238	222	236	234	100	212	256	139	184	245	257	212	21	303	297
cat	92	3	56	52	31	36	55	69	20	54	44	27	42	57	86	21	34	78	21	58
cz	75	75	6	98	75	86	117	113	69	38	77	102	55	55	55	85	104	51	95	99
deu	103	53	70	4	37	72	76	104	46	76	65	68	64	75	76	73	61	109	62	64
eng	125	37	70	49	4	79	83	100	45	87	56	80	86	79	80	77	75	89	61	67
epo	93	40	51	74	52	4	59	63	42	46	57	75	39	47	74	56	70	79	59	81
est	78	57	61	55	63	62	5	33	55	57	46	87	38	53	81	72	68	83	86	61
fin	83	101	98	70	104	95	36	5	65	103	81	128	46	60	84	132	77	101	133	74
fra	90	14	48	48	27	39	55	62	4	53	46	29	45	53	81	28	31	80	31	56
hrv	42	46	23	65	65	47	68	75	51	6	56	62	38	44	46	54	59	43	63	91
hun	115	75	63	63	74	85	77	90	60	84	6	118	75	90	90	102	119	89	104	55
ita	76	17	45	62	37	33	72	71	27	38	55	4	44	50	67	21	32	78	22	56
lat	119	90	90	82	130	100	91	104	69	95	121	167	3	38	101	130	87	137	150	97
lit	142	97	61	101	113	102	89	95	84	92	102	167	26	7	109	109	140	132	147	101
pol	172	167	68	144	122	155	373	311	150	95	143	208	103	141	6	149	203	115	162	228
por	84	14	48	61	36	34	69	77	26	44	49	22	47	53	65	4	36	83	19	63
ron	95	37	72	84	78	73	87	84	36	69	90	51	40	65	67	51	5	124	65	58
rus	26	150	81	199	155	166	189	182	177	74	144	210	127	138	152	207	203	6	208	267
spa	104	11	46	60	33	31	76	70	22	51	43	23	47	49	59	15	38	80	5	56
tur	227	90	108	114	124	135	120	136	85	130	78	199	66	71	167	188	137	175	155	6

Table 1: Bible perplexities with Gamallo's normalization

Matrix 1

Perplexities with modified transcript																				
train–>	bul	cat	cz	deu	eng	epo	est	fin	fra	hrv	hun	ita	lat	lit	pol	por	ron	rus	spa	tur
bul	5	189	110	195	188	209	257	275	192	84	159	174	138	160	121	200	185	20	209	295
cat	71	3	56	48	30	35	46	56	20	55	41	27	41	45	64	21	31	67	20	56
cz	58	71	6	82	73	92	96	103	67	38	61	107	56	52	31	80	95	42	92	99
deu	66	42	56	4	34	59	62	82	38	62	43	56	54	54	64	60	50	69	48	52
eng	81	31	57	42	4	64	51	69	37	71	42	65	68	56	68	59	60	63	50	55
epo	69	39	51	67	59	4	44	46	42	44	52	84	31	35	49	55	64	69	58	65
est	81	45	51	59	51	62	5	33	48	60	46	85	51	40	55	68	60	70	68	56
fin	129	56	50	60	59	93	29	5	60	107	44	131	95	58	60	138	132	66	78	54
fra	63	13	47	44	26	39	43	49	4	53	42	30	44	44	64	27	28	60	30	54
hrv	33	47	23	59	61	51	67	74	49	6	47	74	38	38	28	53	53	36	61	100
hun	89	60	50	55	53	83	64	75	55	74	6	118	72	61	60	103	109	68	90	56
ita	50	17	44	57	36	33	58	55	27	38	47	4	43	40	48	21	29	57	22	55
lat	124	98	91	104	122	98	113	137	84	97	114	167	3	57	95	125	124	143	137	155
lit	75	59	52	68	66	63	58	69	58	53	60	104	28	7	59	73	79	81	82	81
pol	78	103	35	113	128	140	151	156	113	54	83	165	77	64	6	115	158	54	119	209
por	64	14	48	60	37	35	59	64	26	46	45	24	48	45	55	4	34	71	19	61
ron	56	21	42	65	45	41	41	50	26	40	49	28	41	44	50	32	5	62	37	44
rus	27	147	81	171	162	173	193	205	167	77	126	207	131	131	81	189	180	6	184	277
spa	73	11	46	51	33	31	51	52	22	51	39	24	46	40	51	14	35	64	5	53
tur	199	107	100	100	107	144	63	102	80	141	77	214	136	66	114	151	138	135	136	6

Table 2: Bible perplexities with modified normalization

Matrix 2

Perplexities with modified transcript (from Little Prince Corpus)																				
train ->	cat	ron	ita	lat	spa	est	rus	por	hun	fra	tur	cz	deu	eng	fin	epo	hrv	bul	lit	pol
cat	6	32	22	44	4	49	58	15	39	23	47	44	42	30	49	28	43	62	38	51
ron	30	5	31	47	34	39	63	32	48	29	48	47	57	45	60	38	42	56	39	56
ita	18	25	4	40	17	50	50	20	40	25	48	40	46	32	47	28	36	45	39	48
lat	85	107	131	6	97	98	45	96	39	87	97	6	77	72	113	93	32	64	54	32
spa	6	29	20	43	4	49	57	14	39	22	47	43	41	30	49	25	42	58	37	50
est	57	56	85	47	65	5	63	83	41	57	48	47	52	46	32	50	55	80	39	55
rus	129	162	177	62	151	160	6	173	90	160	214	62	121	126	177	148	60	26	108	71
por	14	30	22	48	15	57	62	3	44	27	53	48	48	35	65	31	45	60	42	55
hun	58	91	119	52	78	59	59	108	6	77	52	52	53	47	74	72	58	76	60	53
fra	18	30	32	43	24	42	50	27	38	4	50	43	36	23	49	36	47	56	41	57
tur	64	101	126	63	75	59	84	122	46	75	5	63	56	60	64	80	74	113	56	67
cz	85	107	131	6	97	98	45	96	39	87	97	6	77	72	113	93	32	64	54	32
deu	42	49	57	48	47	60	57	62	37	37	53	48	4	32	74	60	53	56	52	53
eng	30	54	61	45	42	43	51	61	38	33	53	45	38	4	62	56	52	62	65	63
fin	66	119	134	50	82	25	64	172	41	73	46	50	55	55	5	87	75	119	55	58
epo	31	35	48	41	31	40	55	36	38	40	46	41	50	42	42	4	38	47	31	41
hrv	52	51	62	23	53	69	32	49	39	62	71	23	55	58	82	48	6	31	37	30
bul	198	183	189	106	214	245	19	213	140	232	268	106	173	185	258	190	84	5	160	133
lit	68	64	97	48	77	56	74	73	50	77	72	48	68	68	70	58	47	69	6	56
pol	105	145	181	33	115	134	63	127	53	119	169	33	99	112	148	136	43	83	62	6

Table 3: Little Prince perplexities with modified normalization

Table of normalization effects

	bul	cat	cz	deu	eng	epo	est	fin	fra	hrv	hun	ita	lat	lit	pol	por	ron	rus	spa	tur
bul	0	-32	-22	-15	-42	-29	35	39	-42	-16	-53	-82	-1	-24	-124	-57	-27	-1	-94	-2
cat	-21	0	0	-4	-1	-1	-9	-13	0	1	-3	0	-1	-12	-22	0	-3	-11	-1	-2
cz	-17	-4	0	-16	-2	6	-21	-10	-2	0	-16	5	1	-3	-24	-5	-9	-9	-3	0
deu	-37	-11	-14	0	-3	-13	-14	-22	-8	-14	-22	-12	-10	-21	-12	-13	-11	-40	-14	-12
eng	-44	-6	-13	-7	0	-15	-32	-31	-8	-16	-14	-15	-18	-23	-12	-18	-15	-26	-11	-12
epo	-24	-1	0	-7	7	0	-15	-17	0	-2	-5	9	-8	-12	-25	-1	-6	-10	-1	-16
est	3	-12	-10	4	-12	0	0	0	-7	3	0	-2	13	-13	-26	-4	-8	-13	-18	-5
fin	46	-45	-48	-10	-45	-2	-7	0	-5	4	-37	3	49	-2	-24	6	55	-35	-55	-20
fra	-27	-1	-1	-4	-1	0	-12	-13	0	0	-4	1	-1	-9	-17	-1	-3	-20	-1	-2
hrv	-9	1	0	-6	-4	4	-1	-1	-2	0	-9	12	0	-6	-18	-1	-6	-7	-2	9
hun	-26	-15	-13	-8	-21	-2	-13	-15	-5	-10	0	0	-3	-29	-30	1	-10	-21	-14	1
ita	-26	0	-1	-5	-1	0	-14	-16	0	0	-8	0	-1	-10	-19	0	-3	-21	0	-1
lat	5	8	1	22	-8	-2	22	33	15	2	-7	0	0	19	-6	-5	37	6	-13	58
lit	-67	-38	-9	-33	-47	-39	-31	-26	-26	-39	-42	-63	2	0	-50	-36	-61	-51	-65	-20
pol	-94	-64	-33	-31	6	-15	-222	-155	-37	-41	-60	-43	-26	-77	0	-34	-45	-61	-43	-19
por	-24	0	0	-1	1	1	-10	-13	0	2	-4	2	1	-8	-10	0	-2	-12	0	-2
ron	-39	-16	-30	-19	-33	-32	-46	-34	-10	-29	-41	-23	1	-21	-17	-19	0	-62	-28	-14
rus	1	-3	0	-28	7	7	4	23	-10	3	-18	-3	4	-7	-71	-18	-23	0	-24	10
spa	-31	0	0	-9	0	0	-25	-18	0	0	-4	1	-1	-9	-8	-1	-3	-16	0	-3
tur	-28	17	-8	-14	-17	9	-57	-34	-5	11	-1	15	70	-5	-53	-37	1	-40	-19	0

Correlation

- Yves suggested Mantel test!
- It ”measures the correlation between two matrices typically containing measures of distance.” (idre.ucla.edu, 2018)
    - While multiple libraries in R can perform the test, we chose the ncf library (Bjornstad & Bjornstad, 2018), for it outputs easily interpretable numbers like correlation and p-value.
    - This function requires symmetric distance matrices, so we averaged the perplexities to fit it.

Correlation

- mantel.test(lp, bibles, resamp = 1000)
- correlation: 0.8494259
- p-value: 0.001998002

We can say with 95% confidence that the matrix entries are positively associated, i.e. the perplexity values depend on the languages rather than genres.

Visualization

- Yves suggested Hierarchical Clustering!
- We used hclust (Müllner et al., 2013) in R with its default complete linkage method, and got a pair of hierarchical trees from matrices.

Visualization

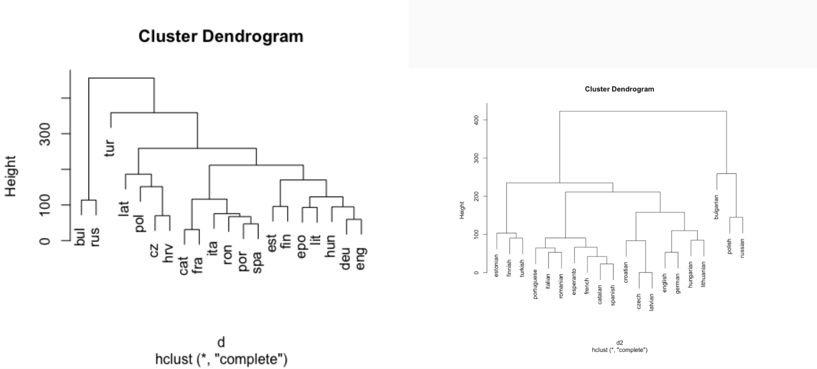


Figure 2: The dendrograms for the Bible (left) and the Little Prince (right) corpora with complete method

Visualization

- Similar matrices, different trees?
- Guess 1: We blame the agglomeration method.
- To test guess 1: We tried different methods in hclust, and got different results than previous ones.

Visualization

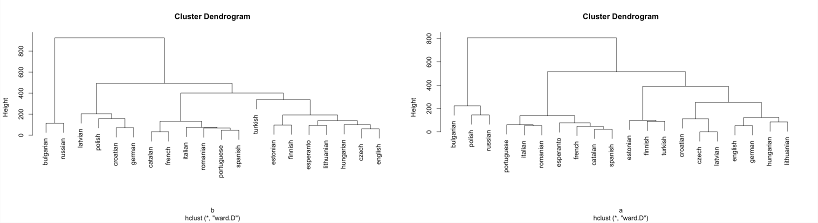


Figure 3: The dendrograms for the Bible (left) and the Little Prince (right) corpora with ward.D method

Visualization

- ... but they still do not agree with each other :(
- Guess 2: We blame the whole algorithm.
- To test guess 2: We can't. Too innocent and too little time.
- On the other hand, the trees are not that different; especially if you compare those produced by ward.D method.

Visualization

- We believe the result from Mantel test can prove this perplexity-based method is stable enough across genres; therefore, Q.E.D.!
- According to Gamallo et al. (2017), ”the language distance we have defined intends to measure interactions among languages from a synchronic perspective. The most suitable representation for this type of data is not a hierarchical tree but rather a network showing language interactions.”

Possible improvements

- If only we knew some methods that could produce networks!
- Normalization
- More consistent material, more material
- Running a new test with fixed mistakes

Thank you!

References

Bjornstad, Ottar N & Maintainer Ottar N Bjornstad. 2018. Package ncf .

Digitais, Livros. 2010. Portuguese online library@ONLINE. https://www.livros-digitais.com/antoine-saint-exupery/o-principezinho/77.

Gamallo, Pablo, José Ramom Pichel & Iñaki Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications* 484. 152–162.

Gardner, Matt, Mark Neumann & Joel Grus. 2018. Writing code for nlp research, emnlp 2018. https://github.com/allenai/writing-code-for-nlp-research-emnlp2018.

idre.ucla.edu. 2018. How can i perform a mantel test in r, r faq. https://stats.idre.ucla.edu/r/faq/how-can-i-perform-a-mantel-test-in-r/.

References ii

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania* 135(273). 40.

Müllner, Daniel et al. 2013. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software* 53(9). 1–18.

Pablo Gamallo, Inaki Alegria, Jose Ramom Pichel. 2018. Language distance measure. `https://github.com/gamallo/Perplexity`.