Statistical Classification of Languages

Generalising Ward's Method for Use with Manhattan Distances

Trudie Strauss[1,2] Michael J. von Maltitz[2]

[1]Institute for Bioinformatics
University of Leipzig
Germany

[2]Department of Mathematical Statistics and Actuarial Science
University of the Free State
South Africa

TBI Winterseminar, Bled, 2016

Outline

Motivation

- The question arises whether groupings of languages, similarities between languages and language traits well known in the field of linguistics can be extracted or independently observed using unsupervised techniques; that is, whether it is possible to autonomously classify languages without any prior linguistic knowledge or assumptions.

Background

- August Schleicher: one of the first to suggest in 1873 that languages follow the same evolutionary process as Darwin suggested biological organisms do in nature.
- This contributed to the foundation being laid for Comparative Linguistics - and Quantitative Comparative Linguistics.
- Methods of language classifications are based on biological classifications.

Quantifying Languages

Suggestion: Quantifiable/statistical language signature (SLS)

- We count the number of bi-gram (adjacent pairs of letters) occurrences in a language
- The matrix of the relative bi-gram frequencies in a language constitutes that language's SLS.

Example: bi-gram frequency

We know that the bi-gram "th" is observed much more in English and "en" much more in German.

- Using this SLS, we compute the distances between languages.
- Based on this distance matrix, we are able to do cluster analysis.

Process

- 32 Indo-European languages are analysed.
- Previous authors suggest the use of translations of the Universal Declaration of Human Rights as corpus.
- Advantage: the different texts are more or less the same in length.
- However: problem of loanwords could bias results when assessing the proximity between languages.
- For this reason, we expand our analysis to a corpus of non-parallel newspaper texts. For languages where newspaper texts weren't available translations from the Universal Declaration of Human Rights or the Bible were used (Asturian, Breton, Friulian, Scottish and Welsh).

- While all the selected languages use the Latin alphabet, we have to include diacritics.
- We introduce an alphabet consisting of 65 characters: the 26 letters of the Latin alphabet, blank spaces between characters and 38 special characters.

Table: Table of characters used for analysis

a	b	c	d	e	f	g	h	i	j	k
l	m	n	o	p	q	r	s	t	u	v
w	x	y	z	_	ä	à	á	â	å	ã
æ	ç	ê	ë	è	é	ì	í	î	ñ	ö
ø	ò	ó	õ	ô	š	ß	ü	ù	ú	û
ý	ž	ś	ź	ð	ż	ł	ć	ą	ę	

- For each language 65×65 matrix of relative bi-gram frequencies, with entries $RF(\alpha, \beta) = \frac{n_{\alpha\beta}}{(n-1)}$
- This SLS is then used in the distance calculation between languages
- Simple distance measurement: Manhattan (l_1 norm) distance

Manhattan Distance

$$D_{Manhattan}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{65} \sum_{j=1}^{65} \left| a_{ij} - b_{ij} \right|$$

where a_{ij} represents the ij^{th} element of the SLS matrix for language **a**.

Statistical Language Signature and Distance Calculation

	a	b	c	d	e	f	g	h	i	j	k	l
a	66	1440	3109	2915	138	924	1564	196	3409	132	995	7145
b	1750	120	68	16	3972	1	3	6	914	34	0	1390
c	3870	23	577	22	4425	15	12	3869	1790	2	1489	1113
d	2123	22	53	347	5388	24	245	29	3061	30	3	236
e	5948	412	3054	8417	3056	1118	907	244	1313	26	308	3631
-												
l	3935	70	66	2018	6153	262	41	15	4743	0	207	4968
m	4059	693	102	21	5801	22	8	7	2529	0	3	32
n	2461	86	2568	8539	5598	485	8358	81	2510	114	563	382
o	608	725	1107	1102	303	5650	538	148	757	78	635	2600
p	2432	11	32	36	3425	18	18	494	930	1	3	2056
q	24	1	0	0	1	0	0	0	2	0	0	0
r	4420	184	907	1572	12631	190	791	90	4974	5	892	727
s	3046	76	1053	317	6126	80	28	2432	3582	1	347	435
t	3718	79	331	33	8393	48	22	23291	7297	3	11	639
u	722	603	1047	695	1022	112	921	18	596	7	60	2083

Statistical Language Signature and Distance Calculation

	Afrikaans	Asturian	Bosnian	Breton	Catalan	Corsican	Czech	Danish	Dutch	English	French	Frisian
Afrikaans	0	0,96515	1,0102	0,89913	0,86591	1,02001	1,02262	0,59663	0,39112	0,6794	0,81761	0,54186
Asturian	0,96515	0	0,99925	1,04379	0,52906	0,79417	1,02723	0,97517	0,97343	0,82293	0,65994	1,00643
Bosnian	1,0102	0,99925	0	1,11275	0,94135	0,85996	0,69827	0,96544	0,99022	0,99511	1,0363	1,03927
Breton	0,89913	1,04379	1,11275	0	0,92933	1,10858	1,13446	0,90414	0,84986	0,96679	0,97215	0,94902
Catalan	0,86591	0,52906	0,94135	0,92933	0	0,6921	1,0281	0,84858	0,85455	0,73635	0,5067	0,92676
Corsican	1,02001	0,79417	0,85996	1,10858	0,6921	0	1,09125	1,01112	1,01452	0,86632	0,8022	1,04947
Czech	1,02262	1,02723	0,69827	1,13446	1,0281	1,09125	0	1,02036	1,03145	1,00025	1,04501	1,06281
Danish	0,59663	0,97517	0,96544	0,90414	0,84858	1,01112	1,02036	0	0,56734	0,70068	0,80991	0,63004
Dutch	0,39112	0,97343	0,99022	0,84986	0,85455	1,01452	1,03145	0,56734	0	0,69883	0,84201	0,48651
English	0,6794	0,82293	0,99511	0,96679	0,73635	0,86632	1,00025	0,70068	0,69883	0	0,67513	0,7221
French	0,81761	0,65994	1,0363	0,97215	0,5067	0,8022	1,04501	0,80991	0,84201	0,67513	0	0,85856
Frisian	0,54186	1,00643	1,03927	0,94902	0,92676	1,04947	1,06281	0,63004	0,48651	0,7221	0,85856	0

Ward's Linkage Algorithm

- Ward's method: Minimisng intra-cluster variation and maximising inter-cluster variance.
- Joins the two clusters A and B that minimise the increase in the sum of squared errors (SSE):

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

- Because we use the Manhattan distance, we change the objective function from minimising SSE to minimising Absolute Deviation.
- We show that the Lance Williams Parameters for this objective function are the same as for the objective function of minimum SSE.

Therefore, the updating function for the distance matrix follows:

Updating the Distance Matrix

If points i and j are combined into cluster ij, then the distance between the new cluster ij and another cluster k, is defined as:

$$d_{k(ij)} = \frac{n_i + n_k}{n_i + n_j + n_k} d_{ki} + \frac{n_j + n_k}{n_i + n_j + n_k} d_{kj} - \frac{n_k}{n_i + n_j + n_k} d_{ij}$$

We use the `hclust` package in `R`, specifying the option `method = "ward.D2"`

- Results from clustering summarized in Dendrograms
- Because we expanded Ward's Method to Manhattan distances, we include Euclidean distance Results for the sake of comparison
- Results are then compared to show that original characteristic of Ward's Method is still in tact with Manhattan distances

Figure: Ward's Linkage using Euclidean Distances

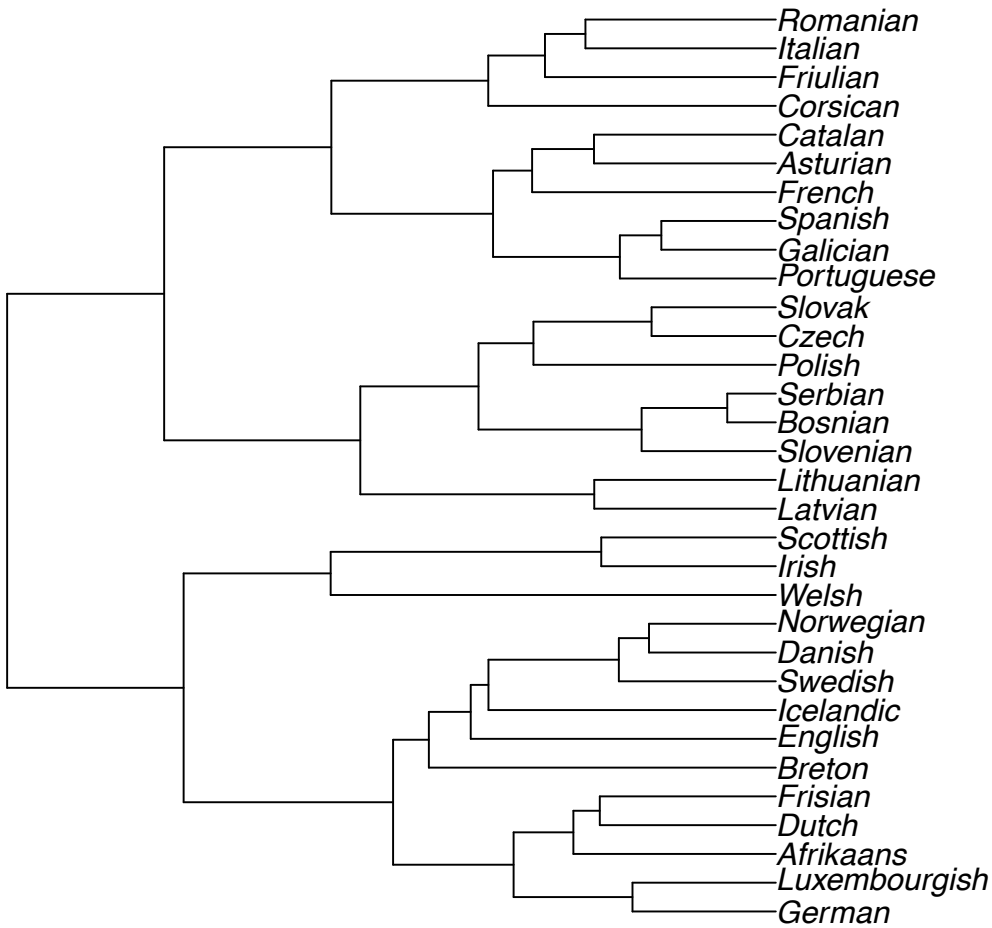
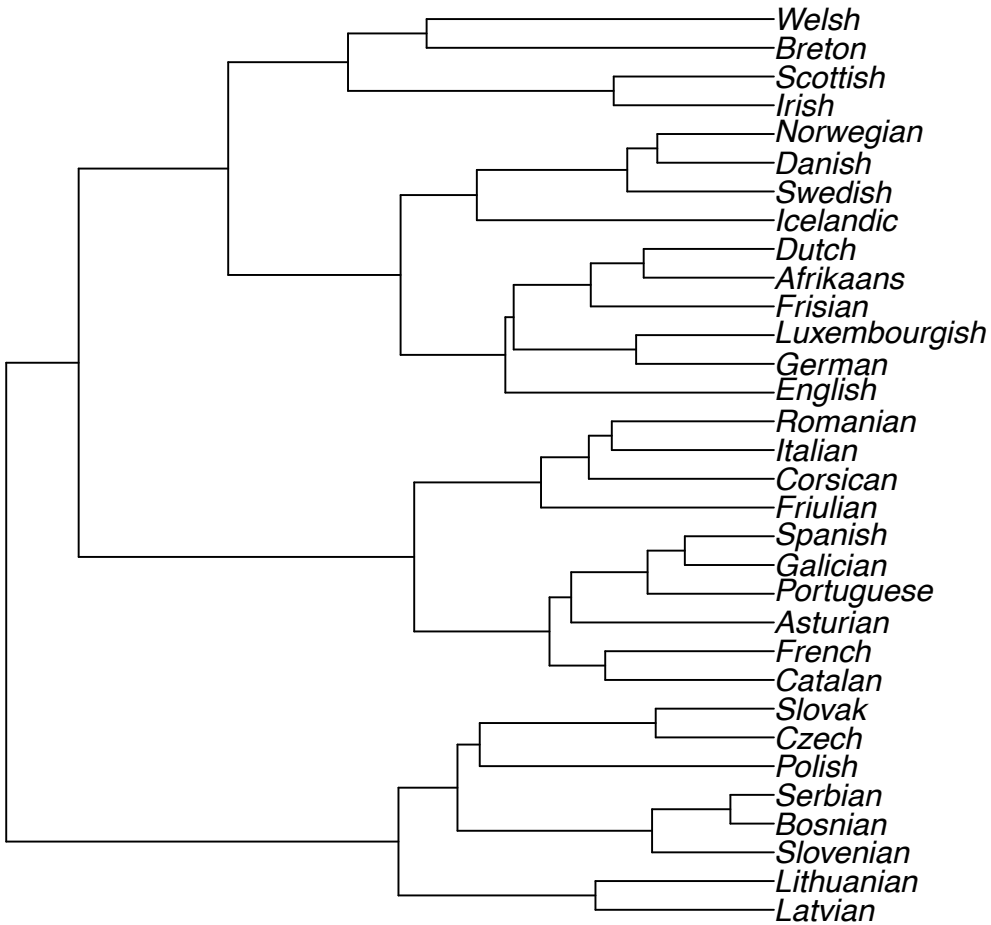
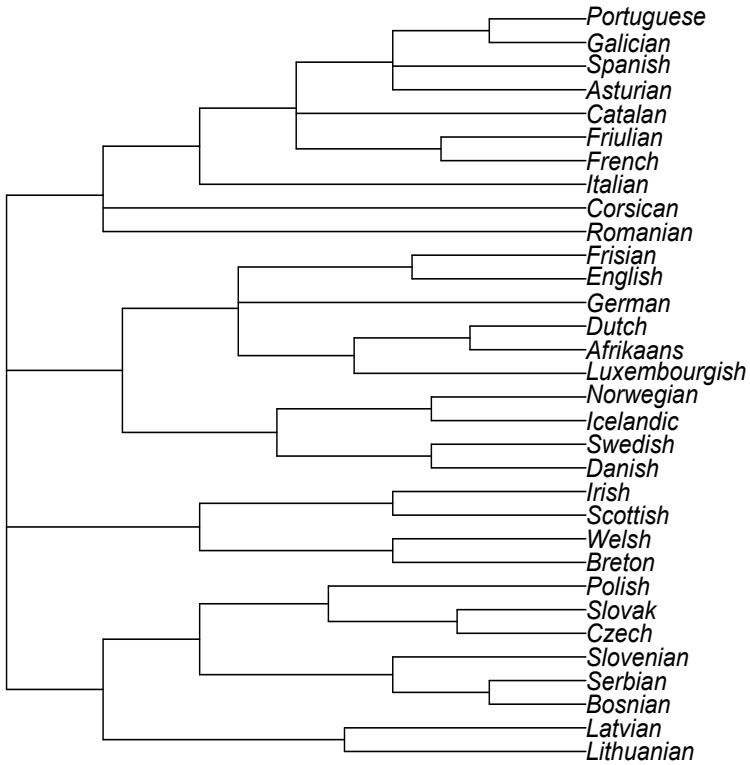


Figure: Ward's Linkage using Manhattan Distance



We compare the trees with a benchmark tree created, without branch lengths from info obtained from Glottolog 2.6.



Robinson-Foulds distance between trees:

- Euclidean: 33
- Manhattan: 21

Cluster Validation Measures

Table: Comparison of Cluster Validation: Euclidean Distance vs. Manhattan distance

Cluster Characteristic	Validation Measure	Euclidean Distance	Manhattan Distance
Compactness and Separation	Silhouette Width	0.2129	0.2571
	Dunn Index	0.5557	0.6246
Connectedness	Connectivity	17.10	16.52

Summary

- We were able to quantify languages, determine distance and cluster them purely statistically.
- We expanded Ward's Method to include use of Manhattan distance matrix.
- We showed that using Manhattan distance doesn't violate the characteristic of Ward's Method (minimising within-cluster variation, and maximising between-cluster variation)

Thank you

University Leipzig

- Peter Stadler
- Nancy Retzlaff
- Christian Höner zu Siederdissen

UNIVERSITÄT LEIPZIG

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA

UFS UV

University of the Free State

- Michael von Maltitz
- Sean van der Merwe