

Validating Dialectometry

John Nerbonne

LSA.107 Dialectology in the Aggregate

Rijksuniversiteit Groningen

j.nerbonne@rug.nl

<http://www.let.rug.nl/alfa/>

Summer, 2005





Overview

- Problem of Validation
- Checking partial results
- Consistency
- Validity vis-à-vis a “gold-standard”
 - Metric Perspectives
 - “ F ” ratio, Fischer linear discriminant
 - Non-Metric Perspectives
- Lay Perception of Dialect Differences



Problem of variants

Have you understood, or have you just worked hard?

- “Embarrassment of riches”: too many variants
- Especially bothersome if some techniques work, others don’t
- Statistician’s admonition: Data snooping!



Lexical Distance: Variants

- treatment of rare variants
- treatment of multiple responses
- treatment of inflectional variants
- treatment of imperfect ling. overlap among sites



Levenshtein distance: variants

- Treatment of diacritics
- Relative costs of indels vs. substitutions
- Diphthongs: one segment or two?
- Refinement: symbol-based vs. feature-based
- Feature systems: Hoppenbrouwers (1988) (SPE-like); Vieregge et al. (1984); Almeida & Braun (1986); LAMSAS; ...?
- Feature weighting, e.g. via information-gain
- Log. correction on sum of feature differences



Strategy: Checking Partial Results

GA1G! Savannah --- PA7C Lancaster Co.

'æ _v əftə, nʌ·n						---						,æftər'nu·n					
æ _v	ə	f	t	ə		n	ʌ·	n									
æ		f	t	ə	r	n	u·	n									
0.0594	0.5144	0.5144	0.5144	0.5144	1.3644	1.3644	1.6152	1.6152									
'	æ	v	{	ə	}	f	t	ə		,	n	u	-	·	n		
,	æ					f	t	ə	r	'	n	u		·	n		
2	2	3	4	5	6	6	6	6	7	9	9	9	10	10	10		

- Sensible to inspect, e.g., edit distance alignments
- Problems: (i) we'd like to quantify quality if possible
(ii) inspection results only in relative checks on analyses



Strategy: Compare to Expert Opinion

- Compare to expert consensus (when available)
- Problem: this occurs most in the form of classification into dialect areas
- “Solution”: cluster first, then compare
- Problem: INSTABILITY of clustering



Data source

- Reeks Nederlands(ch)e Dialectatlassen (RND)
 - Contains 1,956 translations of 141 sentences
 - Variants in the Netherlands and North Belgium
 - Texts from 1925–1982
 - Texts in phonetic transcription
 - About 2-5 informants per site
 - Transcriptions by professional phoneticians



Data source: Problems?

Limits on data quality

- 53 years is a long interval
- Reliability of informants
- Representativity of sentences chosen
- Missing words or phrases can not be supplemented
- Meaning of missing data not always clear

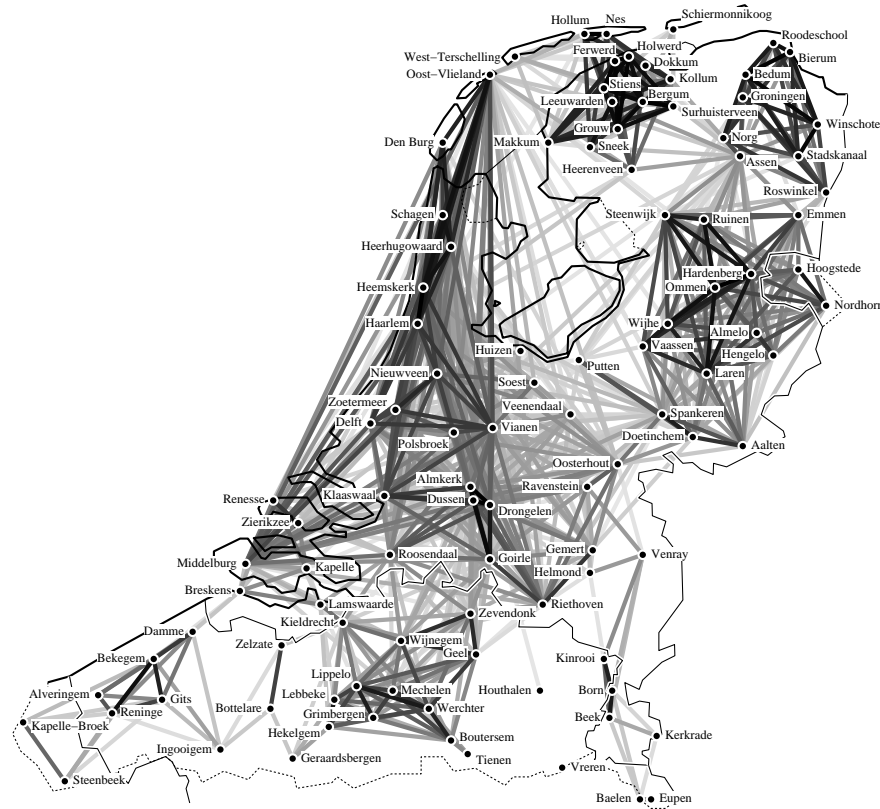


Choice of 104 Dialects





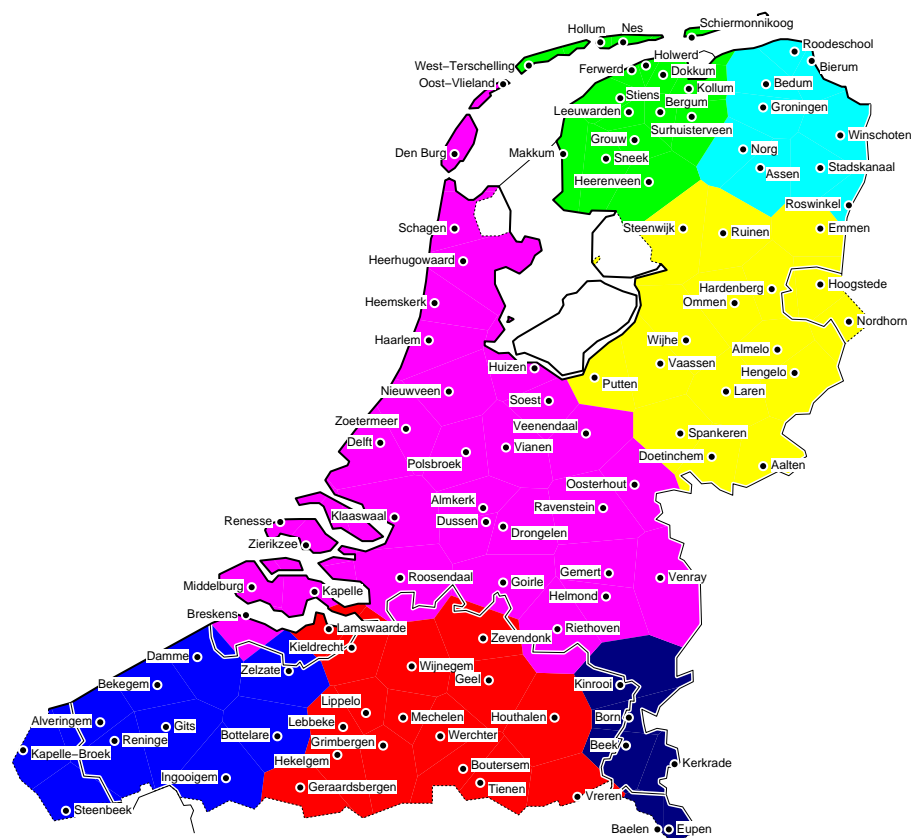
Mapping Distance



No dependence on clustering!



Clustering



7 most significant groups (Ward's method) agree with traditional dialectology.



MDS Separation



$x \rightarrow$ 3rd dim., $y \rightarrow$ 1st, and darkness \rightarrow 2nd. Left above Frisian, Saxon, Franconian.



Consistency

- Interitem correlation (per word pair) is $\bar{r} = 0.15$

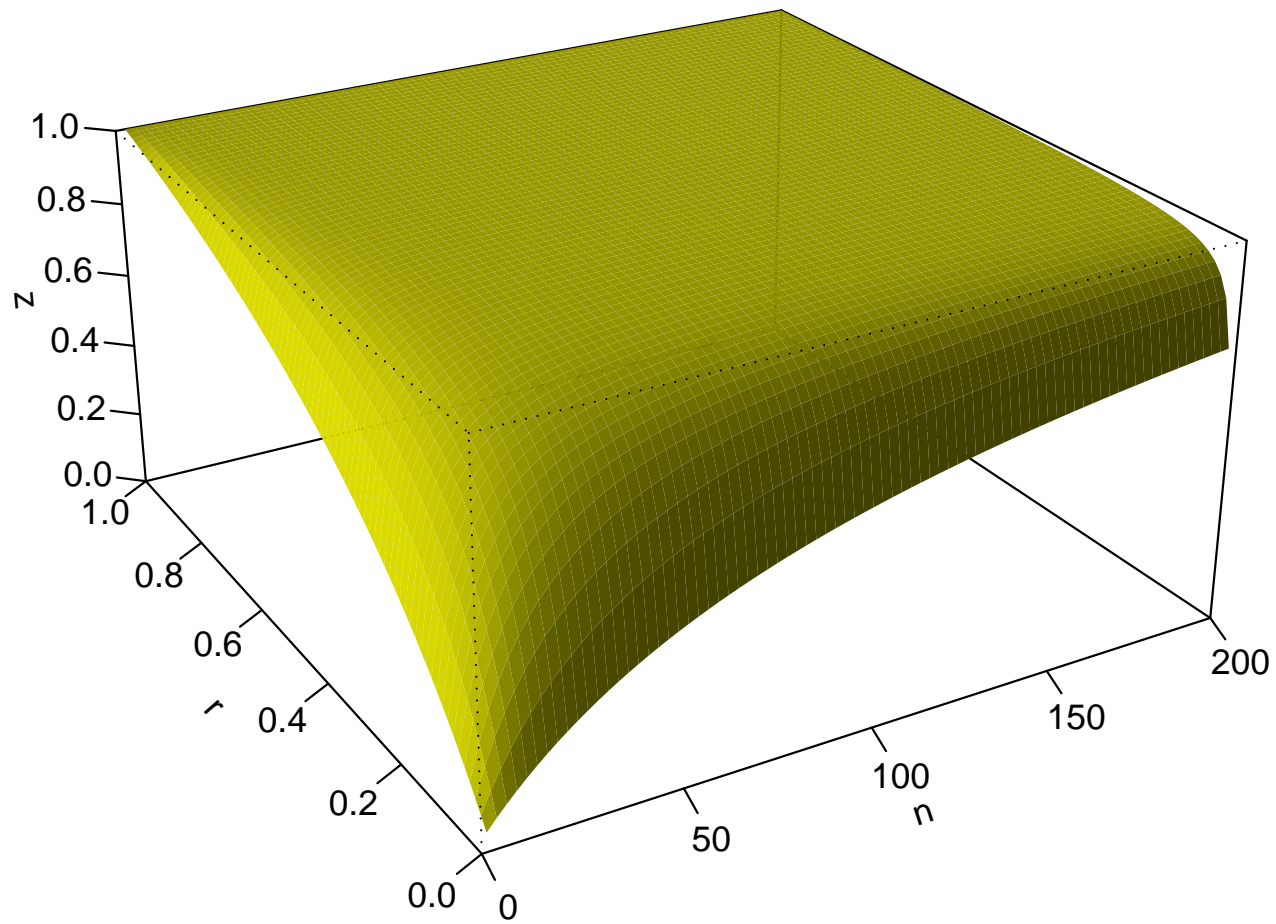
Cronbach's α for 100 items is

$$\begin{aligned}\alpha &= \frac{k*\bar{r}}{1+(k-1)*\bar{r}} \\ &= \frac{100*0.15}{1+99*0.15} \\ &= \frac{15}{15.85} \\ &= 0.95\end{aligned}$$

Very reliable



Cronbach Alpha





Validation

- Agreement with new data

Master's Thesis by Johan Dijkstra used exactly the same parameters (feature vector distance, weighting, edit distance, minimal error clustering), chosen earlier and replicated good classification on 40 previously unstudied Dutch varieties.

- **later:** 200 more varieties



Validation

- Agreement with expert opinion

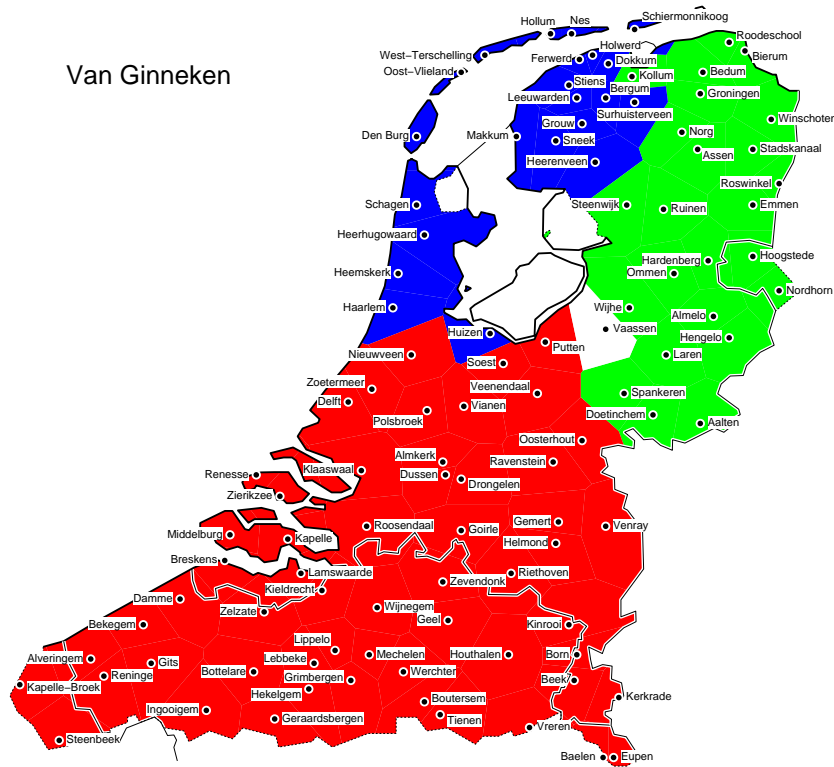
idea: define “gold standard” where expert opinion agrees

test agreement with distance metric

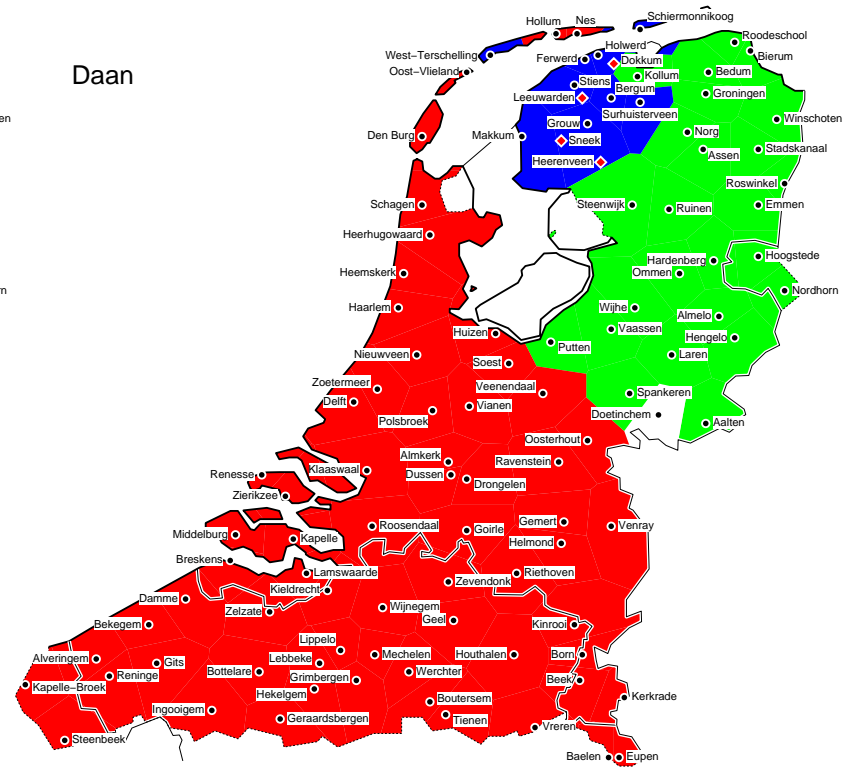


Van Ginneken/Wijnen vs. Daan

Van Ginneken



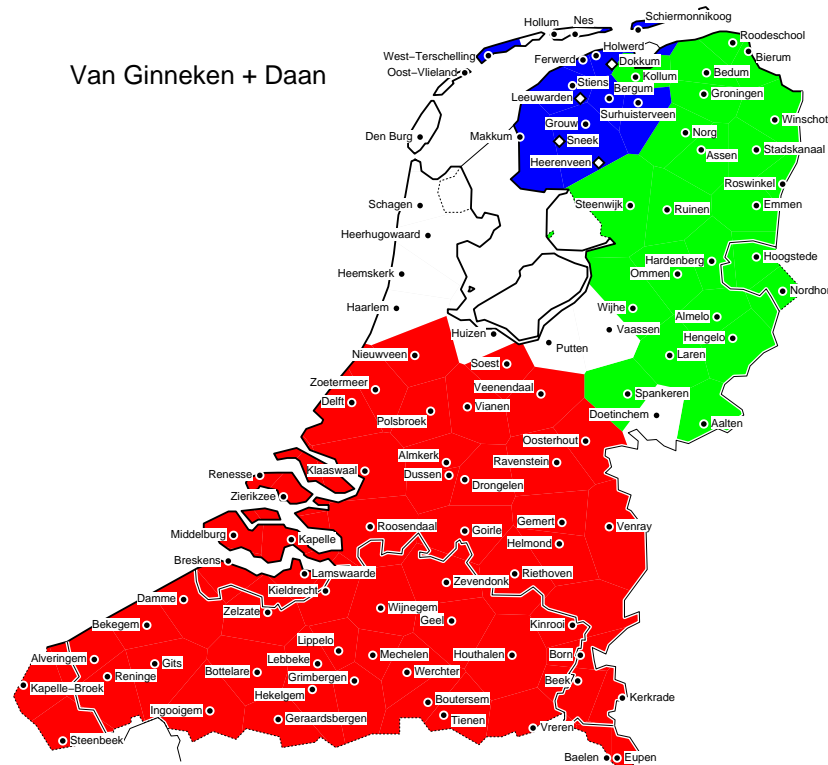
Daan





Toward Gold Standard

Van Ginneken + Daan



Classification à la van Ginneken/Wijnen \cap Daan



Toward GS Validation

- Define “gold standard” where expert opinion agrees
- Metric perspective:

test agreement with distance metric

all like varieties, v, v' , are closer to each other than to unlike varieties v_d

$$d(v, v') \ll d(v, v_d)$$

- Non-metric perspective:

test overlap of grouping



Metric Validation

- F ratio

Given partition of dialects into groups, distinguish ave. **within-group distance** 'In' and ave. **without-group distance** 'Out'

Evaluation: minimize $F = \frac{\text{In}^2}{\text{Out}^2}$



Metric Validation 2

- Fischer's Linear Discriminant

Adopt vector-space perspective: each dialect variant identified with position in 100-dimensional space (distances to 100 dialects)

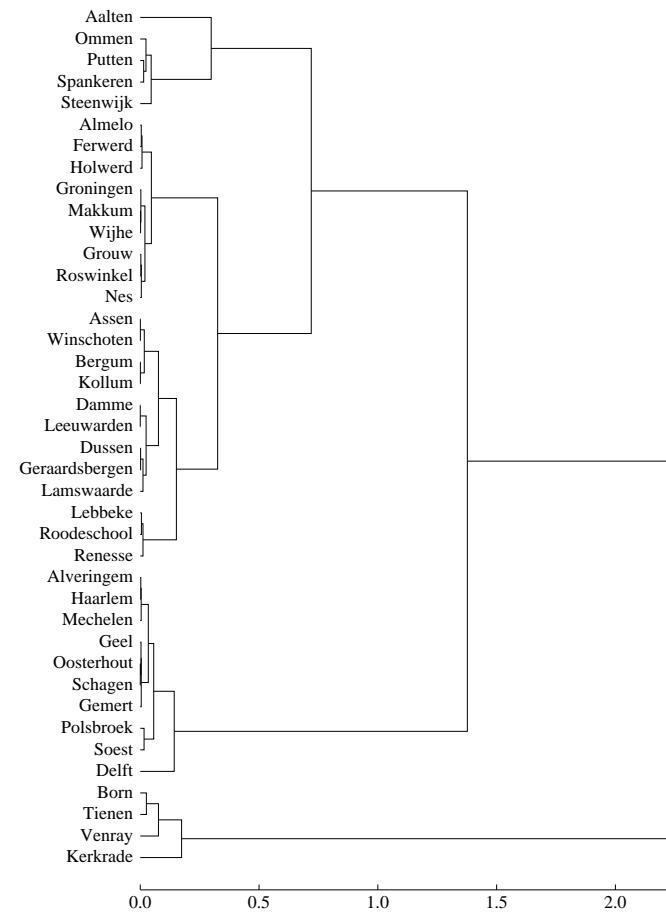
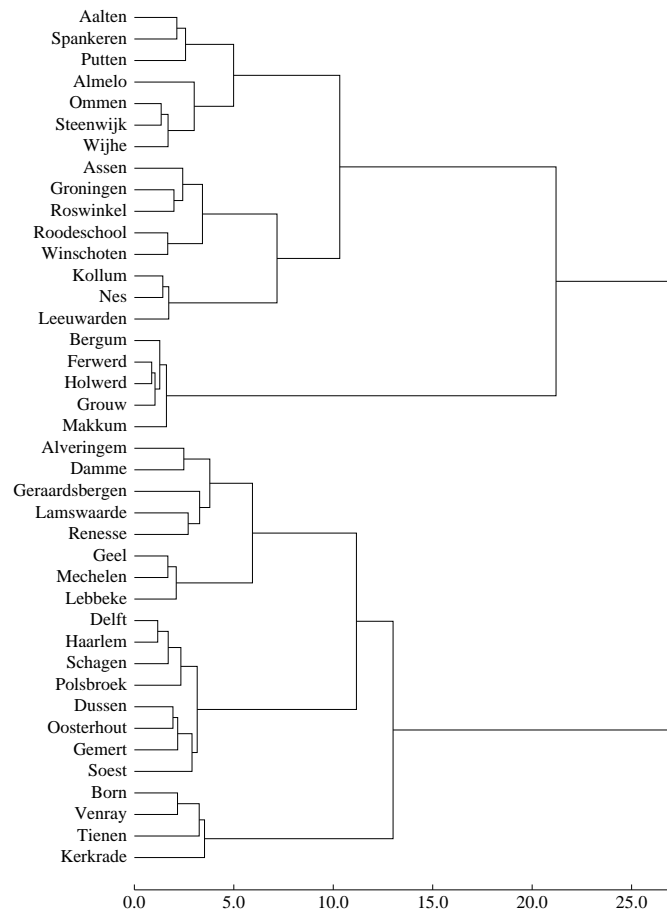
- Calculate mean, variance of each group (given by gold standard)
- For each pair of groups, each dimension, calculate

$$J(i, j) = \frac{(\bar{m}_i - \bar{m}_j)^2}{\sigma_i^2 + \sigma_j^2}$$

Evaluation: maximize the discrimination



Metric Evaluation (Right) Emphasizes Contrast





Non-Metric Evaluation

Two bases for evaluation

- Clustered results of distance measure
- n -nearest elements wrt distance

Two evaluation methods

- Choose method which best respects gold standard groupings, respects most classification pairs in gold standard
- Matrix comparison of gold-standard grouping to method under evaluation



Fowlkes/Mallows's Matrix Comparison

imperfect grouping

	1(2)	2(35)	3(3)
A(22)	2(0.09)	17(0.43)	3(0.14)
B(13)	0	13(0.37)	0
C(5)	0	5(0.14)	0
Total	0.09	0.94	0.14

score = 1.17/3

perfect grouping

	1(2)	2(35)	3(3)
A(22)	0	22(1)	0
B(13)	13(1)	0	0
C(5)	0	0	5(1)
Total	1	1	1

score = 3/3

Non-metric evaluations consistently choose techniques found linguistically sound (see paper).



First Conclusions

- Comparison with expert opinion valuable even if it's not an unconditional sign of quality.
- Consistency straightforward to assay.
- Non-metric validation needed to avoid emphasis on contrast.
- Instable clustering undesirable element in validation
- Correlation with perceptive distance?



Perceptual Distance as Quality Criterion

Two traditions in dialectology

- Linguistically based: phonemes, features, words, ..
- Layman's judgement: what's strange is strange

Leading idea: let's test the linguistic approach using lay judgements

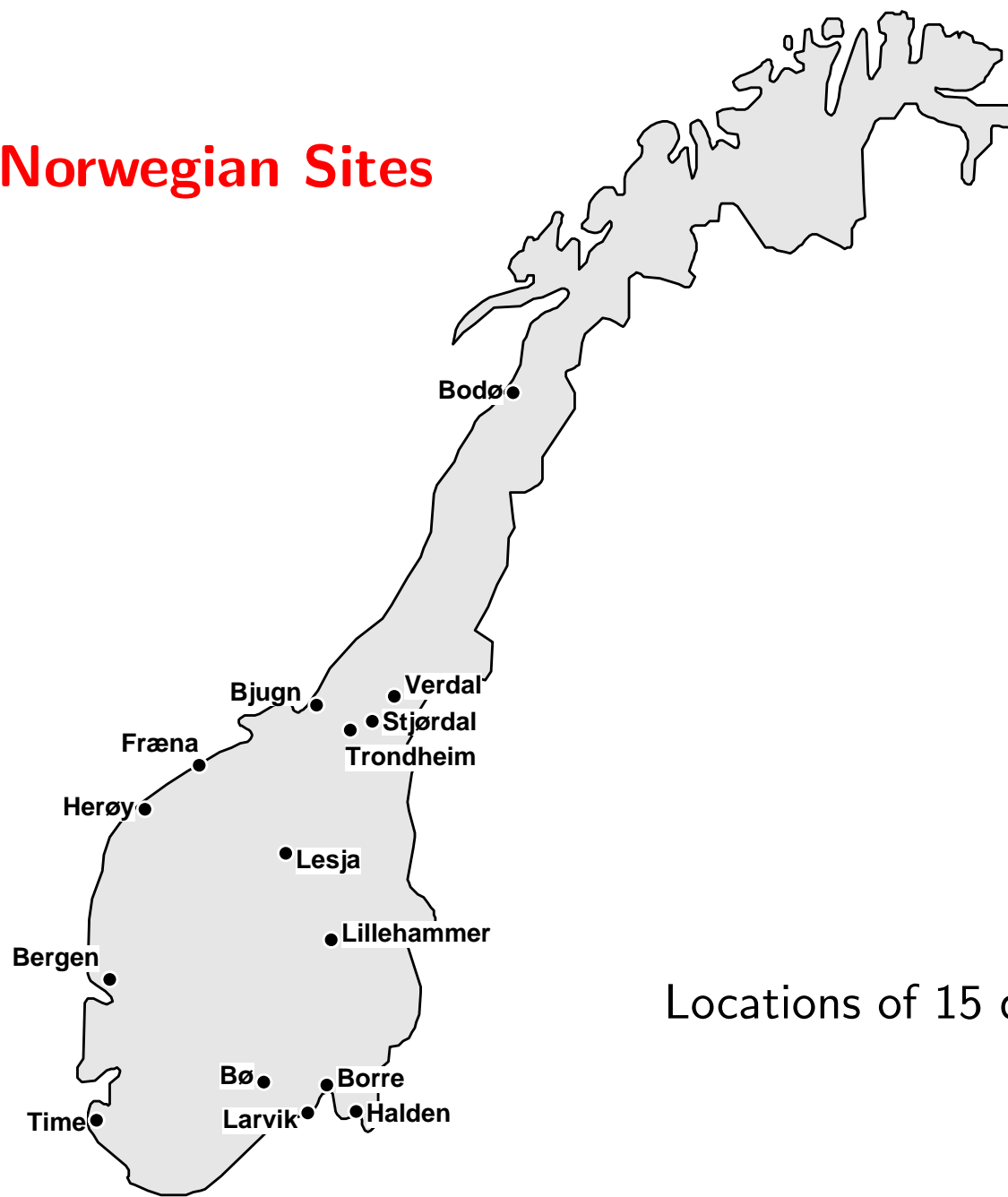


Perceptual Validation

- Results of computational methods are compared to distances according to the perception of dialect speakers.
- Jørn Almborg's material, Norwegian dialect translations of the 'The North Wind and the Sun' avail. at <http://www.ling.hf.ntnu.no/nos/> (audio files *and* transcriptions)
- For the perceptual distance measurements the complete texts were used.
- For the computational distance measurements we used the 58 different words from the text.
- Multiple pronunciations of one word are processed.



Norwegian Sites



Locations of 15 dialects.



Perceptual “Distance”

- Charlotte Gooskens conducted a perception experiment using the audio files.
- In each of the 15 locations, a group of 16 to 27 high school pupils listened to all 15 texts, presented in random order.
- Task: each pupil notes for each text the “distance” of the dialect from his own dialect.
- Scale from 1 (similar to own dialect) to 10 (not similar to own dialect).
- Final result: a 15×15 perceptual distance matrix.



Correlation with perception

We correlate the distance matrices of the various computational methods with the perceptual distance matrix.

	Freq. corp.	Freq. word	Lev. lin.	Lev. log.
phones	0.66	0.66	0.67	0.67
features	0.46	0.59	0.62	0.64
acoustic			0.64	0.66

The effect of methods, segment representations and segment distance metrics is shown in the average correlation coefficients of computational distances with respect to perceptual distances.



Validation Results and Conclusions

- Phone-based methods correlate most strongly!
- After that, Levenshtein distances using acoustic segment distances correlate most strongly.
- Examining feature-based methods:
linear Levenshtein distance $<$ logarithmic Levenshtein distance
- It is most important *that* segments differ, not to *what extent*.
- Details may be more important in a network with a higher density.