How representative are linguistic perceptual distance measurements based on a short text?

Judging the whole on the basis of a part

Wilbert Heeringa
Meertens Institute
Variationist Linguistics

Charlotte Gooskens
University of Groningen
Scandinavian Department

Methods XIII
Workshop on Measuring Linguistic Relations between Closely Related Varieties
Leeds, Monday, August 4, 2008

Research question

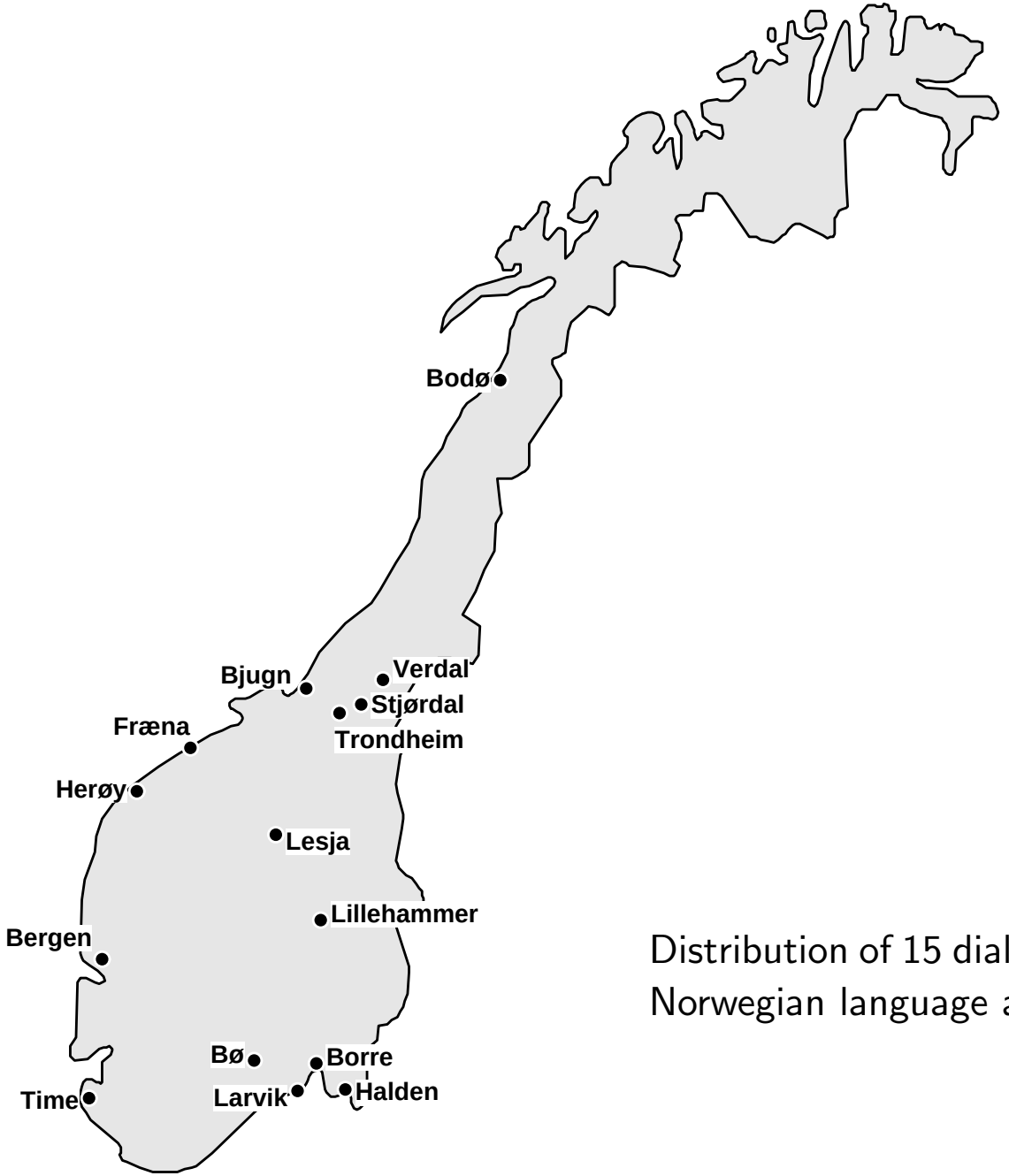
When dialect speakers are asked to judge distances
from their own dialect to other dialects
by listening to recordings of those dialects,
do they use *only* the information they hear,
or do they also use *extra* information
not represented by the recordings?

Perception experiment

- Database of Norwegian dialect recordings collected by Jørn Almberg and Kristian Skarbø in 1999–2002.
- Translations in Norwegian dialects of the fable 'The North Wind and the Sun'.
- Audio files *and* transcriptions available via:

`http://www.ling.hf.ntnu.no/nos/`

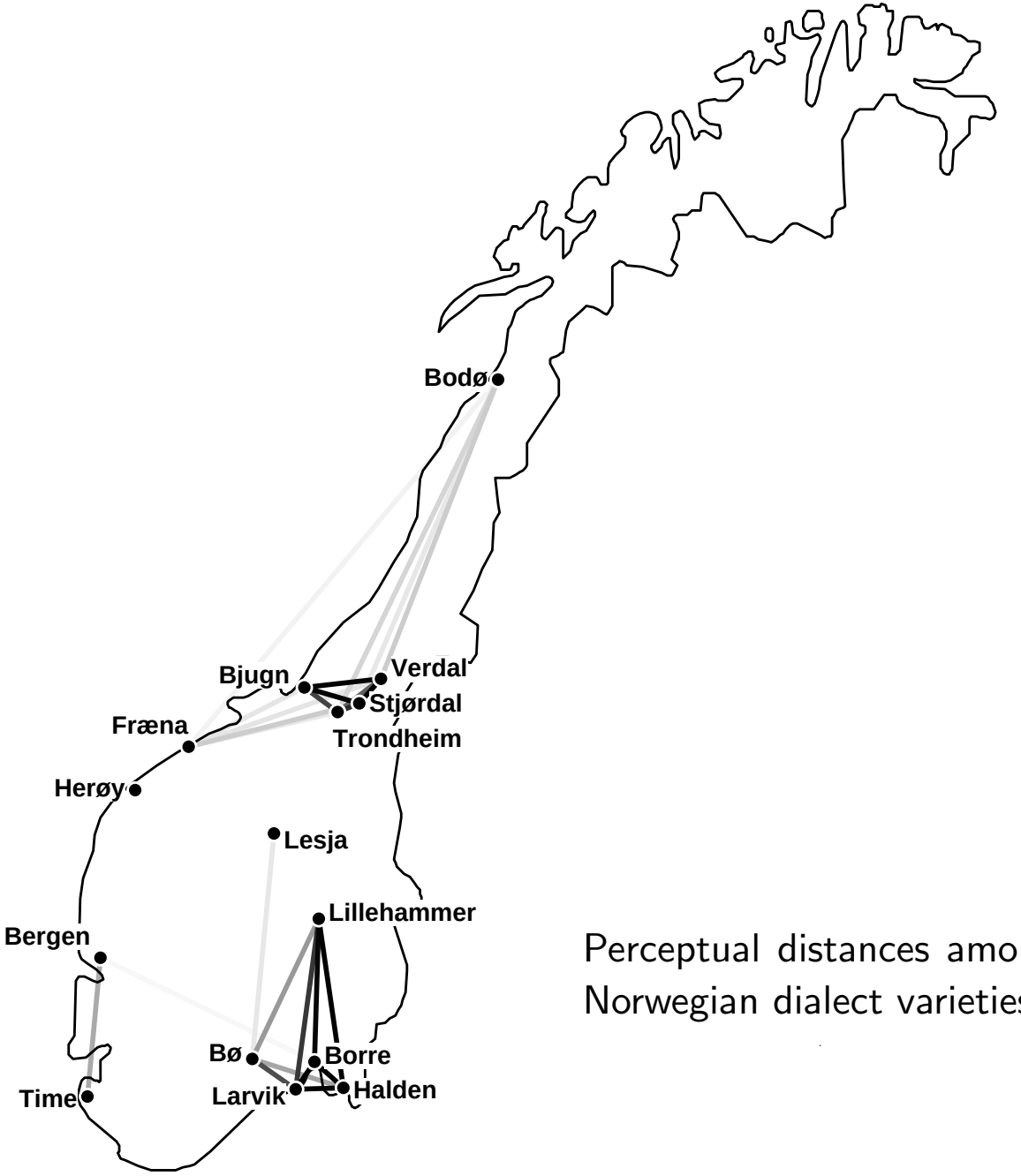
- Perception experiment based on recordings of 15 varieties.
- In each of the 15 locations, a group of 16 to 27 high school pupils listened to all 15 recordings.



Distribution of 15 dialects in the Norwegian language area.

Perception experiment

- The recordings were presented in a randomized order.
- Task for each pupil: to note for each dialect recording the distance to his own dialect.
- Scale from 1 (similar to own dialect) to 10 (not similar to own dialect).
- Final result: a 15 $\times$ 15 perceptual distance matrix.



Perceptual distances among 15 Norwegian dialect varieties.

Perception experiment

- The pupils also put a cross at a map of Norway in the county where they thought that the speaker came from.
- 25% of the crosses were placed in the correct county (with recognition), 75% of the crosses were placed in the wrong county (without recognition).
- Correlations:

	PERC	PERC *with recognition*	PERC *without recognition*
PERC	-	0.84	0.99
PERC *with recognition*	0.84	-	0.78
PERC *without recognition*	0.99	0.78	-

Perception experiment

- Hypotheses:
  1. Subjects do not only use the information they hear, but use extra information not represented by the recordings.
  2. Subjects who guessed the right county are more likely to use extra information when judging the distances.

Data sources

- We measure pronunciation distances on the basis of:
  - The transcriptions of the recordings used for the perception experiment, 58 words per variety. Narrow transcription (NOS)
  - *Atlas Linguarum Europae*. Per variety 547 words, we randomly selected 60 words, rather broad transcription. Material collected end of the seventies in the previous century (ALE).
  - Same as ALE, but transcriptions in Norwegia system. Lillehammer reconstructed (ALENOR).
  - Norwegian dialect atlas. Per variety 1500 words, we randomly selected 60 words. Norwegia system used. Material collected between the forties and the seventies in the previous century (NOR).
- For each dataset the pronunciation distances are correlated with the perceptual distances.

Measuring pronunciation distances

- Levenshtein distance: calculates the cost of changing one string into another. Levenshtein distance = cost of cheapest mapping.
- Example: *gåande* or *gående* 'going' may be pronounced as [ˀgoːɑns] in the dialect of Bø and as [ˀgɔːnə] in the dialect of Lillehammer.
- Alignment (ignoring suprasegmentals and diacritics):

g	o	ɑ	n	∅	s
g	ɔ	∅	n	ə	∅
	1	1		1	1

  We obtain a total cost of 4 operations.
- Distance between two varieties: average Levenshtein distance of 58/60 cognate word pairs.
- Final result: a 15 × 15 pronunciation distance matrix.

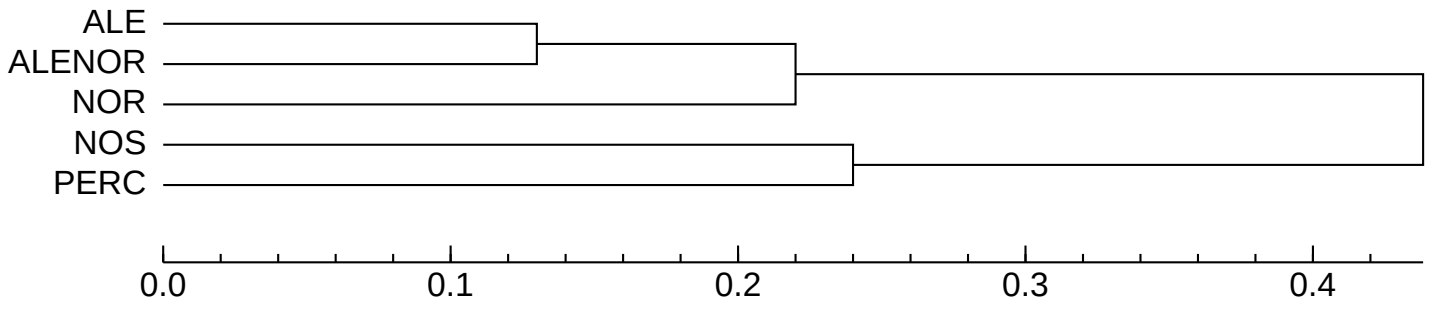
Correlations

- Correlations among perceptual distances and pronunciation distances of different data sets:

	PERC	NOS	ALE	ALENOR	NOR
PERC	-	0.76	0.61	0.60	0.55
NOS	0.76	-	0.49	0.56	0.56
ALE	0.61	0.49	-	0.87	0.75
ALENOR	0.60	0.56	0.87	-	0.81
NOR	0.55	0.56	0.75	0.81	-

Correlations

- From the correlations, distances can be derived by calculating $1 - r$.
- We applied cluster analysis (UPGMA) to these distances.
- We obtained the following dendrogram:



Correlations

- Correlations:

	PERC	PERC *with recognition*	PERC *without recognition*
PERC	-	0.84	0.99
NOS	0.76	0.58	0.78
ALE	0.61	0.61	0.57
ALENOR	0.60	0.59	0.57
NOR	0.55	0.56	0.53

Multiple regression analysis

- We performed multiple linear regression analysis (stepwise):
  - Independent variables: NOS, ALE, ALENOR, NOR
  - Dependent variable: PERC
- Results:

			correlation	significant variables
PERC			0.81	NOS, ALE
PERC	*with*	*recognition*	0.68	ALE, NOS
PERC	*without*	*recognition*	0.81	NOS, ALE

- ALE is a significant variable, probably contains information used by the subjects but not found in the recordings.

Multiple regression analysis

- Correlations:

	PERC	PERC *with recognition*	PERC *without recognition*
NOS+ALE	0.81	0.68	0.81
NOS	0.76	0.58	0.78
increase	0.05	0.10	0.03

- Largest increase for the perceptual measurements based on subjects who correctly guessed the county where the variety is spoken.

Conclusions

- Subjects do not only use the information they hear, but use extra information not represented by the recordings.
- Subjects who guessed the right country seems to use extra information when judging the distances.

Acknowledgements

We thank:

- Benedikte Tobiassen (digitalization of ALE, ALENOR and NOR data)
- Peter Kleiweg (visualization software)