Measuring Socially Motivated Pronunciation Differences

John Nerbonne, Sandrien van Ommen, Charlotte Gooskens & Martijn Wieling
j.nerbonne@rug.nl

Center for Language and Cognition, University of Groningen

Workshop on Comparing Approaches to Measuring Linguistic Differences
University of Gothenburg, 24-25 Oct. 2011
Groningen variationist team!

Charlotte Gooskens, Peter Houtzagers, Hermann Niebaum, Wilbert Heeringa, Jelena Prokić, Therese Leinonen, Martijn Wieling, Marco Spruit, Peter Kleiweg, Christine Siedle, Jens Moberg, ...

... 

Sebastian Kürschner, Alexandra Lenz, Bob Shackleton, Renée van Bezooijen, ...

... 

Bob de Jonge, Agnes de Bie, Cornelius Hasselblatt

... 

Simonetta Montemagni, Franz Manni, Petja Osenova, Esteve Valls, Lucija Šimičić, Kristel Uiboaed, Boudewijn van den Berg
Mutual Comprehensibility of Dutch Dialects

- NWO-FWO project 2007-2010
- Collaboration Groningen, Nijmegen and Leuven
- Groningen: John Nerbonne (co P.I.), Sandrien van Ommen, Sebastian Kürschner, Charlotte Gooskens, Renée van Bezooijen
- Nijmegen: Roeland van Hout, Stef Grondelaers, Mieke Steegs
- Leuven: Dirk Geeraerts (co P.I.), Dirk Speelman, Leen Impe
Mutual Comprehensibility of Dutch Dialects

Key questions:

- How well do Dutch speakers from different regions understand each other?
- Are lexical or pronunciational differences more the important predictors of comprehensibility?
- This paper is a “spin-off,” looking at how “regional speech,” fits into the spectrum of varieties.
Overview

- Measuring sociolinguistic differences
  - Comparison to dialectology
  - Measuring pronunciation differences
  - Validating measures

- Regiolects between standard language and base dialects
  - Auer & Hinskens (1996) “cone model”
  - Distance metaphor

- Data
- Results
- Reflections on this undertaking
Why dialectometry?

- Strengthen geographic signals by aggregating
- Solve problems of earlier dialectology
  - Deal with massive variation
  - Non-overlapping distributions
  - Selection of features too arbitrary
  - “Atomism” (Coseriu), idiosyncratic words (Bloomfield)
- Introduce replicable procedures
- Seeking law-like relations in linguistic variation (Seguy’s sub-linear curve)

J. Nerbonne “Data-Driven Dialectology” *Language and Linguistic Compass* 3(1), 2009. 175-198
The current consensus is that social and geographical variation should be studied using the same linguistic methods.


“Social Dialectology”

Interpretations (extra-linguistic correlates of variation) are of course different.

Probably uncontroversial, included for sake of explicitness.
Calculating varietal distances

- To determine the aggregate distance between varieties:
  - For every single linguistic element (in sample), we determine the distance between the realizations of the element in each pair of varieties.
    - Perhaps just same (0) vs. different (1)
    - ... but we’ve developed more sensitive measures (below)
  - We sum these distances for every element (hundreds of them)
  - Immediate result: place × place table of varietal differences

- Séguy (1971), Goebl (1980s and on), many others
Chambers & Trudgill (1998) ask for a ranking of features (and isoglosses) in order to identify dialect boundaries.

Implicit “feature ranking” in dialectometry: a feature that’s instantiated $n$ times in dialect atlas material is weighted $n$ times more heavily than one that appears once.

- Lexical items uniformly weighted
- Phonetic segment distances weighted in proportion to their frequency in the word list
- More sophisticated options possible (Wieling, this workshop)

Note that Goebl has also experimented with “inverse frequency” weighting of responses.
Levenshtein distance enables analysis of phonetic transcriptions without manual alignment —move from categorical to numerical analysis of data.

One of the most successful methods to determine sequence distance (Levenshtein, 1964)
- biological molecules, software engineering, ...

Levenshtein distance: minimum number of insertions, deletions and substitutions to transform one string into the other
Syllabicity constraint add: vowels never substitute for consonants
Example of the Levenshtein distance

\[
\begin{array}{c|c|c}
\text{mælkə} & \text{delete ə} & 1 \\
\text{mɔlkə} & \text{subst. ɔ/ɛ} & 1 \\
\text{mɛlkə} & \text{delete ə} & 1 \\
\text{mɛlk} & \text{insert ə} & 1 \\
\text{mɛlək} & & 4
\end{array}
\]

\[
\begin{array}{c|c|c|c|c}
n & 1 & 1 & 1 & 1 \\
m & m & c & l & k & e \\
\text{m ɛ l ə k} & & & & & 1
\end{array}
\]
Levenshtein validly measures pronunciation difference

- Measures *must* be validated, shown to measure what they claim to. Hard problem, in general.
- Gooskens and Heeringa (2004) show that aggregate Levenshtein distances correlate strongly with Norwegian dialect speakers judgments of “how differently people speak” ($r = 0.67$)
- Sanders & Chin (2009) show that aggregate feature-based Levenshtein distances correlate very strongly with naive listeners’ judgments of the degree to which the speech of cochlear implant users differs from normal speakers’ ($r = 0.92$).
- Naturally, further validation efforts are welcome, especially concerning different (sub-)fields of inquiry.
Greenhill challenges the use of Levenshtein distance as a method of detecting genealogical relations among languages. *Computational Linguistics, to appear*

Kessler, Wichman and others have used Levenshtein distance for this purpose.

But Greenhill’s results may just confirm the standard wisdom in historical linguistics that genealogical relatedness is shown not by similarity, but rather by *shared innovation*. (Campbell, 2004)

Note that the Levenshtein algorithm may still be of diachronic service for its alignments (Prokić 2010).
Regiolects

What is happening to local variation?

- Nearby varieties remain similar to accommodate, to maintain communication
- In the modern DIAGLOSSIC situation, virtually all dialect speakers are competent in the standard, as well.
- This gives rise to regional speech, which is more standard-like than the traditional base dialects, but which still shows signs of local affinity.
- Auer and Hinskens (1996) provide a cone model.
Auer & Hinskens’ Cone
Auer & Hinskens’ Cone

- Geometrical model of relations among varieties cries out for measures!
- “Regiolects” intermediate between standard and basilects, represent choices closer to standard, etc.
- It’s our task to engage the sociolinguists, however.
Regions

standard (spoken/written)

regional standards

regiolects

base dialects
Data

- Dutch pronunciation data from the Goeman-Taeldeman-Van Reenen-Project (GTRP; Goeman and Taeldeman, 1996)
  - We use 37 words for relevant basilectal varieties in the Netherlands
- Additionally, eight announcers from regional radio stations were recorded and transcribed.
  - More needed to say something with more certainty.
  - But “regional speech” is difficult to operationalize
  - And professional announcers expect to be paid!
- Finally, we include both Dutch and Flemish standard pronunciations, aka Netherlandic and Belgian Dutch.
Regions

Mix of central, peripheral regions.
Hypotheses

1. Is the regiolect a faithful representative of its region, i.e. closer to its own basilects than any other regiolect?
2. Is the regional speech closer to the basilects than the standard is?
3. Is the regional speech intermediate between the standard and the basilects (thereby facilitating communication)?

The last two test the degree to which the regional speech conforms to the conical model.
Regional speech faithful to region?

In fact, all the samples of regional speech reflect the speech of their region more than the speech of any other region. The central, *Randstad* was (insignificantly) closer to the speech of Dutch Brabant.
Regional speech closer to basilects than standard is?

Netherlands: where left box plot higher than right, then regional speech is closer to the standard — only in Limburg and Groningen.
Flanders: where left box plot higher than right, then regional speech is closer to the standard — only in West Flanders.
Where left box plot above dashed line, then regional speech is closer to standard than basilects are — everywhere except in Groningen.
Flem. regional speech between basilects & standard?

Where left box plot above dashed line, then regional speech is closer to standard than basilects are — only in Brabant and West Flanders.
Summary of results

1. Is the regiolect a faithful representative of its region, i.e. closer to its own basilects than any other regiolect?
   —yes

2. Is the regional speech closer to the basilects than the standard is?
   —Three of eight are closer.

3. Is the regional speech intermediate between the standard and the basilects (thereby facilitating communication)?
   —Five of eight are intermediate.

In fact, only one regional sample is both closer to the standard and genuinely intermediate, the West Flanders sample.
Regional Speech

- Background
- Variationist measures
- Regiolects
- Data
- Results
- Reflections

Regional Speech Diagram:

- Randstad
- North Brabant
- BE Brabant
- Standard
- Antwerp
- Regiolects
- West Flanders
- Basilect
- Groningen
- NL Limburg

Authors:
John Nerbonne, Sandrien van Ommen, Charlotte Gooskens & Martijn Wieling
j.nerbonne@rug.nl
Performing regionally?

- Our results should not be taken to challenge the conical model, as our sample is neither large nor representative.
  - Still, surprising that regional announcers speak as they do.
  - The conical model correctly identifies the standard as influencing regional speech.
- Our announcers are professionals, and yet they do not strike a balance intermediate between basilects and the standard.
- Either this is difficult, or they are not trying to strike such a balance.
  - Two (Groningen and Dutch Limburg) overemphasize local elements.
- Sociolinguists also study how identity is projected through choices in language behavior (Eckert)
  - They speak of performing according to a linguistic repertoire.
Since sociolinguistic variation involves the same linguistic elements found in dialectology, we expect to transfer measurement techniques.

This work focused on a sociolinguistic topic close to dialectology, but that is not essential.

Some sociolinguistic work appears to cry out for measurement, as it is expressed in explicit terms of distances.

Sociolinguists are not, in general, immediately convinced/aware of the value of measurement.

Lots more linguistic subfields might benefit from measurement, e.g. second-language learning, language pathology (aphasiology, specific language disorders), stylometry, ...
Questions?

Thank You!

Try Gabmap! www.gabmap.nl