Measuring Socially Motivated Pronunciation Differences

John Nerbonne,

j.nerbonne@rug.nl

Center for Language and Cognition, University of Groningen

FRIAS Colloquium, Universität Freiburg
Jan. 16, 2013
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, Sandrien van Ommen...

... 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.
“Dialectometry” is now fairly successful: used by those researching geographic variation.

But geographic and social variation are theoretically closely related (Chambers-Trudgill fusion). So dialectometric methods should make sense for social variation, too.

Either “lectometric” techniques make sense for use in the study of both sorts of variation, or they make sense in neither.

We discuss one exceptional perspective below.
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
Setting our sights on social variation

- 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.

Note that some sociolinguists are primarily interested in catching single phoneme changes *in statu nascendi*. Dialectometric aggregating makes no sense from this perspective.
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 $\times$ place table of varietal differences

- Séguy (1971), Goebl (1980s and on), many others
A natural “feature ranking”

- 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, Margarethe & Nerbonne, *J.Phonetics* 2012)
- Note that Goebl has also experimented with “inverse frequency” weighting of responses.
A more sensitive pronunciation distance measure

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

We’ll apply this pronunciation difference measure to the question of how regiolects are constituted, but first two highlights from earlier work.
Dutch Dialect Data (also used in regiolectal study)

- Dutch pronunciation data from Goeman-Taeldeman-Van Reenen-Project data (GTRP; Goeman and Taeldeman, 1996)
  - We use 562 words for 424 varieties in the Netherlands

- Calculating Levenshtein distances yields interesting sound correspondences
  - Note that a 100-word comparison already yields about 500 sound correspondences; 562 words imply 2810.
Distribution of sites
Align each of the $\sim 90,000$ pairs of varieties $= \binom{424}{2}$

- involving $2810$ segment correspondences each
- $\approx 2.5 \times 10^8$ segment correspondences in total

For dialect geography we’re primarily interested in the site $\times$ site matrix of aggregate distances, which we analyze further.

Here we’re interested in a comparison to socially influenced varieties.

But first, the two highlights of earlier work.
Dialect Geography: MDS of aggregate distances

w. first 3 MDS dimensions → blue, green and red → map.
Seguy’s Curve: A Dialectological Law?
Levenshtein validly measures pronunciation difference

(\textit{end of excursus})

- Measures \textit{must} be validated, shown to do what they claim.
- Gooskens and Heeringa (2004) show that aggregate Lev. distances correlate well ($r = 0.67$) with Norwegian dialect speakers’ judgments of “how differently people speak”.
- Aggregate feature-based Lev. distances correlate very strongly ($r = 0.92$) with naive listeners’ judgments of the deviance of cochlear implant users’ speech (Sanders & Chin 2009).
- Strong correlation ($r \approx 0.8$) with judgments of foreign accents wrt Am. English (ongoing work).
- Naturally, further validation efforts are welcome, especially concerning different (sub-)fields of inquiry.
Greenhill’s paper


- 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 DIGLOSSIC 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

- Very influential model!
- Very plausible in situations with a dominant standard language, compulsory schooling, mobile population.
- “Regiolects” intermediate between standard and basilects, represent choices closer to standard, etc.
- Geometrical model of relations among varieties cries out for measures!
- It’s our task to engage the sociolinguists, however.
Note that claims aren’t about distances directly, but rather about intermediate positions of regiolects wrt standard and base dialects. Since we measure distances, we need to be careful (below).
Data

- GTRP described above
  - 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.
Checking intermediate position via distances

There’s no simple check on intermediacy using distances. We need to test intermediacy in two steps.

1. Is regional speech closer to the basilects than the standard is?
   - Still allows regional identification

2. Is regional speech closer to the standard than the basilects are?
   - “Easier” than standard, better for interregional communication
$d(B, R) \ll d(B, S)$?

Is the regiolect closer to the basilects than the standard is?

John Nerbonne, j.nerbonne@rug.nl
\[ d(S, R) \ll d(S, B) ? \]

Is the regiolect closer to the standard than the basilects are?
Intermediacy (Model-Conforming Regiolects)
Hypotheses

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

The last two test the degree to which the regional speech conforms to the conical model.
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.

John Nerbonne, j.nerbonne@rug.nl
Regional speech closer to basilects than standard is?

Flanders: where left box plot higher than right, then regional speech is closer to the standard — only in West Flanders.

John Nerbonne, j.nerbonne@rug.nl
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.

John Nerbonne, j.nerbonne@rug.nl
Regional speech faithful to region?

- (One example) Each col. (except leftmost) shows distances from base dialects to selected regiolect (here Randstad). If regiolect is faithful, we expect the region’s own base dialects to be lowest.

- 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.
Summary of results

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

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

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

In fact, only one regional sample is both closer to the standard and genuinely intermediate, the West Flanders sample. If we suppose that there’s an 0.5 of a regiolect being intermediate, then the chance of seeing only one (or fewer) intermediate regiolects in a group of eight is approx. 0.036 (binomial calculation).
Regional Speech

Diagram:
- **Randstad**
  - North Brabant
  - BE Brabant
  - Standard
  - BE Limburg
  - Antwerp
  - Regiolects
  - West Flanders
  - Basilects
  - NL Limburg
  - Groningen
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.
“Regiolects” are not like base dialects in that they aren’t acquired by anyone natively or used by anyone exclusively.

- Auer: “the implication that we are dealing with a variety is not necessarily justified” (2005) See Hoppenbrouwers, too.
- RL’s result when dialect speakers aim to make their speech more accessible to speakers of neighboring dialects (our announcers), or to speakers of other regions or standard speakers.
- But lots of modern speakers of RL’s are more than capable of shifting completely to standard (all of the radio speakers). They often retain local features by choice.
- Note that I haven’t emphasized the effects of external speakers (Thomason’s “shifters”) adopting (some dialect)
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