Challenges in Finding the Determinants of Dialect Variation

John Nerbonne and Hermann Niebaum

Rijksuniversiteit Groningen

Kick-off Meeting, Determinants of Dialect Variation

Groningen, Jan. 9, 2004

Determinants of Dialect Variation (*DDV*)

- Point of Departure
- Goals
- Special Challenges
- Schedule
- NWO project 2003-2007
  - Project Team
  - Institutions

DDV Point of Departure — General

- Dialectometry is successful
  - Jean Séguy, initiator, 1971
    attention to aggregate dialect differences
    data viewed categorically—same or different
    attention to pronuciation, lexis, morphology, syntax
  - Hans Goebl, canonizer
    broadened linguistic application
    deepened connections to numerical taxonomy, cartography
    importance of unlikely overlap: “G.I.W.” ‘weighted similarity’

DDV Point of Departure — Local

- Levenshtein Measure of Pronunciation Difference
  - Wilbert Heeringa, Diss. 2004
    pronunciation difference measured numerically
    attention to reliability, validity
  - Wilbert Heeringa and John Nerbonne, 2002
    geographic explanation of dialect difference
    heralded by Seguy, 1971

- Syntactic Atlas of the Netherlandic Dialects
  - Hans Bennis, Hans den Besten, Johan Rooryck, Johan Van de Auwera, and Magda Devos
    first compendium of Dutch syntactic variation
    one of the earliest syntactic atlases anywhere

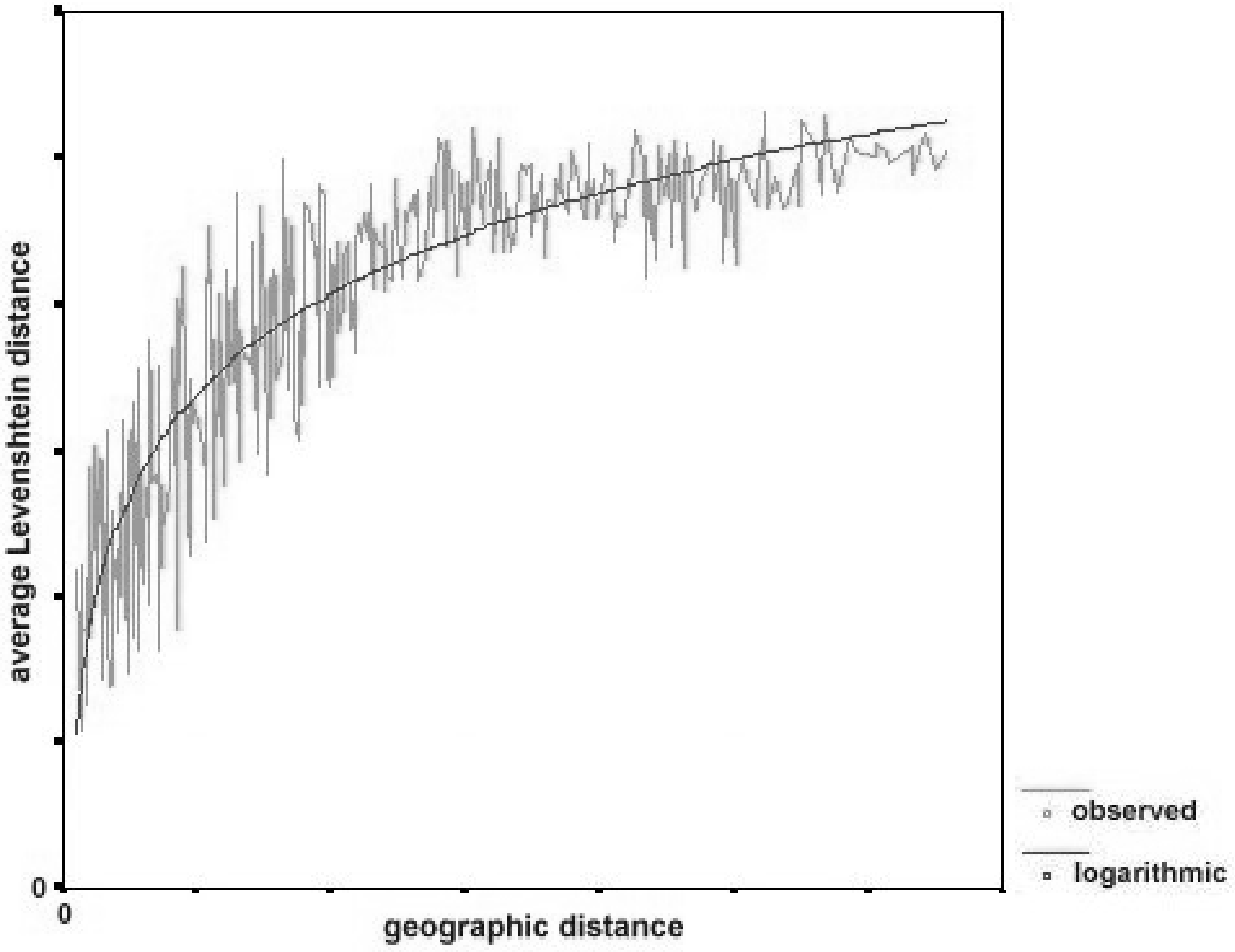
DDV Goals

- Validate pronunciation approach on German
  - test with same parameters, representation type
  - analyse sample—choice of sites
  - compare dialects to standard German
    —compare to Herrgen-Schmidt *Dialektaliät*
  - collaborate on analysis of determinants

- Apply nominal measure to Dutch syntax
  - see Nerbonne & Kleiweg for example of nominal measure
  - check geographic cohesion of syntax variables
  - check systemic affinities (θ: parametric variation)
    idea: see Agrawal on discovering associations in databases
  - collaborate on analysis of determinants

DDV Goals

- Analyze determinants: what predicts linguistic distance?
- **idea:** regression analysis with linguistic distance as dependent variable
- Investigate independent variables:
  - geographic distance, travel distance (and/or trade routes, pilgrim routes?)
    - water vs. land
  - population size (product of populations)
    —together with geography yields Trudgill's "gravity model"
  - tribal history
  - other? trade? on route?


average Levenshtein distance
0
0
geographic distance
observed
logarithmic

DDV Goals

- Establish a dialectometric package, Kleiweg's "Lev"
- Encourage experimentation with techniques
- Internal use, tutorial presentation
  - Compare to Goebl's package
- Models: Rumelhart & McClelland's PDP, Sankoff's VARBRULE, Shieber's PATR

DDV Challenges/Opportunities

- How well do linguistic levels correlate?
  - pronunciation, lexis, syntax, ?morphology?
  - do some theoretical views imply high correlation?
  - LAMSAS pronunciation, lexis correlate $r = 0.65$
- Effective check on systemic affinities (θ: parametric variation) would open door to data-intensive language typology.
  - Typology DB's in Amsterdam, Leipzig,
- Can we contrast similarity (dialectological measure) with degree of shared innovation (historical measure)?
  - Wang on regularity of sound change

Levenshtein Sequence Comparison

Basic idea: Distance between two sequences is the sum of the "costs" needed to rewrite one string into another

kœstə	delete ə	1
kœst	replace œ by ɔ	2
kɔst	insert r	1
kɔrst		
		4

Lots of (sets of) operations can transform one string into another.

Definition: Levenshtein distance between two sequences $L(s_1, s_2)$ is the minimal cost needed to rewrite s_1 into s_2.

Algorithm notes correspondences

Levenshtein distance(Lat.*pater*, Germ. *faðir*)

		f	a	ð	i	r
	0	1	2	3	4	5
p	1	0.4	1.4	2.4	3.4	4.4
a	2	1.4	0.4	1.4	2.4	
t	3	2.4	1.4	0.9	1.9	
e	4	3.4	2.4		2.1	
r	5	4.4				2.1

identifies correspondences in individual word pairs p:f, t:ð, e:i

historical linguistics requires *regular* correspondences, e.g., p:f in *pisces:fish*, *plenum:full*, *primus:first* as evidence of relatedness

Challenge

- Lift the identification of individual correspondences
- Obtain identification of *regular* correspondences
  - how often is “regular”?
  - what to do with near-correspondences?
  - how to incorporate effect of phonetic environment?
- Enable studies in language contact and historical linguistics

DDV Project Team 10/2003–9/2007

- Staff:
  - Wilbert Heeringa, postdoc, Groningen
    background: Dialectometry
  - Christine Siedle, graduate student, Groningen
    background: phonetics, computational linguistics
  - Marco Spruit, graduate student, Meertens
    background: computational linguistics

- Supervision
  - Sjef Barbiers, Syntactic Theory, Meertens
  - Hans Bennis, Syntactic Theory, Meertens
  - John Nerbonne, Computational Linguistics, P.I., Groningen
  - Hermann Niebaum, Dialectology, Groningen

- Advisors: t.b.a.