Linguistic Challenges for Computationalists

John Nerbonne

University of Groningen

Sept. 23, 2005

Recent Advances in Natural Language Processing

Borovetz

RuG

Structure of Talk

Thesis: CL is poised to contribute to Linguistics

- Some preliminaries
- Science and Engineering in CL
- Dialectology
- Diachronic Linguistics
- Language Acquisition
- Language Contact
- Other areas, conclusions

RuG

CL poised to contribute to Linguistics

Some Preliminaries

- Argument by way of convincing examples
  - areas with ongoing contributions
  - concrete examples potentially stimulating
- Avoid well-known examples such as grammatical theory
- No attempt at comprehensive list (for reasons of time, information)
- No attempt at comprehensive descriptions
- No plea to ignore (practical) applications!

Familiar Characterization (Joshi, Kay, Shieber)

- Science (theory)
  - Language, grammar & automata classes; parsers, transducers, ...
- Engineering
  - translation, lexicography, speech understanding, foreign language tools & instructions, dictionary & theosaurus strcuture & access, IR (incl. term extraction, summarization, text mining, & question answering), information systems, grammar checking, controlled language, handicapped aids, ...
- Large software infrastructure useful in science and engineering!

CL *applied* **to science**

Driven by curiosity, not practical gains.

- Genetics applies biochemical techniques ...
- Archaelogy applies radiochemistry (carbon dating)
- Astronomy applies optics, electromagnetics (radio)
- ...

- X applies computational linguistics
- X $\stackrel{?}{=}$ dialectology, diachronic linguistics, language acquisition, language contact, ...

RuG

Measuring Segment Differences

- Phonetics, CL shows how to measure differences in segments, e.g. as city-block distance in *features*

  Example: difference ([i], [e]) much smaller than difference ([i], [u]).

	i	e	u	i-e	i-u
advancement	2(front)	2(front)	6(back)	0	4
high	4(high)	3(mid high)	4(high)	1	0
long	3(short)	3(short)	3(short)	0	0
rounded	0(not rounded)	0(not rounded)	1(rounded)	0	1
				1	5

- Diacritics [ĩ,e:,ə^{r}] can also be taken into account
- Vieregge-Cucchiarini system used, also Almeida-Braun
- Chomsky-Halle (SPE) system less useful (clever features for making rules compact)

Levenshtein Distance

Idea: *lift* segment distance to sequence distance.

Standard American	sɔəgIrl	delete r	0.5
	sɔəgIl	replace I/ɜ	0.1
	sɔəgɜl	insert r	0.8
Bostonian	sɔrəgɜl		
		Sum distance	1.4

- L-distance $=^{\text{df}}$ *minimal cost* of operation to rewrite one string to another.
- Insertions and deletions compare segment to silence

Levenshtein Distance aka edit distance, string distance also used in CL (bilinguial alignment), bioinformatics, software engineering.

`http://www.let.rug.nl/~kleiweg/lev/`

RuG

Dynamic Programming Algorithm

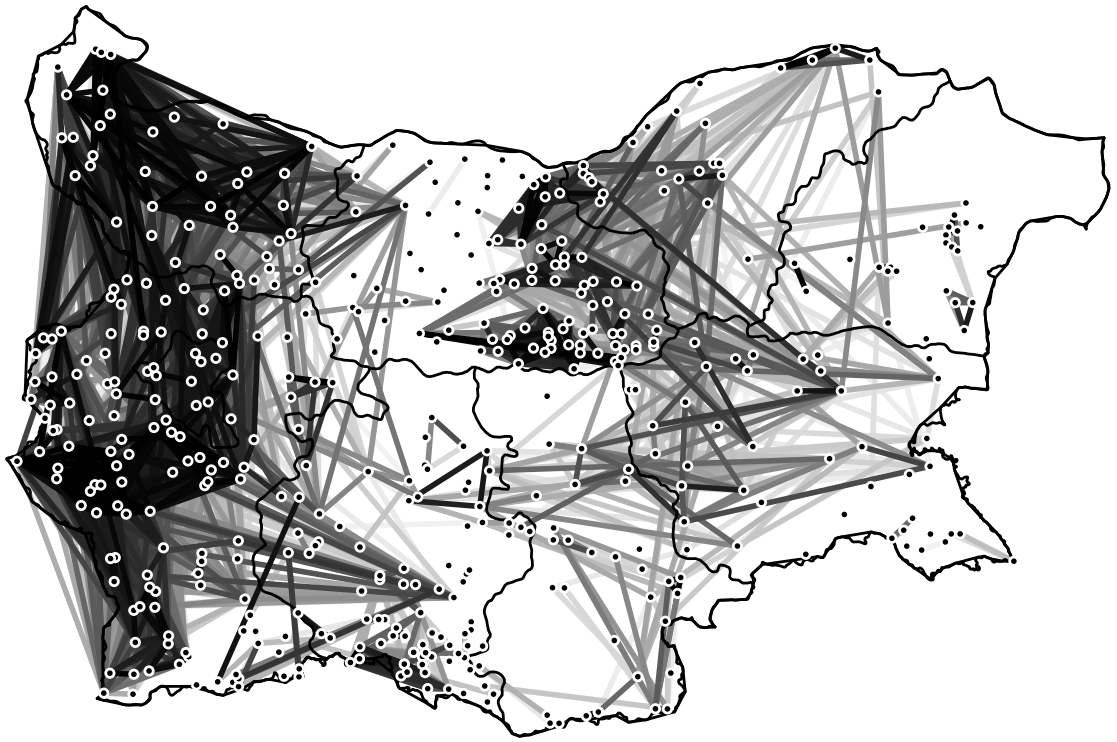
Levenshtein-distance(*adresse*,*address*)

		a	d	d	r	e	s	s
	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4		
d	2	1	0	1	2			
r	3	2	1	2	1			
e	4	3	2			1		
s	5	4					1	
s	6							1
e	7							2

- Use 100-word sample in large number of varieties
- dialect distance is equal to the sum of the word distances (distance is *additive*)
- Kessler, EACL ’95 — application to Irish dialectology

RuG

Average Distance Between Bulgarian Varieties



Collaboration with Petja Osenova (LML, Sofia) & Wilbert Heeringa (Groningen)

RuG

Further Applications of CL Techniques

- Lexical analyser for parsing phonetic transcriptions
- Lemmatizing word forms as step in measuring lexical variation
  fair off, fairing, fairing off, faired off, fairs off, ...
- Lifting edit distance from strings to sets of strings (to measures differences involving multiple responses).
- Assessing measurements (consistency, validity vis-à-vis dialect speakers'perceptions)
- Exploratory statistics (clustering)
- (Inverse) frequency-based weighting

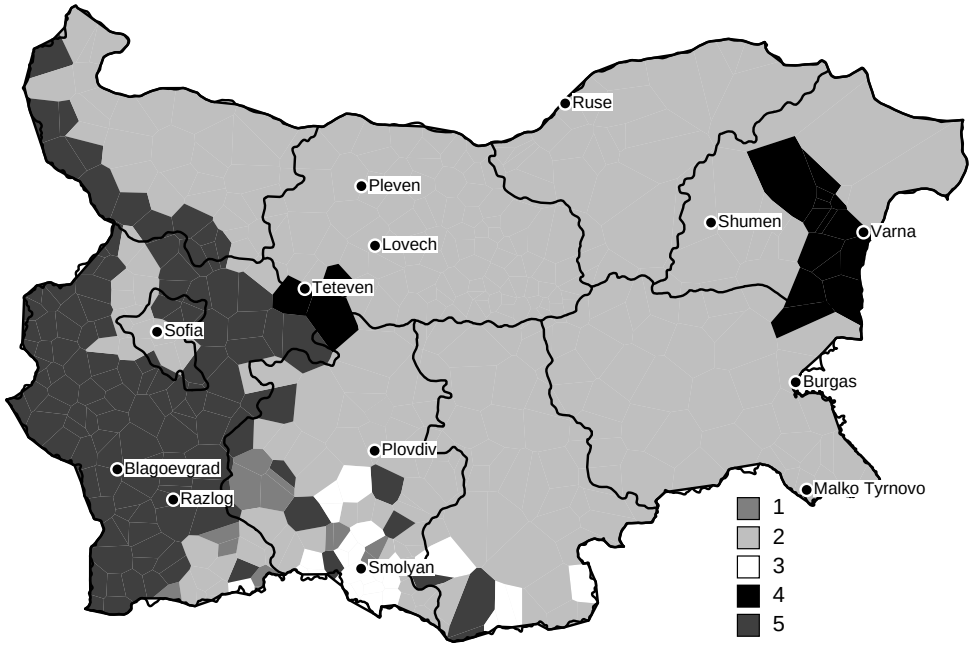
Novel Steps in Dialectology

- Areas vs. continua as organizing principle
- Convergence and divergence
- Validation versus dialect speakers' intuitions
- Quantify importance of geography ($\approx$ 15-60% variance)
- Dynamics (Gravity Hypothesis)

RuG

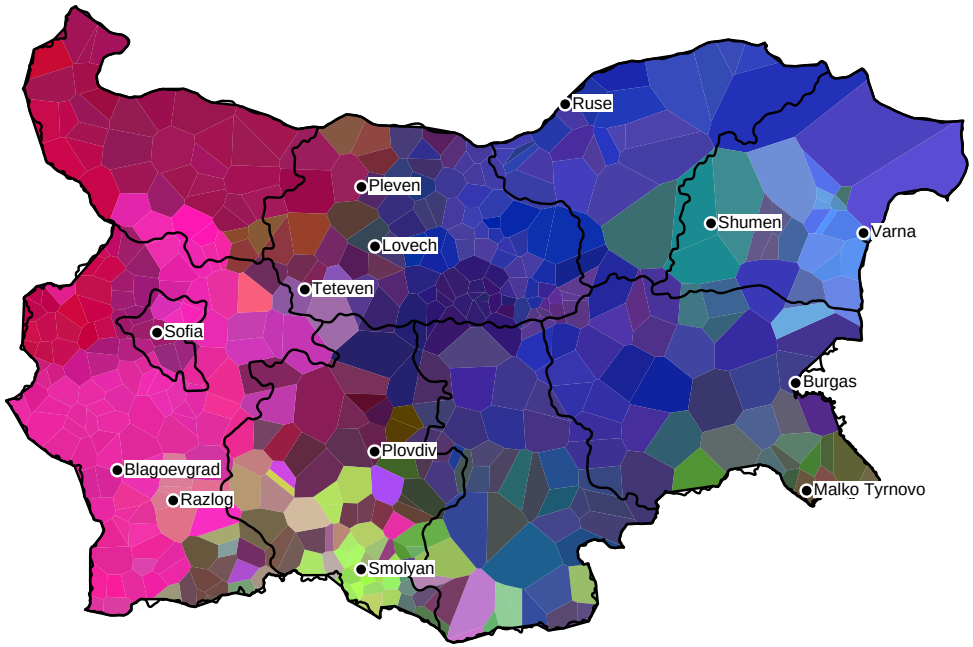
Areas

- Via clustering (weighted average)



Continuum?

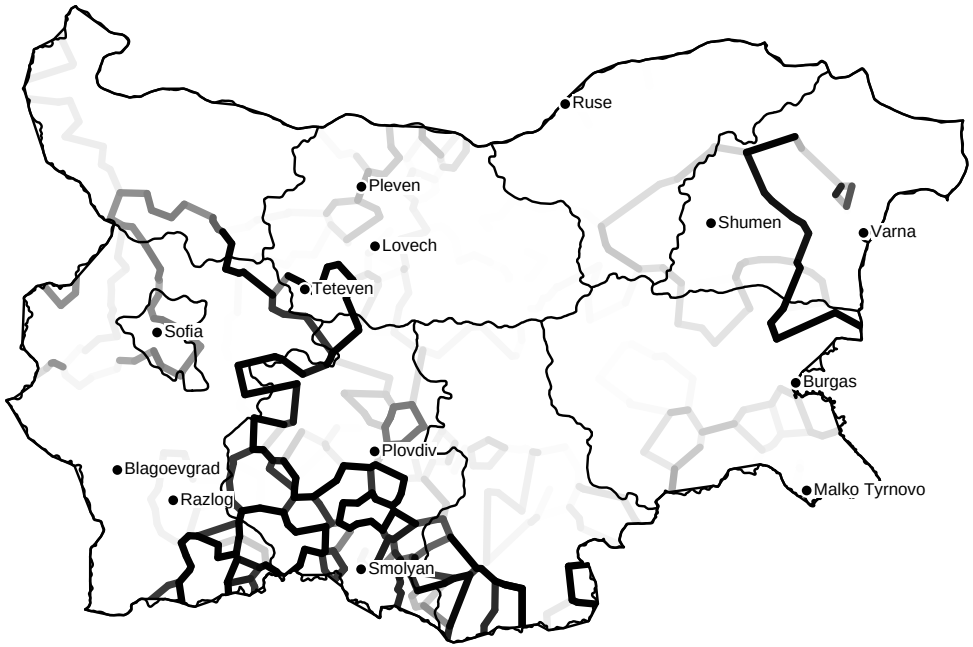
- Multi-Dimensional Scaling: given distances, ideal coordinates can be inferred.
- From $\binom{n}{2}$ distances we infer 3-dim. coordinates accounting for 95% of variance.



RuG

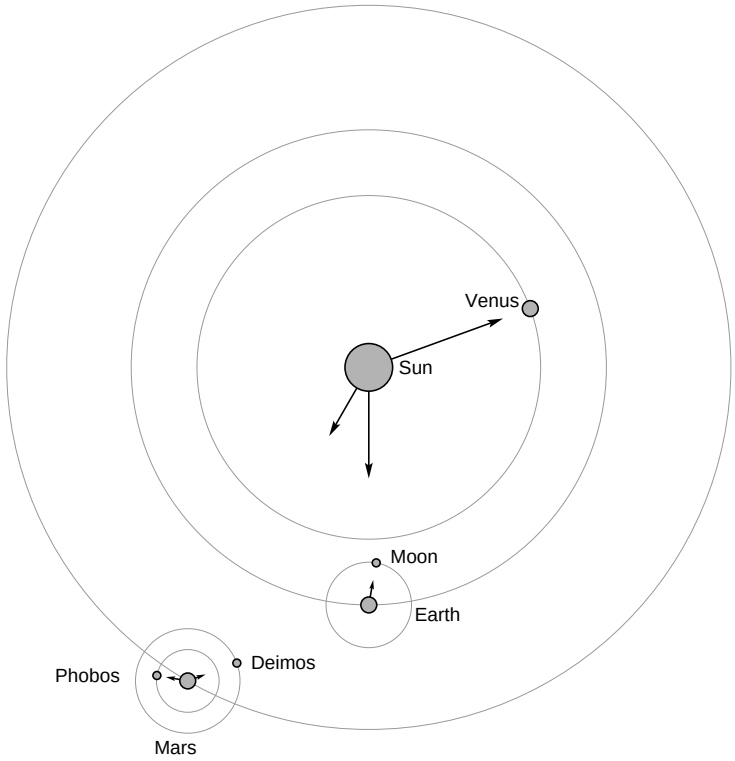
Visual Reconciliation

- Repeated clustering with noise



RuG

Gravity Hypothesis of Linguistic Diffusion



RuG

Linguistic Cohesion via Gravity

$$F = G\frac{m_1 m_2}{r^2}$$

F is the attractive force,

m_1, m_2 the population masses of the two settlements,

r the distance between them, and

G won't be speculated on

Idea: social contact promotes linguistic accommodation and linguistic similarity.

Detecting Effects of Linguistic Gravity

$$F = G\frac{p_1 p_2}{r^2} = 1/D$$

$$D \propto 1/G\frac{r^2}{p_1 p_2}$$

F is ling. attraction, which should produce similarity

D is ling. dissimilarity

p_1, p_2 the population masses of the two settlements, and

r the distance between them

RuG

Linguistic Cohesion via Gravity

$$D \propto 1/G\frac{r^2}{p_1 p_2} \propto \frac{r^2}{p_1 p_2}$$

$$D \propto r^2 \text{ AND } D \propto -p_1 p_2$$

D is linguistic distance,

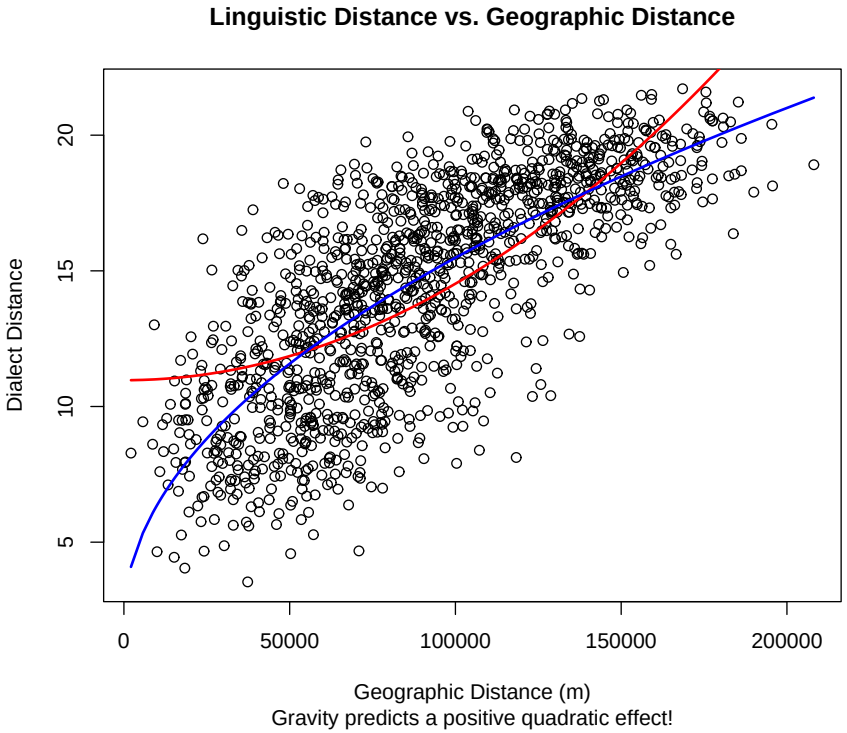
p_1, p_2 the populations of the two settlements, and

r the distance between them

Notate bene: we measure linguistic dissimilarity, which we postulate stands in inverse relation to the attractive force of social contact.

RuG

Function of $\sqrt{x}$?



Quadratic vs. root: Shape? Zero? ($r^2 = 0.57$ for root)

RuG

Conclusions on Dialectology and CL

- Levenshtein distance measures dialectal pronunciation differences reliably, validly
- Aggregations (sums/averages) of linguistic distances characterize entire varieties.
- Dialect continua and dialect areas may be characterized from one theoretical perspective.
- New questions are enabled: quantifying effect of geography, etc.

- CL technique is foundation of Levenshtein measure of pronunciation difference
- Many other CL techniques turn out to be useful
  - Lemmatizing/stemming, inverse frequency weighting, regular expression grammars for transcriptions, clustering, problems of evaluation/assessment, ...

Diachronic Linguistics: Regular Correspondences

- Historical linguistics notes parallel pronunciations in cognates

Latin	p	a	t	e	r	
Greek	p	a	t	e	r	a
Engl.	f	a	ð	e	r	
Indic	p	i	t	ā		
Irish		a	h	i	r	

  which Kondrak (2002) systemizes and aligns via a variant of Levenshtein
- Tricky issue: avoid false cognates Eng. ‘have,’ Lat. *habere*
- Solution: focus on regular correspondences, e.g. /p:f/ (Eng. ‘fish’, Lat. *pisces*; Eng. ‘full’, Lat. *plenus*)
- Computational puzzle: how to identify global regularity?

RuG

MT alignment & Regular Sound Correspondences

Kondrak (2002) notes parallel

- MT aligns parallel sentences, looking for regular lexical correspondence (in order to identify translation equivalents). See Tiedemann (1999, 2003)
- Diachronic linguistics aligns cognate words, looking for regular segmental correspondence (in order to identify sound equivalences)

- In both cases, one needs to generalize from local alignments to global ones
- Kondrak applies Melamed's ideas on identifying translation equivalences to the problem of obtaining sound correspondences
- Kondrak tests algorithm on Bloomfield's Algonquian data with precision and recall near 90%

RuG

More: CL & Diachronic Linguistics

- Kessler measures statistical significance of regular correspondences using permutation measures
- In spotting cognates, Kondrak enlists WordNet as a means of quantifying the semantic overlap one would like to see in cognates. He concludes that its contribution is minor. Can the recognition of semantically related words be improved?
- Can alignment be made more sensitive to phonetic conditioning?
- Can models for identifying correspondences be generalized to dozens, or even hundreds of related varieties?
- Can borrowings be identified along with cognates?
- Why is computational biology (**PHYLIP** by Felsenstein) the most popular source of ideas (see Gray & Atkinson, *Nature*, 2003)?

CL & Language Acquisition

- Lg. Acquisition “central problem of linguistic theory”
- Huge interest in machine learning techniques in CL

- Obvious match?

RuG

CL & Language Acquisition

- Pioneer work in mid 90's by Michael Brent
- Asked whether child-directed speech (corpora available) allow segmentation

- Words in /dɔginaɨsdɔgiwʌrənaɨsdɔgi/ ?

- Idea: use phonotactics (beginnings and endings of utterances), and minimize sum:

  number of tokens in experience
  number of types in lexicon
  length of word types postulated
  entropies of word types

- Link lg. acquisition and minimal description length learning!

RuG

CL & Language Acquisition

- /dɔginaɬsdɔgiwʌrənaɬsdɔgi/ → ‘Doggie. Nice doggie. What a nice doggie!’

- Tjong Kim Sang, Stoianov, Konstantopoulos (Groningen) studies of phonotactics (what syllables occur?)
- /vstrɛtʃ/ OK in Russian, not in Dutch
  —How is this learned?
- Rule-based techniques compact, statistics required to separate well-formed from ill-formed effectively

CL & Language Acquisition: Growing Interest

- *Psycho-Computational Models of Human Language Acquisition*
  COLING ’04, ACL ’05
- Computational simulations operationalize innateness assumptions (in bias).
- Interest on the part of linguistic theory (Albright & Hayes)
- Linguistic focus on error profile, differentiating among material not in experience. Testing on possible vs. less possible forms.
- Huge horizon of unsolved problems, gradually coming within range.

CL & Language Contact

- Borrowing of words, sounds, structures, ...
- Mixing in koinés, pidgins, creoles, dialect leveling
- Area growing in interest, perhaps due to interest in cultural contact and mixing.
- Linking to (imperfect) second-language acquisition
- Data situation: corpora available, no systematic “atlases”
  —techniques from dialectology of limited use

Measure of Syntactic Infection

- Idea: take two corpora, one candidate for “infection”
- Tag both corpora with smallish tag set, collect POS-trigrams into vector, measure histogram difference, assess significance via permutation test
- Hypothesis: infection will be reflected in degree of deviation
- Expected result: numerical *measure* of deviation
- Preliminary result (with Wiersema (Groningen), Opas Hänninen, Lauttamus, and Hirvonen (Oulu)): we can show speech of late immigrants to deviate very significantly from speech of child immigrants.
- Still need (automated) techniques to attribute sources of deviation

Other Linguistic Topics

- Grammar—well-known, established, but limited in theoretical impact
- Psycholingistic processing—earlier center of attention (psychological parsing, dis-ambiuation), but perhaps worthy of revival.
- ?? (question for discussion)

Aside: Can Engineering Illuminate Science?

- Nerbonne & Kleiweg use a Porter stemmer to identify forms of different lexemes (to detect lexical overlap in dialectology).
- Kondrak adopts Melamed's work on identifying translation equivalences to the problem of finding regular sound correspondences in historical linguistics.
- Several learning experiments apply ML techniques to child-direct speech to demonstrate that input data contains sufficient information to support learning (with specific biases).

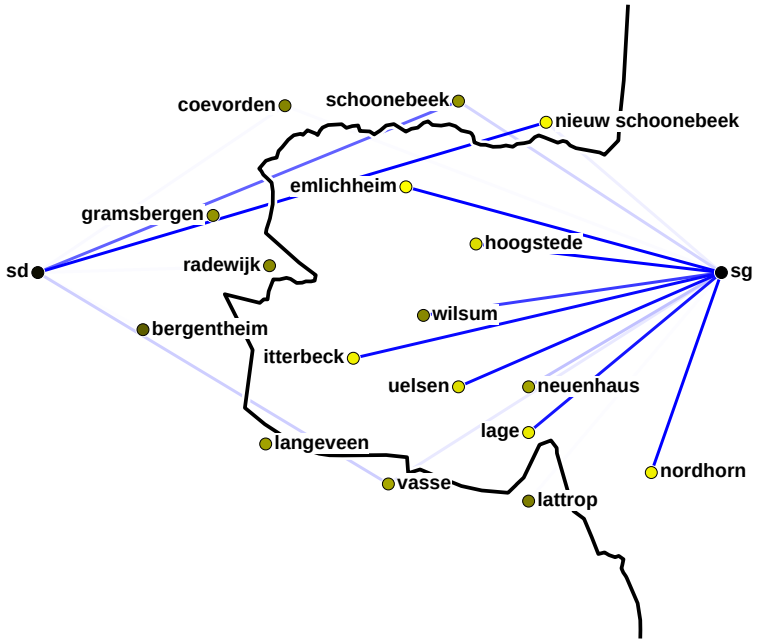
Computation Illuminates Linguistics

- Well-known opportunities for CL contributions
  - Grammars (CG, HPSG, LFG)
  - Psycholinguistics (Crocker, Kempen, ...)

- Emerging opportunities for CL contributions
  - Dialectology (own work, Heeringa 2004 *et passim*)
  - Optimality Theory (Karttunen, van Noord & Gerdemann, Eisner, ...)
  - Language Acquisition (Brent, 1997 et seq.)
  - Historical Linguistics (Kondrak 2002)
  - Language Contact (potential)

Application: Borders & Standards

Heeringa et al. 2000: Divergence Dutch-German border in Bentheim, 1974-2000



Blue convergence toward standard Dutch (sd) vs. standard German (sg).

RuG

RuG