Modeling Scandinavian Semi-Communication

The Conditional Entropy of the Phoneme Mapping

John Nerbonne[1]

[1]Alfa-informatica
University of Groningen

Seminar für Sprachwissenschaft
Alumni Meeting,
Tübingen, June 21, 2007

Collaborators

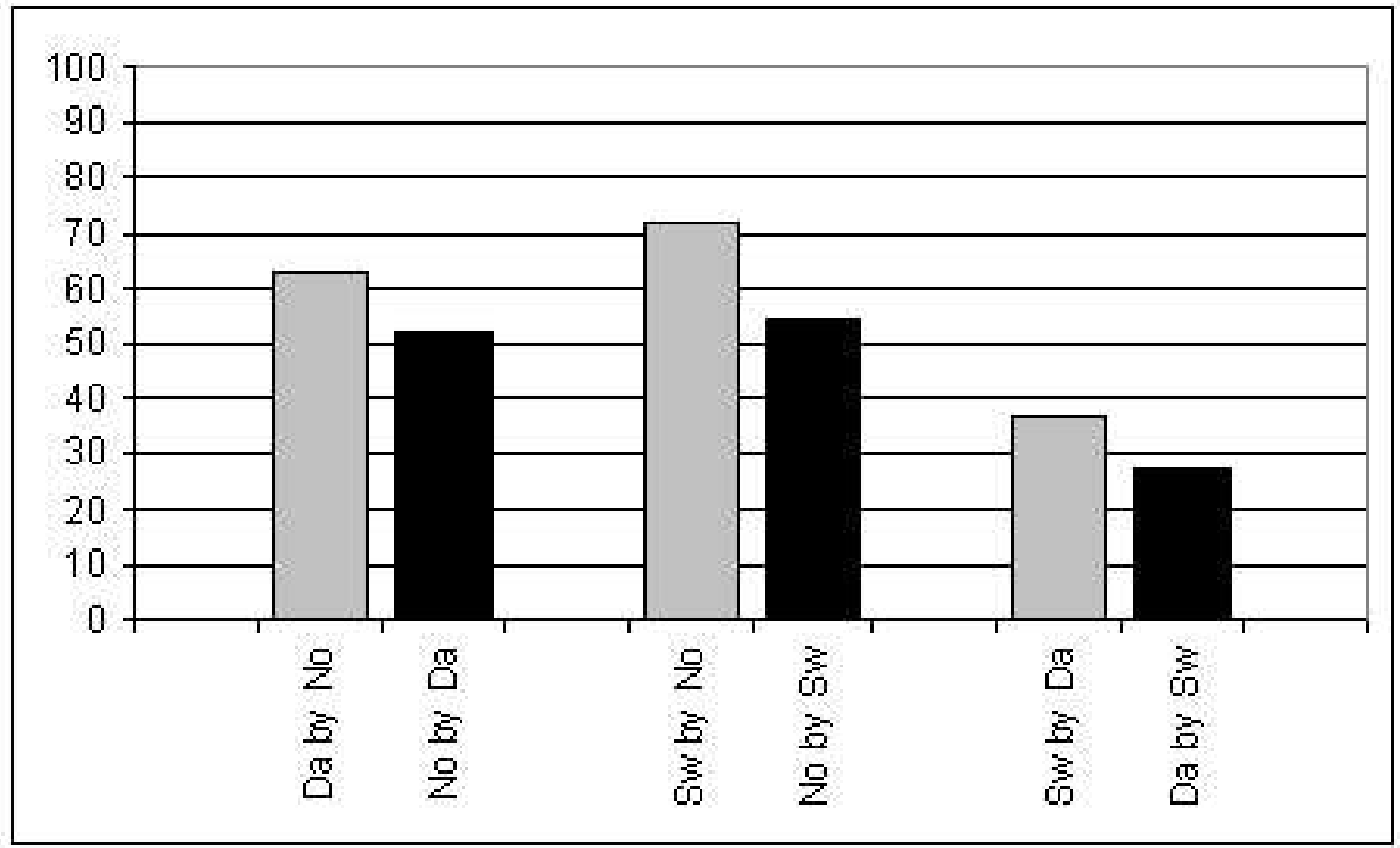
- Jens Moberg, Charlotte Gooskens CLIN paper
- Charlotte Gooskens's NWO Vidi project modeling comprehensibility, including linguistics, attitudes, and experience
  Sebastian Kürschner, Renée van Bezooijem
- **Nathan Vaillette**, Massachusetts
  VW project proposal – Groningen, Tübingen, Sofia
- Erhard Hinrichs, Kiril Simov, Petya Osenova, Jelena Prokić, Thomas Zastrow

Background

Scandinavian “Semicommunication”

- Scandinavians—e.g., Swedes and Danes—hold conversations
  - where each speaks his own language, the Swede Swedish, and the Dane Danish
  - comprehend one anothers’ languages, but imperfectly and asymmetrically
    Danes understand Swedes better than *vice versa*
- Haugen 1966: “semicommunication”, Braunmüller, ca. 2004 “receptive multilingualism”
- Research has focused on attitudes and experiences als explanatory factors
- What about linguistic structure?

Comprehension



Sources: Maurud (1976), Bø (1978), Delsing & Åkesson (2005)

Idea

Danish	j	ɑ	i	
Swedish	j	ɑː	g	

Danish	l	ɑ	ŋ	ʔ
Swedish	l	ɔ	ŋ	#

Swedish problem: map Danish to Swedish
Danish problem: map Swedish to Danish

Map Foreign to Native

Idea, part 2.

Problem: Map Foreign to Native

Hypothesis: Comprehensibility is mirrored by complexity of mapping

Complexity of Mapping

Danish	j	ɑ	i	
Swedish	j	ɑː	g	

Danish	l	ɑ	ŋ	ʔ
Swedish	l	ɔ	ŋ	#

Note: Swede needs to map ɑ $\longrightarrow$ ɑː *and* ɑ $\longrightarrow$ ɔ

Swede must decide, increasing complexity

Conditional Entropy

$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) \log_2 p(x|y)$$

$$H(\text{Native}|\text{Foreign}) = -\sum_{n \in N, f \in f} p(n,f) \log_2 p(n|f)$$

Given foreign words, how hard is it to map to native words?

Comprehensibility

Hypothesis: Comprehensibility is modeled by $H(\text{Native}|\text{Foreign})$

- Acoustic experience is given, we map to native categories
- Comprehensibility is a relation between linguistic varieties
  - Individual words are more or less comprehensible against the background of the varietal mappings

Sub-vocabularies

- Cognates (since proto-Scandinavian)
- German loans
- Greek, Latin, French loans

Expectation:
Entropy(Cognate Mappings) $>$ Entropy(German Loans) $>$ Entropy(Other Loans)

The longer a word is in a language, the more time it has to drift in pronunciation

Sub-vocabulary

- Function words vs. content words

Expectation:
Entropy(Function Words) $>$ Entropy(Content Words)

Function words are more frequent, less important and therefore reduced more in pronunciation. Function words are also immune to (some) sound changes, therefore exceptional.

Data

- 1,500 most frequent word meanings in *Corpus Spoken Dutch*, informal speech
- 1,500 most frequent Swedish words in *Europarl* (more formal)
- 1,500 most frequent words in intersection of above, translations into Scandinavian, West Germanic languages

Preprocessing

- All words transcribed phonemically
- All words aligned via edit distance, C/V matched prohibitively costly
- Counts used to estimated conditional probabilities

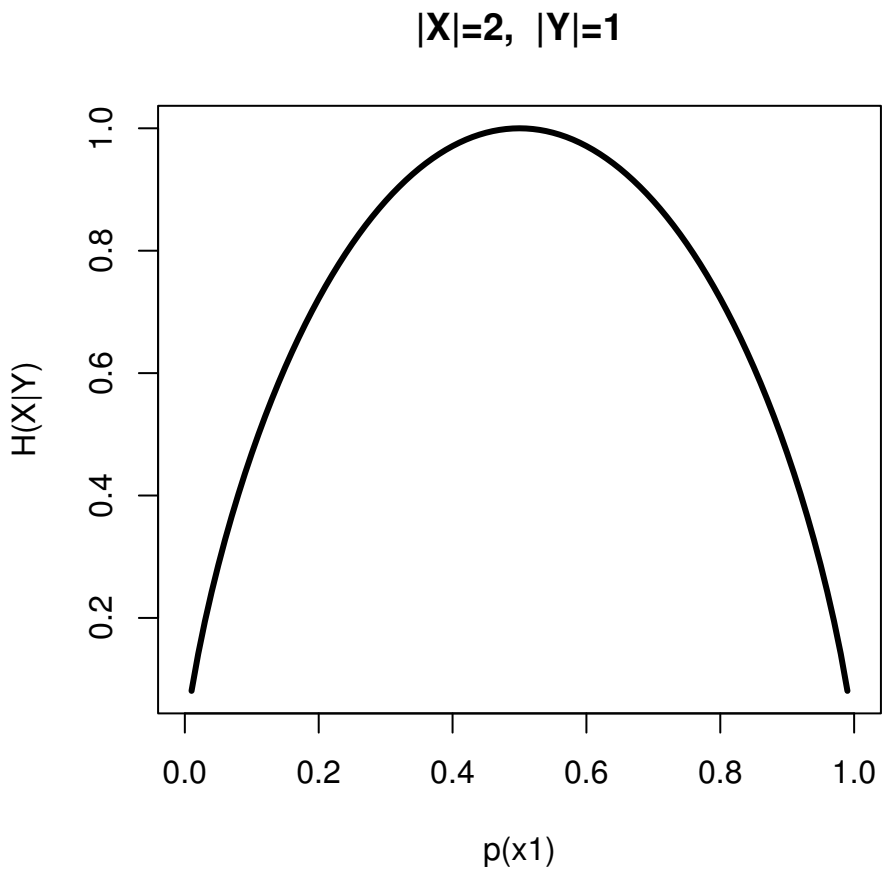
Calculations

Danish	j	ɑ	i	
Swedish	j	ɑː	g	

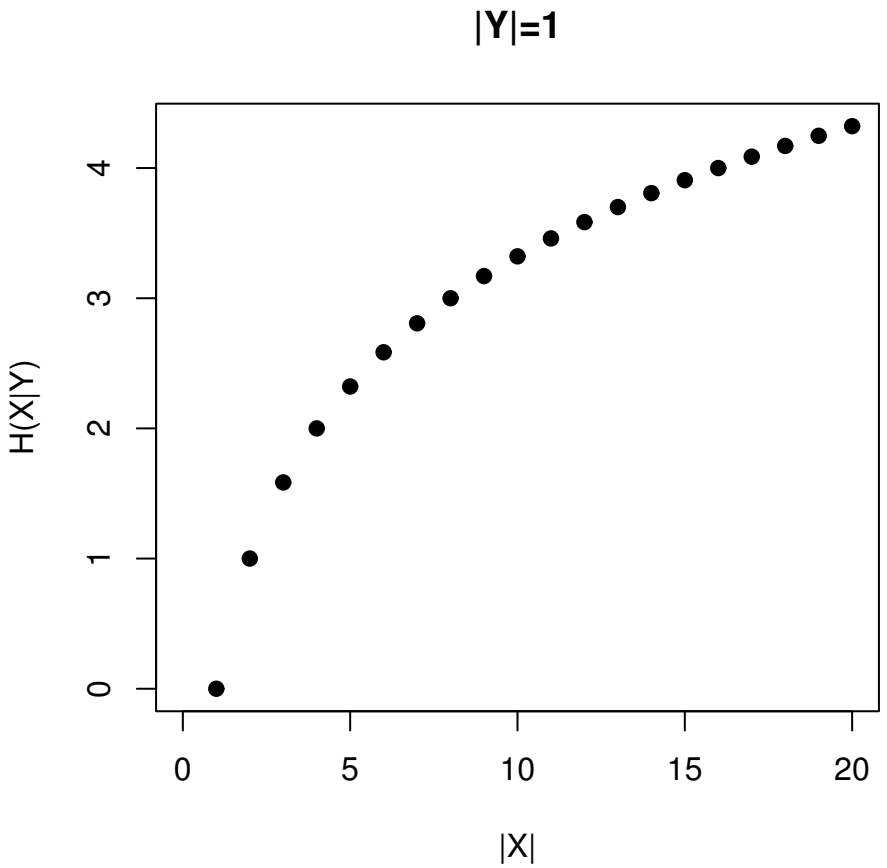
Danish	l	ɑ	ŋ	ʔ
Swedish	l	ɔ	ŋ	#

Except for $p(S|\text{ɑ}_{\text{Dane}})$, all conditional probabilities (mappings) are certain, $p(n|f) = 1, log_2(p(n|f)) = 0$, contributing nothing to entropy.

Conditional Entropy

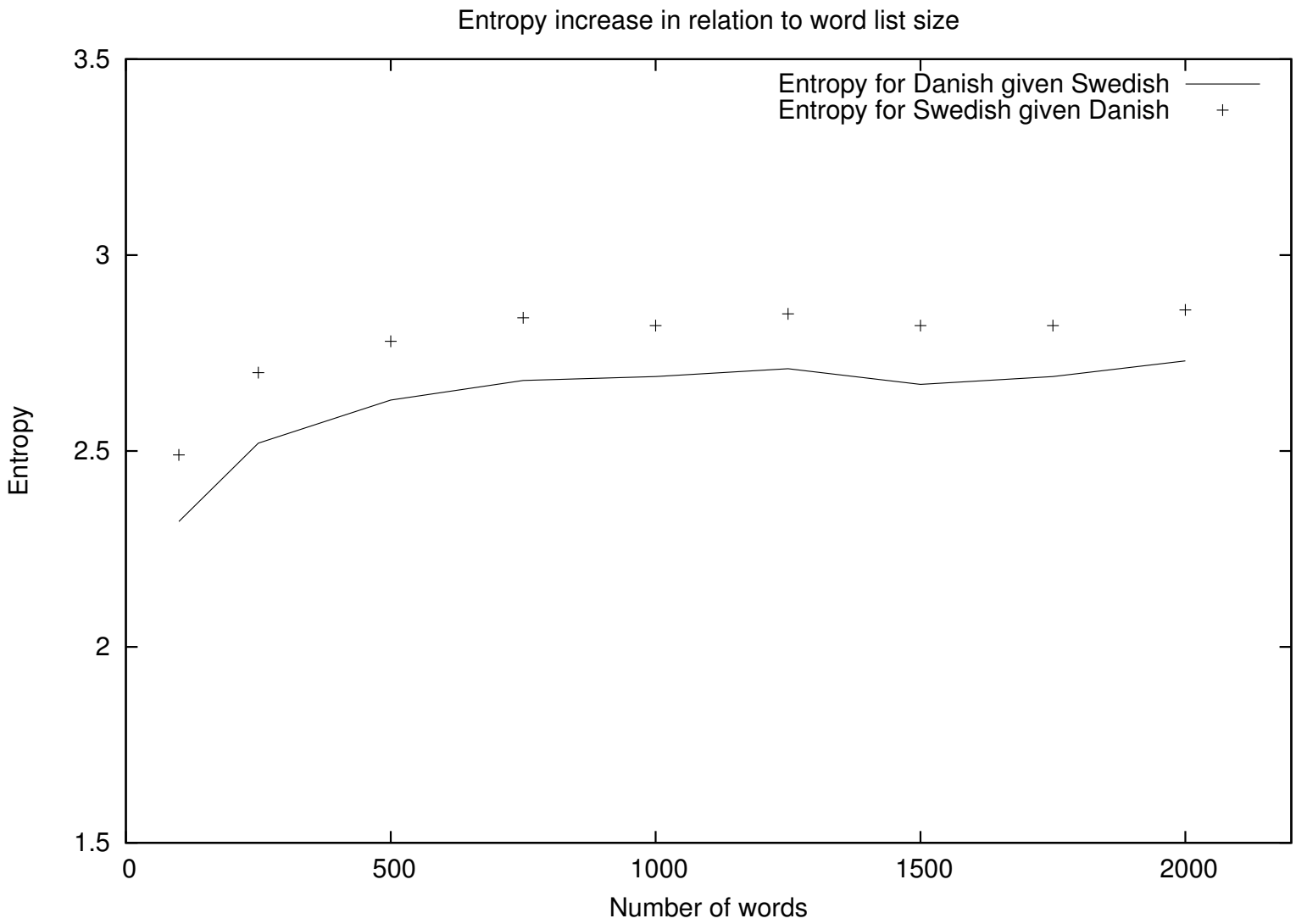


Conditional Entropy

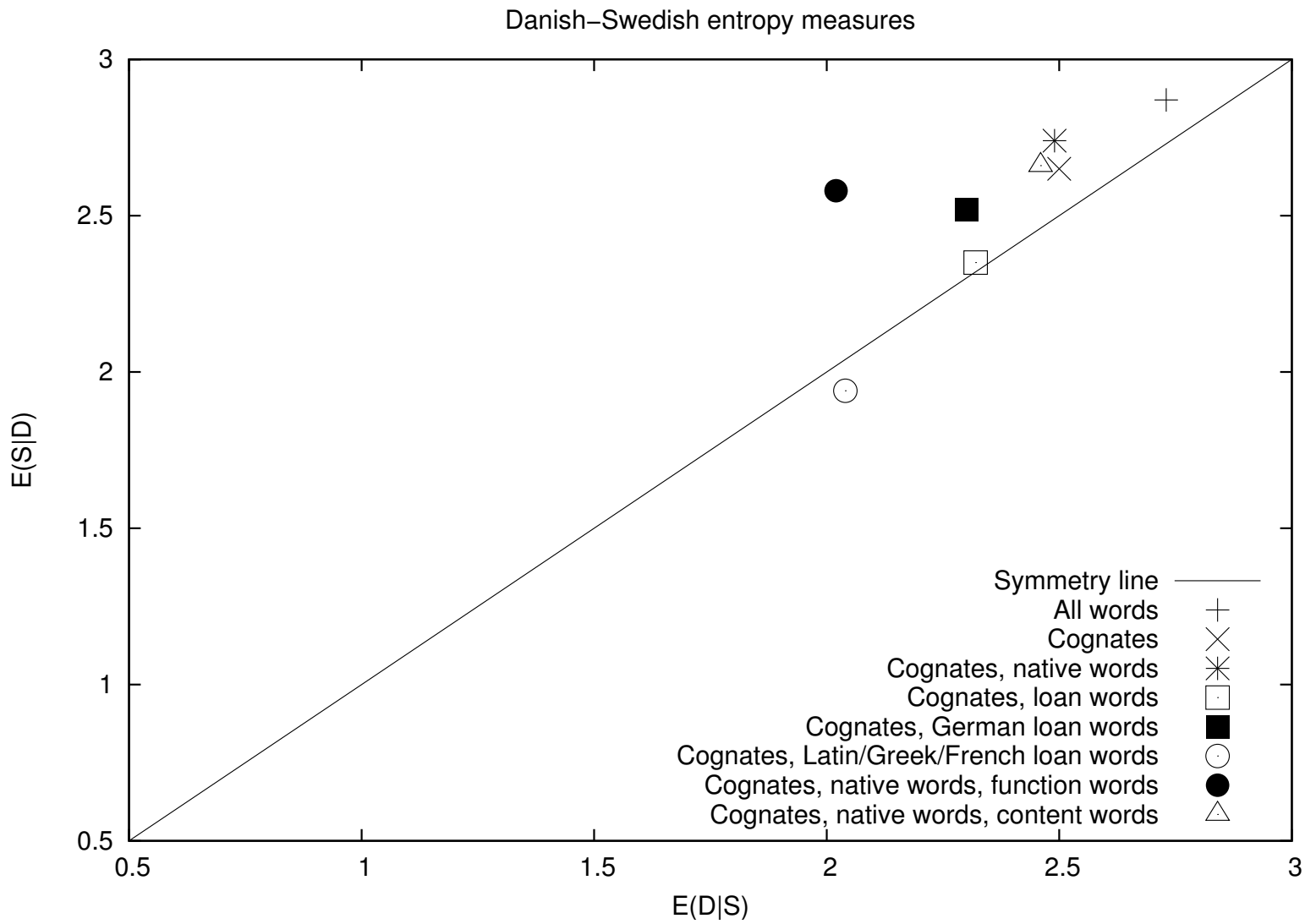


Conditional entropy as a function of the number of **equally likely** images x in mapping $Y \rightarrow X$.

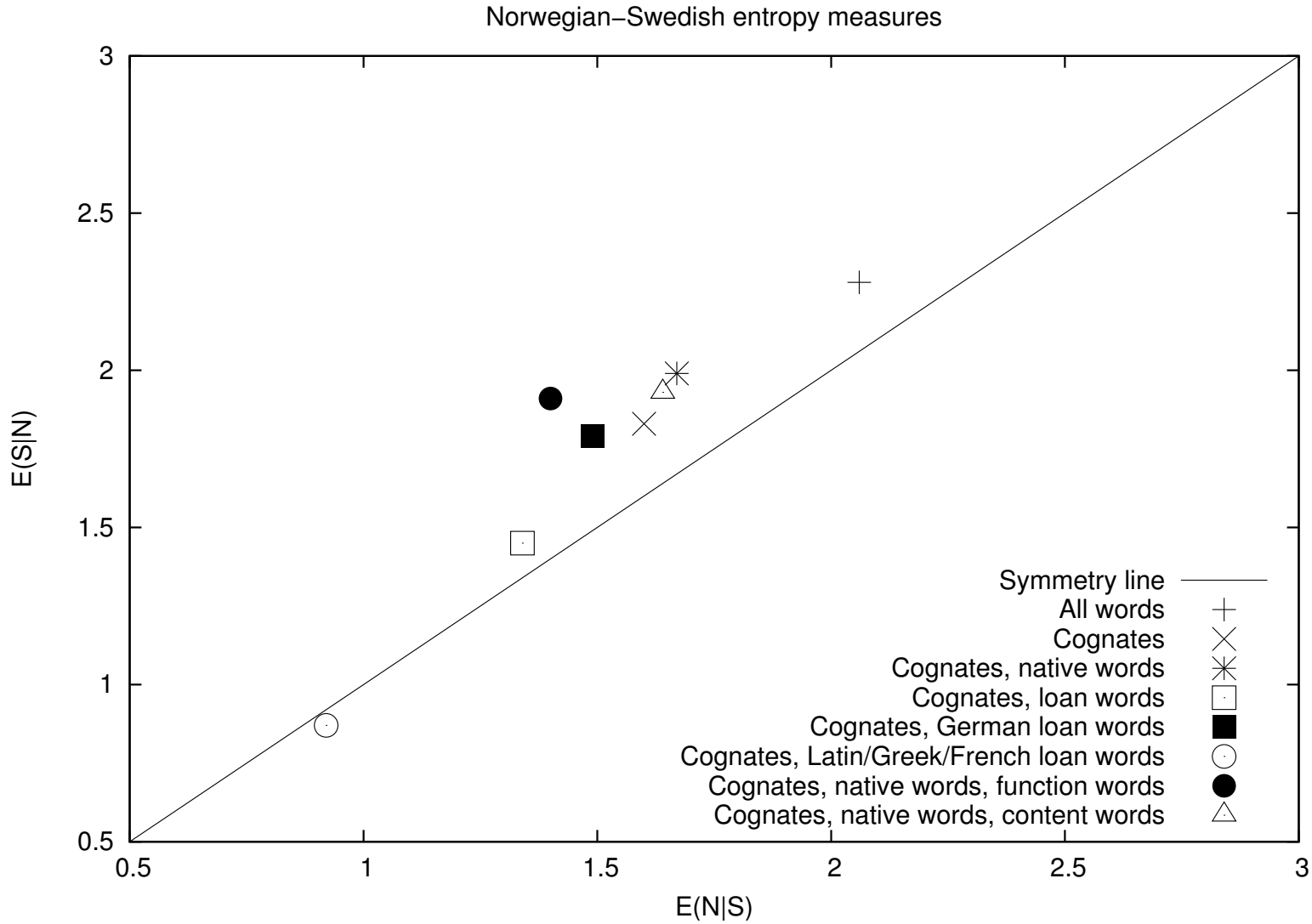
How Much Data for Estimation?



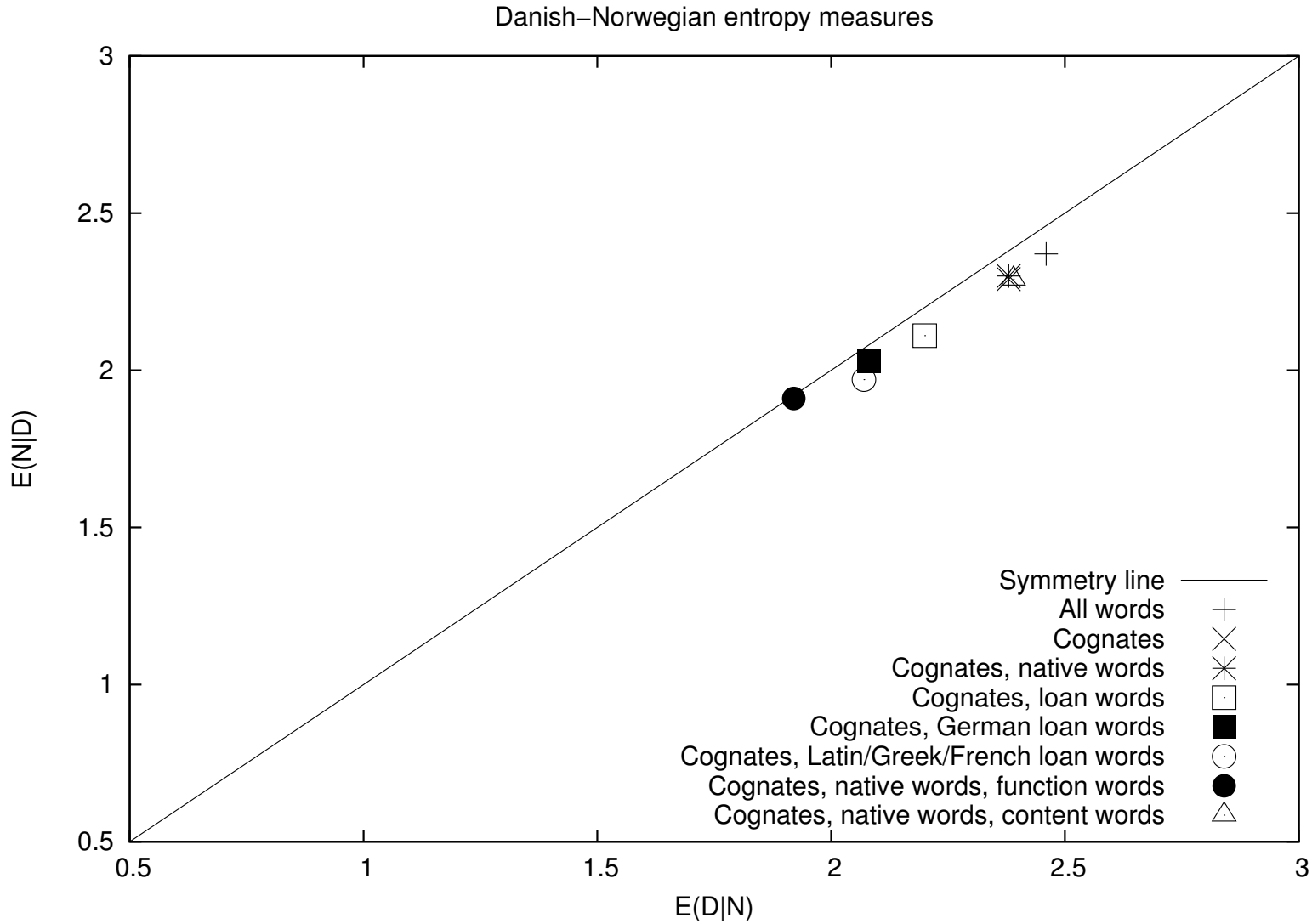
Danish – Swedish



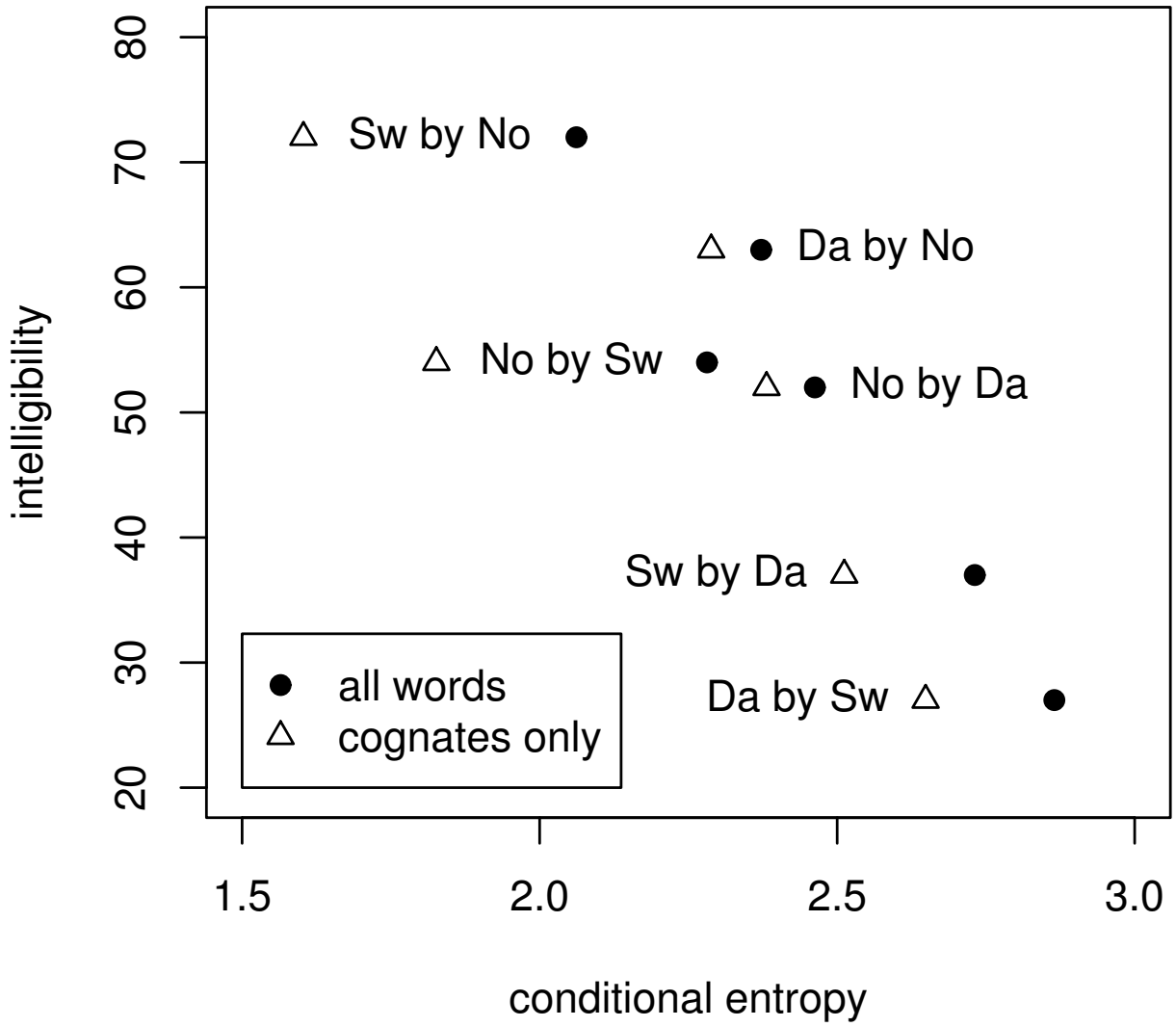
Norwegian – Swedish



Danish – Norwegian



Intelligibility



Conclusions and Discussion

- Intelligibility correlates negatively, and nearly perfectly with the conditional entropy of the phoneme mapping.

- Only six data points—pairs of Scandinavian languages —More needed!
- Technical refinements possible, but difficult: contextual sensitivity, special status of identity mapping, phonetic detail
- Dutch varieties in sights!

General Conclusions

- Linguistic structure vindicated as explanatory.
- Need comparison to attitude, experience.

- Quantitative, computational techniques essential to operationalizing the influence of phoneme structure.
- Perhaps “comprehensibility” rewarding because we needed to measure properties of entire languages in relation to one another.
- Linguistic, psycholinguistic techniques shy away from such properties of “aggregates”.

Reflection

- Only accomplished “semi-communicators” will command the entire mapping (all conditional probabilities).
- Won’t accomplished semi-communicators simply know the words they hear, obviating the need for a phoneme mapping?
- Yes, but perhaps we filter everything through our grid of our native phoneme inventory, even as accomplished semi-communicators

Reflection

- Only accomplished “semi-communicators” will command the entire mapping (all conditional probabilities).
- Won’t accomplished semi-communicators simply know the words they hear, obviating the need for a phoneme mapping?
- Yes, but perhaps we filter everything through our grid of our native phoneme inventory, even as accomplished semi-communicators

Reference

Jens Moberg, Charlotte Gooskens and John Nerbonne. Conditional Entropy Measures Intelligibility among Related Languages. Accepted (5/2007) to appear in: Frank Van Eynde, Peter Dirix, Ineke Schuurman & Vincent Vandeghinste (eds.) *Proceedings of Computational Linguistics in the Netherlands 2006* Amsterdam: Rodopi. ca. 2007
See http://www.let.rug.nl/nerbonne/papers/