Conditional entropy as a measure of comprehension ease between related languages

Jens Moberg, Charlotte Gooskens and John Nerbonne
j.moberg@rug.nl, c.s.gooskens@rug.nl, j.nerbonne@rug.nl
Alfa-Informatica, University of Groningen, The Netherlands

Background

The three mainland Scandinavian languages (Danish, Norwegian, Swedish) are very similar. This enables **semicommunication**, where each interlocutor speaks in his/her mother tongue. The quality of semicommunication varies throughout Scandinavia, but is crucially asymmetric: e.g. Danes generally understand Swedes better than *vice versa*.

Idea

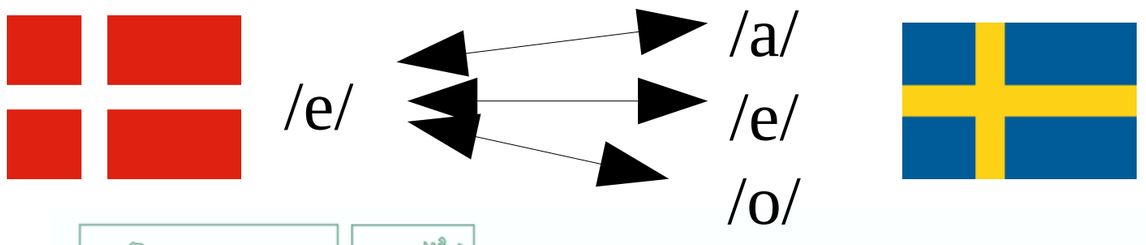
If language A distinguishes two versions of a single phoneme in language B, then speakers of B hearing A will be able to map it to their own system quite easily, while speakers of A hearing B will be puzzled: which is the corresponding element?

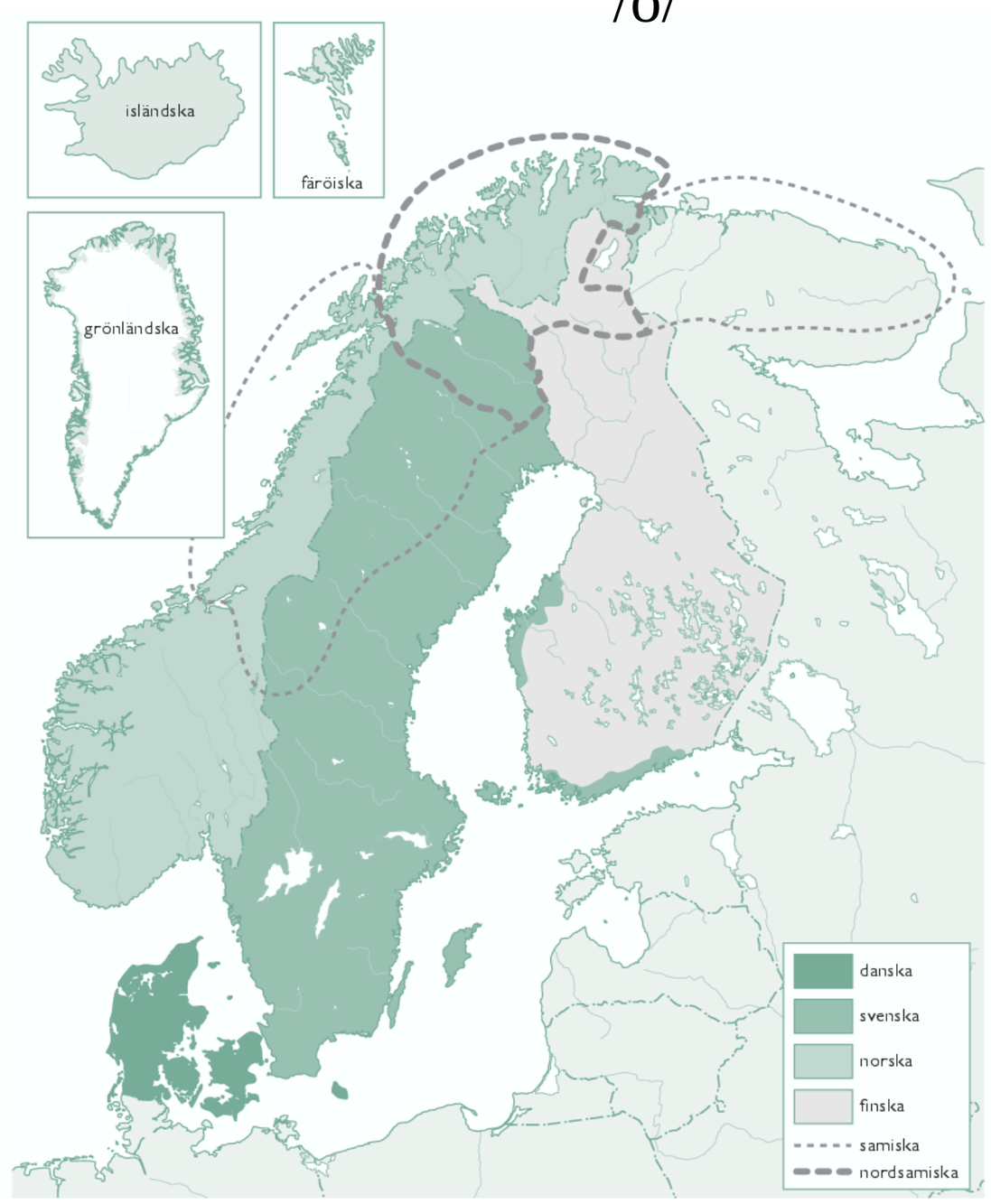
Hypothesis

The more complex the mapping from phoneme set A to B, the more difficult it is for speakers of B to understand language A

Mapping

- Written Danish uses the letter [e] in grammatical endings
- Written Swedish uses both [a], [e] and [o]
- Three choices for the Swede, one choice for the Dane





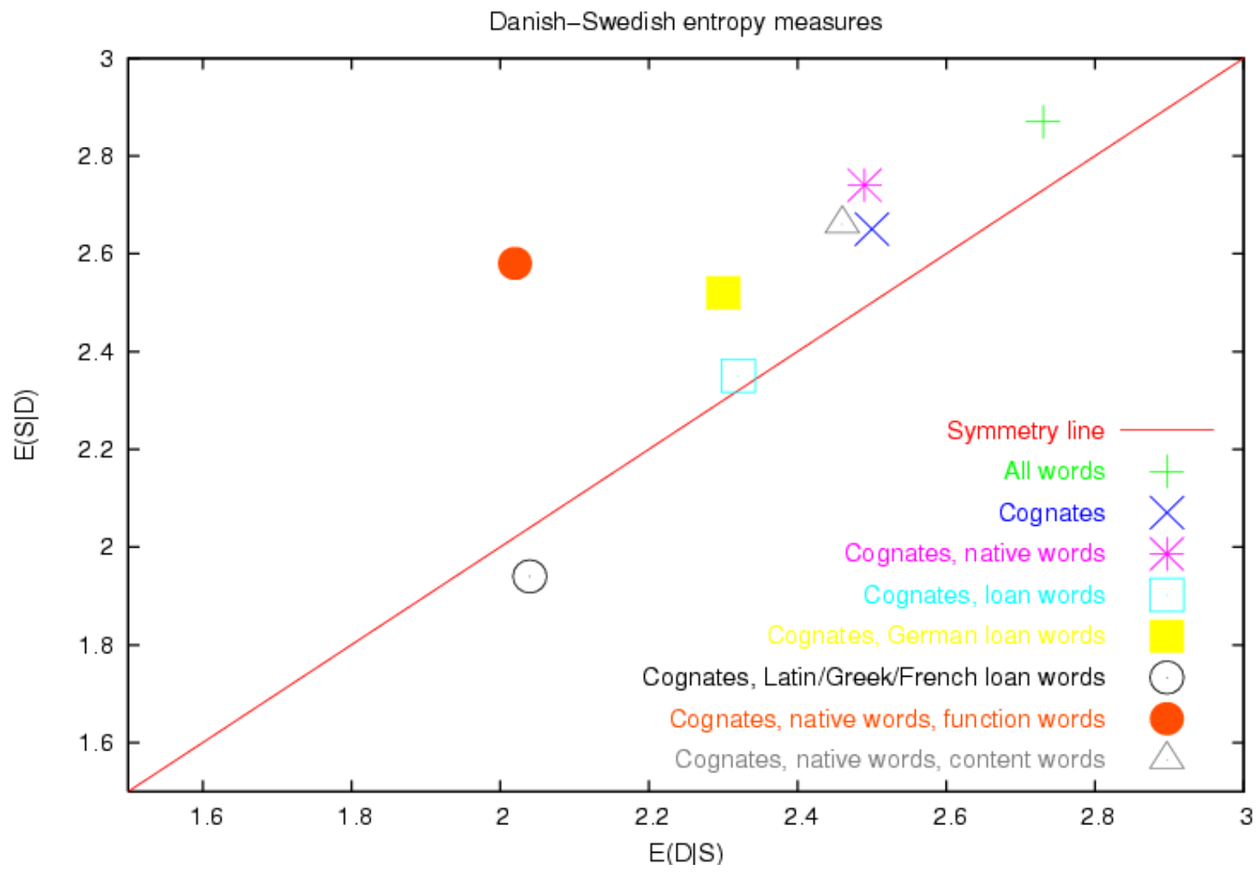
Measure mapping complexity via conditional entropy:

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

Material

Europarl: formal speech corpus
CGN: informal speech corpus

Results (Danish and Swedish)



Next steps

- Look at other language pairs
- Behavioral test
- Include context (bigrams)