Onderlinge verstaanbaarheid van Nederlands en Duits

Onderlinge verstaanbaarheid van Nederlandse en Duitse cognaten gemodelleerd met behulp van automatische spraakherkenning

Universiteit Leiden

rijksuniversiteit groningen

Vincent J. van Heuven, Charlotte Gooskens, Renée van Bezooijen

Inleiding: onderzoeksthema

- Onderzoeksthema onderlinge verstaanbaarheid
- Hoe meten onderlinge verstaanheid?
  - Oordelen (opnion testing, judgments)
  - Functionele tests (woordherkenning, dictee)
- Hoe verklaren we onderlinge verstaanbaarheid?
  - Linguistisch verschillen/overeenkomsten
  - Extralinguistische factoren

Inleiding: asymmetrie

- Asymmetrie in literatuur
  - Braziliaans-Portugeessprekenden verstaan Argentijns-Spaans beter dan omgekeerd (Jensen 1989)
  - Zuid-Chinese dialectsprekers verstaan noord-chinese (Mandarijn) dialecten beter dan omgekeerd (Cheng 1997, Tang & van Heuven 2009)
  - Denen verstaan Zweden beter dan omgekeerd (Gooskens et al. 2010)

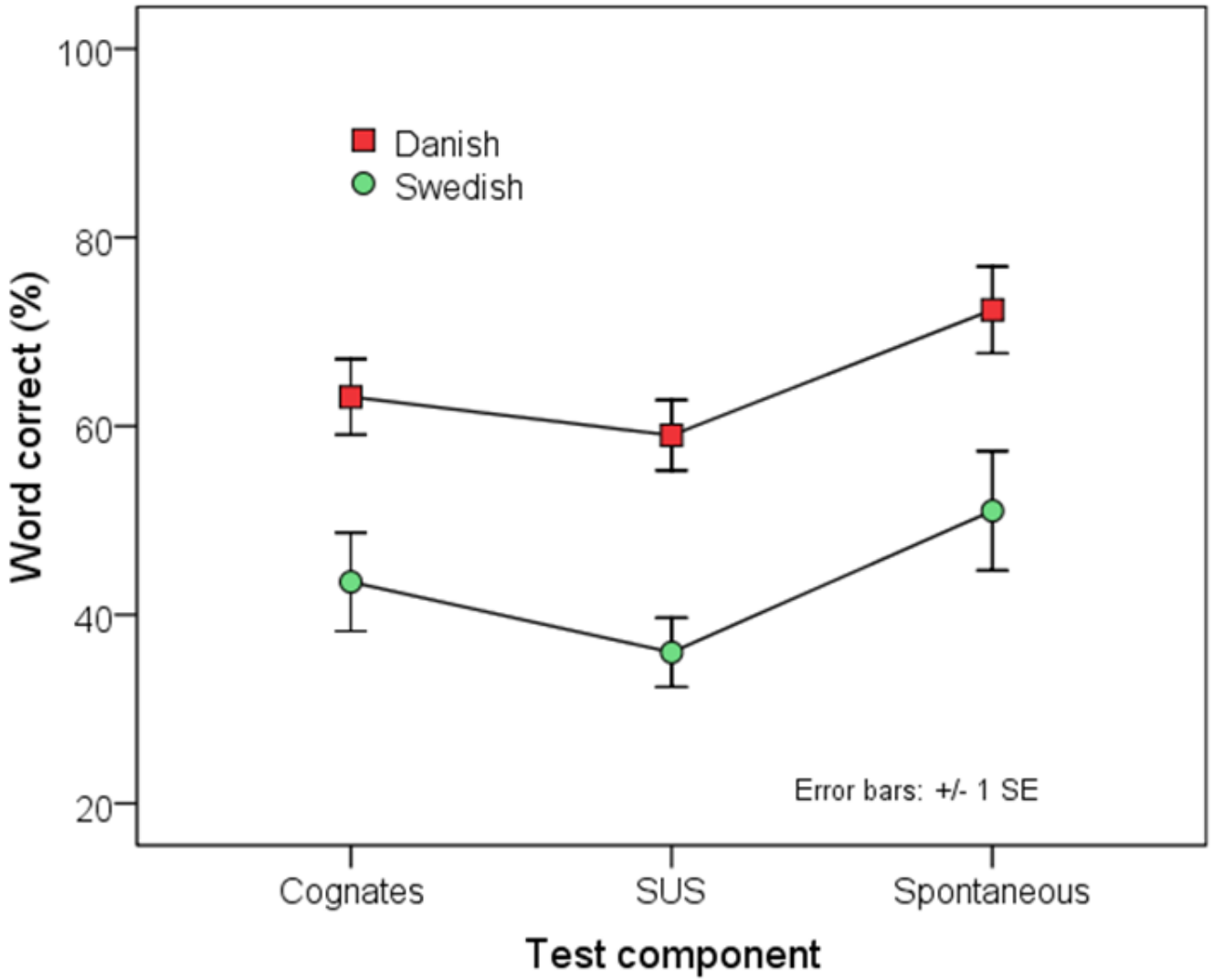


Figure 3. Intelligibility scores (percent correctly recognized words) obtained on three test components by Danish listeners decoding Swedish (squares) and by Swedish listeners decoding Danish (circles). Listener groups (20 Danes, 20 Swedes) were matched with respect to lexical knowledge of and familiarity with the non-native language. Error bars are +/− 1 standard error of the mean.

Inleiding: asymmetrie

- Asymmetrie vaak extralinguistisch verklaard
  - Verschil in ervaring met de andere taal
    - Vaak ook ogv geografie
  - Een taal is sociaal dominant
  - Asymmetrie in attitude tav de talen

Inleiding: asymmetrie

- Wij willen weten in hoeverre asymmetrie in onderlinge verstaanbaarheid een linguistisch verklaard/voorspeld kan worden, in abstractie van extralinguistische factoren
- Experimentele methoden vereist
  - Bv. mbv automatische spraakherkenning

Case van vandaag

- Onderlinge verstaanbaarheid NL ~ D
  - Enige verwante talenpaar waarvoor een commerciële multilinguale spraakherkenner beschikbaar is
- Is asymmetrisch
  - Nederlanders verstaan beter Duits dan Duitsers Nederlands (data?)
  - Kan gemakkelijk extra-linguistisch verklaard worden (geografie, dominantie, onderwijs, media)
- Maar is er ook taalkundig asymmetrie?
  - Geen reden om dat te veronderstellen

Vraagstelling

- Hoe moeilijk is het voor een Nederlander verwante Duitse woorden te herkennen, en vice versa, als zij de andere taal voor het eerst in hun leven zouden horen?
  - Alleen cognaten
  - Hoe groter het klankverschil, des te moeilijker de herkenning
  - Kan asymmetrisch liggen (neutralisaties, bv. gevonden voor Chinese dialecten)
  - Maar niet voor NL ~ D

Methode: ASH

- Automatische spraakherkenning
  - Trainingsfase:
    - systeem leert NL klanken en opeenvolgingen
      - Simuleert een NL luisteraar (zonder kennis van D)
    - systeem leert D klanken en opeenvolgingen
      - Simuleert een D luisteraar (zonder kennis van NL)
    - zgn. Hidden Markov klankmodellen (HMM's)

Methode: ASH

- Automatische spraakherkenning
  - Testfase (na voltooide training van systeem):
    - NL systeem hoort NL testmateriaal (hoge score?)
    - D systeem hoort D materiaal (hoge score?)
  - Testfase 2: luisteren naar de andere taal
    - NL systeem hoort D testmateriaal (lage score?)
    - D systeem hoort NL testmateriaal (lage score?)
  - Wel/geen asymmetrie in testfase 2?

Methode: ASH

- Praktisch probleem
  - HMM Klankmodellen sterk sprekerafhankelijk
  - Daarom per spreker aparte training nodig
  - Gekruiste test alleen mogelijk als de NL en D spreker dezelfde persoon zijn
  - Vereist een perfect bilinguale spreker

Excursie

- De jacht op de perfecte bilinguaal
- Gebruik van voice line-up
  - Bilinguaal mag niet aangewezen worden als afwijkend in een rij monolinguale sprekers
    - Door NL beoordelaars
    - Door D beoordelaars
- Heeft heel veel moeite gekost…
  - …maar is gelukt

Excursie

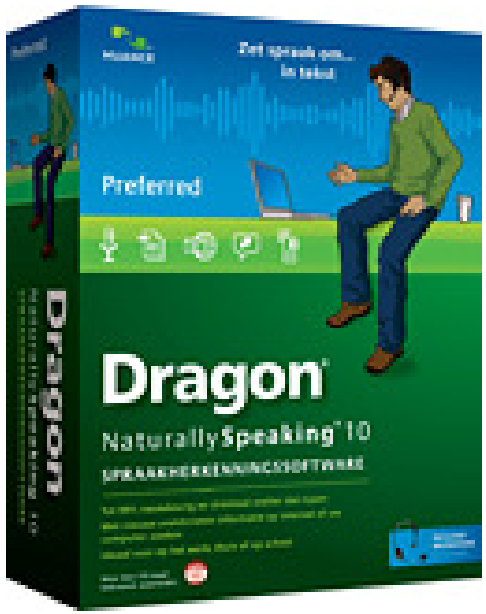
- Onze bilinguale spreekster MM
  - Geb. 1976 in Zwitserland uit NL ouders
  - Thuistaal NL, op school Zwitser-D
  - Vanaf 1996 in NL (studeerde NL en D)
  - Vanaf 2000 werkzaam in Duitsland (Berlijn, Potsdam, Dortmund) maar met onderbreking
- Kwam als enige bilinguaal ongedetecteerd door zowel de NL als de D voice line-up
  - Geluidsdemonstratie ahv stukjes trainingstekst

Excursie

- Bilinguale spreekster MM
  - Nederlands 
  - Duits 

Spraakherkenner

- Nuance (vroeger Lernout & Hauspie Speech Products) Dragon NaturallySpeaking version 10 voor NL en voor D, ca. € 100 per taalmodule
- Standaardversie (geen gespecialiseerde lexica)



Testmateriaal

- 3000 meest frequente nomina
  - Celex NL, Celex D
- Cognaatparen met de hand toegewezen
  - Ca. 750 paren, rest valt af
- Geordend op gemiddelde tokenfrequentie per paar
- Als losse woorden ingesproken door bilinguaal MM
- Los aangeboden aan herkenner, steeds gevolgd door “punt/Punkt” (uitschakelen “taalmodel”)
- Herkenning in batch mode (niet interactief)

Resultaten (1)

- Training van herkenner met vooraf opgenomen trainingsteksten verliep vlekkeloos
- Testfase 1 (testen in getrainde taal)
  - NL: 220 correct uit 768 (29%)
  - Woorden uit begin van de test gaan belangrijk beter (hogere tokenfrequentie?)
  - Daarom analyse voorlopig beperkt tot top-200

Resultaten (2)

- Testfase 1 (test in getrainde taal)
- Top-200 woorden
  - NL: 131/200 = 66% correct
  - D: 146/200 = 73% correct
- Controle:
  - Eigen stem getraind en getest op top-200
  - NL: 128/200 = 64% correct

Interimconclusie (1)

- Dragon NaturallySpeaking
  - Is niet goed in herkenning van losse woorden
  - Kan geen gebruik maken van context-afhankelijkheden
  - Heeft te weinig voorbeelden gezien van begin- en eindklanken
    - Eerste en laatste klanken van testwoorden gaan opvallend vaak fout

Resultaten (3)

- Testfase 2: training- en testtaal gekruist
  - NL (na D-training): 9/200 correct (5%)
  - D (na NL-training): 7/200 correct (4%)
- Na top-200 stimulus-responsie niet op te lijnen (herkenningen schijnbaar random)
- Correcte herkenningen alleen bij (vrijwel) identieke cognaten

Correct herkende cognaten

	NL	> D		D	> NL
1.	broeder	Bruder	1.	Bruder	broeder
2.	radio	Radio	2.	Radio	radio
3.	loon	Lohn	3.	Lohn	loon
4.	idee	Idee			
5.	artikel	Artikel			
6.	roman	Roman			
7.	ingenieur	Ingenieur			
8.	winter	Winter			
9.	bier	Bier			
			4.	Frau	vrouw
			5.	Werk	werk
			6.	Vater	vader
			7.	Ring	ring

Conclusie (2)

- Herkenningsresultaten fase 2 geven geen aanleiding een asymmetrie te veronderstellen in onderlinge verstaanbaarheid NL ~ D
  - 7 vs 193 en 9 vs 191 verschilt statistisch niet significant (chi kwadraat)

Discussie

- Prestaties verbeteren?
  - Lexicon opschonen
    - Alle eigennamen verwijderen
  - Ontbrekende woorden toevoegen
  - Minimale context toevoegen bij inspreken
    - Lexicale categorie inperken tot nomen, maar zonder geslachtinformatie, bv.
    - *Er hat “ohne X” gesagt*
    - *Hij heeft “zonder X” gezegd*

Slot

- Wordt vervolgd...
- Dank voor uw aandacht
- Referenties

  - Jensen, J. B. (1989). On the mutual intelligibility of Spanish and Portuguese. *Hispania* 72, 848-852.
  - Tang, C. & V.J. van Heuven (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119, 709-732.
  - Gooskens, C., V.J. van Heuven, R. van Bezooijen & J.J.A. Pacilly (2010). Is spoken Danish less intelligible than Swedish? *Speech Communication*, (in press).