Zo kun je het ook zeggen!

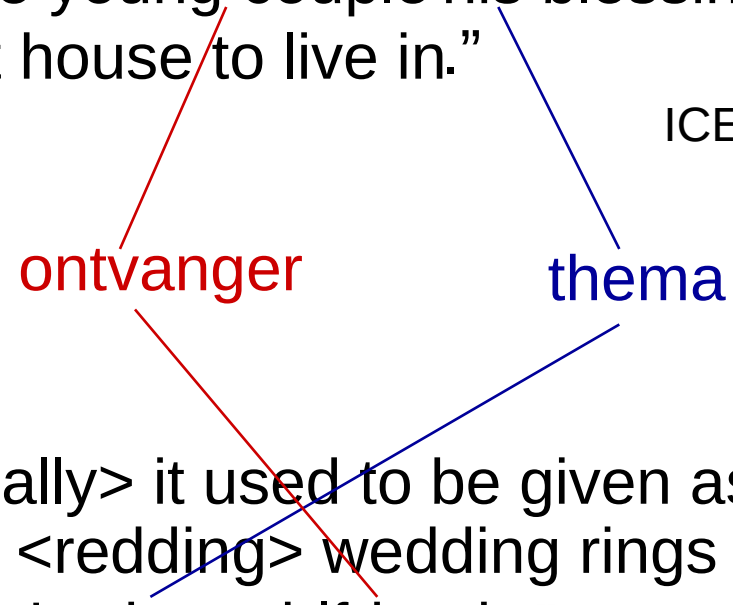
Het modelleren van syntactische taalvariabiliteit

Daphne Theijssen
Promovenda
Afdeling Taalwetenschap
Radboud Universiteit Nijmegen
d.theijssen@let.ru.nl

Datiefalternantie

- “But Isabel talked him round in the end, and he gave the young couple his blessing and a rather elegant house to live in.”

ICE-GB W2F-011_52:1



- “<It's really> it used to be given as fourteenth-century <redding> wedding rings and nowadays blokes give it to girlfriends ”

ICE-GB S1A-047_216:1:B

Datiefalternantie

- “But Isabel talked him round in the end, and he gave the young couple his blessing and a rather elegant house to live in.”

  ICE-GB W2F-011_52:1

- “But Isabel talked him round in the end, and he gave his blessing and a rather elegant house to live in to the young couple.” ?

Datiefalternantie

- “<It's really> it used to be given as fourteenth-century <redding> wedding rings and nowadays blokes give girlfriends it”

?

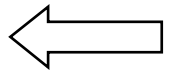
- “<It's really> it used to be given as fourteenth-century <redding> wedding rings and nowadays blokes give it to girlfriends”

ICE-GB S1A-047_216:1:B

Datiefalternantie

Welke constructie past hier het best?

- “I mean you aren't going to honestly give them any priority”

- “I mean you aren't going to honestly give any priority to them” 

ICE-GB S1A-047_216:1:B

Onderzoeksvraag

Kunnen we de datiefalternantie voorspellen?

Deze presentatie

- Gerelateerd onderzoek door Bresnan et al. 2007: gesproken Amerikaans Engels
- Onderzoeksdoelen
  - Toepassen van Bresnan et al.'s model op gesproken Brits-Engels
  - Het gevonden model toepassen op geschreven Brits-Engels
- Experimentele opzet
- Doel 1: Gesproken Brits-Engels
- Doel 2: Geschreven Brits-Engels
- Conclusie en aanbevelingen
- Computerlinguïstiek over 20 jaar
- Vragen

Gerelateerd werk door Bresnan et al. (2007)

- 2360 voorkomens in Switchboard (Godfrey et al. 1992)
- Lineaire regressie-modellering
  - Eigenschappen uit de literatuur
  - 95.0% van de data is goed voorspeld (5.0% onverklaard)
- Toevoeging van geschreven data (financiële teksten)
  - 905 voorkomens uit het Wall Street Journal (Penn Treebank)
  - 93.4% goed voorspeld

- Conclusie:
  - In gesproken Amerikaans-Engels is de datief-alternantie goed te voorspellen
  - Het model generaliseert naar geschreven Amerikaans-Engels

Gerelateerd werk door Bresnan et al. (2007)

Harmonische oplijning

dubbel-objectconstructie

hand **the student** *the book*

ontvanger < *thema*
discourse gegeven niet-gegeven
pronominal niet-pronominaal
bepaald onbepaald
animate

prepositionele datiefconstructie

hand **the book** to *the student*

thema > *ontvanger*
discourse gegeven niet-gegeven
pronominal niet-pronominaal
bepaald onbepaald
inanimate

Onderzoeksdoelen

1. Bresnan et al.'s (2007) regressiemodel voor gesproken Amerikaans-Engels toepassen op gesproken Brits-Engels

2. Het gevonden model voor gesproken Brits-Engels toepassen op geschreven Brits-Engels

Experimentele opzet: Data

- Syntactisch geannoteerde ICE-GB corpus (Greenbaum 1996)
- Gesproken teksten
  - dialogen (prive en publiek)
  - monologen (ongeschreven en geschreven)
- Geschreven teksten
  - ongedrukt (werkstukken door studenten en brieven)
  - gedrukt (academisch, populair, reportage, instructief, betogend en creatief)
- De voorkomens worden gevonden met een Perl-script

Experimentele opzet: Data

- Weggelaten (volgens Bresnan et al. 2007):
  - **andere preposities dan *to***

    e.g. “nobody buys me a book and I can't buy them for myself <,>”

    S1A-013_118:1:A
  - **passieven**

    e.g. “Dido 's pride has been dealt a severe blow .” W1A-010_81:1
  - **clausale objecten**

    e.g. “so doctors will tell you that they 've only just discovered this idea”

    S2B-038_4:1:A
  - **omgekeerde constructies**

    e.g. “lending to the houses and pedestrians a faintly unreal or even theatrical quality” W2B-006_106:1

Experimentele opzet: Data

- Ook weggelaten:
  - **gecoordineerde werkwoorden of werkwoordclusters**
    e.g. “However, anyone caught importing or supplying large quantities of the drug to others will invariably be prosecuted.” W2B-020_47:1
  - **partikelwerkwoorden**
    e.g. “I 'll send you out that” S1B-074_46:1:B
  - **vraagzinnen en gebiedende wijs**
    e.g. “Please don't tell me that” S1B-042_133:1:A
  - **werkwoorden die met slechts één constructie voorkomen**
    e.g. “With the skill of many years of negotiation behind him , Dennis stalled long enough to pass a message to Lynne , giving her the option to call Pete . W2B-004_19:1

Experimentele opzet: Data

Uiteindelijke datasets:

Type	Prep.dat.	Dubb.obj.	Totaal
gesproken Am.	501	1859	2360
gesproken Br.	110	386	496
geschreven Br.	67	227	294

Experimentele opzet: Data

Handmatige annotatie van:

- Pronominaliteit van ontvanger + thema (pronominaal, niet-pronominaal)
- Bepaaldheid van ontvanger + thema (bepaald, onbepaald)
- Animacy van ontvanger (animate, inanimate)
- Persoon van ontvanger (1e/2e persoon, 3e persoon)
- Getal van ontvanger + thema (enkelvoud, meervoud)
- Concreetheid van thema (concreet, inconcreet)
- Bereikbaarheid van ontvanger + thema (discourse gegeven, niet-gegeven)
- Lengteverschil tussen het thema en de ontvanger (log. schaal)
- Semantische werkwoordsklasse (abstract, communicatie, overdracht)
- Structuurparallelisme / syntactische priming (ja, nee)

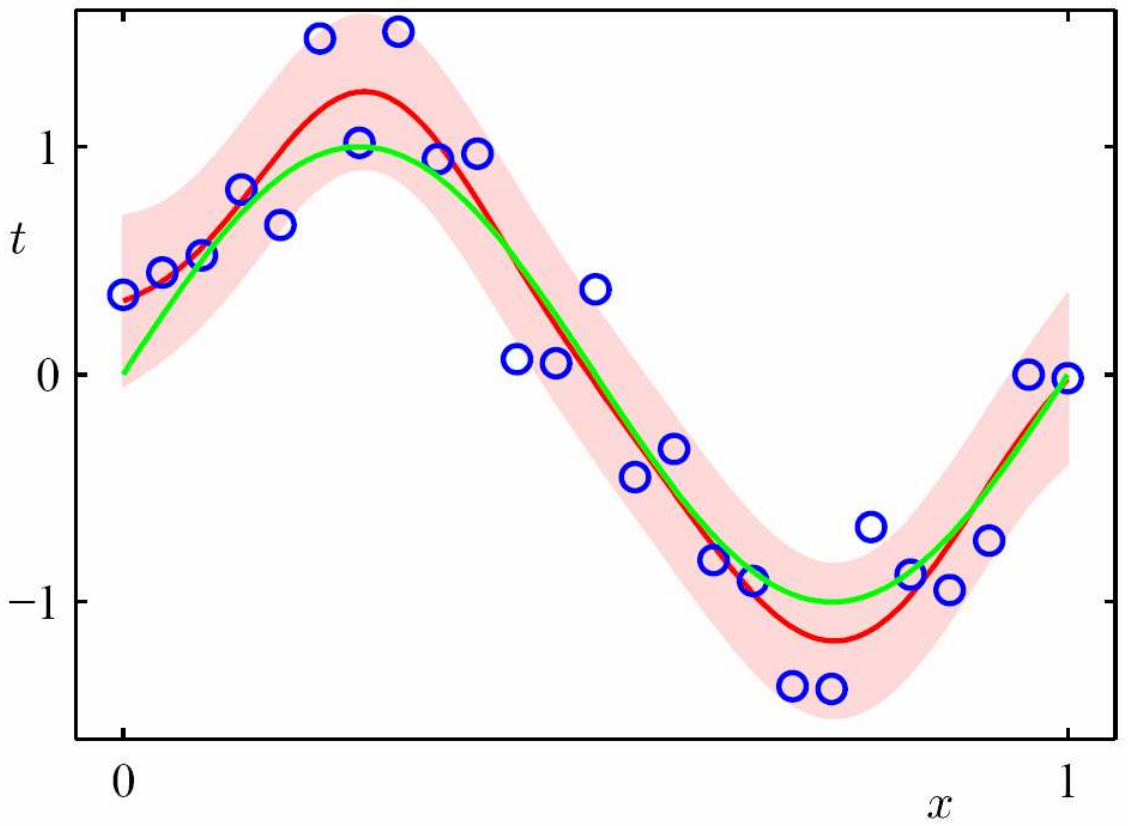
Experimentele opzet: Regressie

- Linear Mixed-Effect Modelling (Bates 2005)
  - Onafhankelijke effecten: 14 eigenschappen van vorige sheet (dezelfde als in Bresnan et al. 2007)
  - Random effect: werkwoord

- Werkwoord => aanname: lexicale voorkeur

  (Bresnan et al. 2007, Gries en Stefanowitsch 2004)

Experimentele opzet: LMER



bron: Biship (2006)

Experimentele opzet: LMER

De regressiefunctie:

- Voorspelt de kans dat constructie prep.dat. is
  Dus als de kans groter is dan 0.5, voorspelt het model prep. dat.
  - Model fit: trainen en testen op dezelfde data
  - Voorspellingskwaliteit: trainen op A, testen op B

  Hoe groter de kans, hoe zekerder is het model dat het prep. dat. is:
  - Prototypische en onverwachte constructies

- Geeft gewichten aan de variabelen
  - harmonische oplijning in Bresnan et al. (2007)

Gesproken Brits-Engels: Resultaten

Trainingset	model-fit	testset	voorspellingskwaliteit
Gespr.Am.	**94.6%**	Gespr.Am.	**90.1%**
Gespr.Am.	**94.6%**	Gespr.Br.	**87.5%** ⇦
Gespr.Br.	**94.1%**	Gespr.Br.	**84.7%**
Gespr.Am. sub	**95.7%**	Gespr.Br.	**85.4%**
Gesproken	**93.1%**	Gesproken	**89.5%**

Gesproken Brits-Engels: modelanalyse

Harmonische oplijning

dubbel-objectconstructie

hand **the student** *the book*

ontvanger < *thema*
discourse gegeven niet-gegeven
pronominal niet-pronominaal
bepaald onbepaald
animate
enkelvoud
1e/2e persoon
structuurparallellisme

prepositonele datiefconstructie

hand **the book** to *the student*

thema < *ontvanger*
discourse gegeven niet-gegeven
pronominal niet-pronominaal
bepaald onbepaald
inanimate
meervoud
3e persoon

Gesproken Brits-Engels: modelanalyse

Prototypische gevallen:

(1) the laryngograph shows us some information about how the vocal folds actually make contact whether they make contact with just a small portion of them and we have low current flow or whether the fat vocal folds actually make complete contact and we have a high current flow (ICE-GB S2A-056_60:1:A)

(2) I 've given it to another friend of mine who said she 's certainly going to keep it till the summer

(ICE-GB S1A-022_163:1:D)

Gesproken Brits-Engels: modelanalyse

Meest overwachte gevallen:

(3) because they gave the musical world the opportunity to reassess La Finta Giardiniera

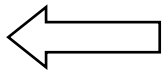
(ICE-GB S1B-044_44:1:B)

(4) But I lent a book on painting to her

(ICE-GB S1A-013_161:1:C)

Geschreven Brits-Engels: Resultaten

Trainingset	model-fit	testset	voorspellingskwaliteit
Gesproken	**94.3%**	Geschr.Br.	**85.4%**
Geschr.Br.	**93.0%**	Geschr.Br.	**88.1%**
Gesproken sub	**95.6%**	Geschr.Br.	**80.1%**



Geschreven Brits-Engels: Resultaten

Harmonische oplijning

Dubbel-objectconstructie

hand **the student** *the book*

ontvanger < *thema*
discourse gegeven niet-gegeven
animate

Prepositionele datiefconstructie

hand **the book** to *the student*

thema < *ontvanger*
discourse gegeven niet-gegeven
inanimate

Geschreven Brits-Engels: modelanalyse

Prototypische gevallen:

(1) to give you an opportunity to elect two representatives for 20th March, and decide on any matters you wish to put to the national meeting (ICE-GB W1B-024_6:1)

(2) will be giving it mainly to first aiders and appointed persons (ICE-GB W1B-018_20:4)

Geschreven Brits-Engels: modelanalyse

Meest overwachte gevallen:

(3) to give the system headroom (ICE-GB W2B-038_45:1)

(4) actually write a letter to you (ICE-GB W1B-002_3:1)

Conclusie

- De datiefalternantie in het Amerikaans- en Brits-Engels is goed te modelleren met behulp van regressie
- De gevonden modellen laten dezelfde ‘harmonische oplijning’ zien als die gevonden in de literatuur
- De datiefalternantie lijkt hetzelfde te werken voor gesproken Amerikaans- en Brits-Engels
- Voor geschreven Brits-Engels lijkt het model voor gesproken Engels niet zo goed te werken

Toekomstig onderzoek

- Het effect van genre
- Het modelleren van de ‘weggelaten’ gevallen, bijvoorbeeld die in passieve zin
- Modelleren met een andere techniek: Bayesiaanse Netwerken

Computerlinguïstiek over 20 jaar

Een toekomstdroom of werkelijkheid?

- Genoeg data om alle taalkundige theorieën te testen
- *Armchair linguists* zijn ‘uitgestorven’
- Nog geavanceerdere technieken die ingewikkelde datasoorten (zoals taal) feilloos kunnen analyseren
- Samenwerking tussen de verschillende vakgebieden (taalverwerving, psycholinguïstiek, corpuslinguïstiek, …)

Vragen?

Referenties

- Baayen, R. H. (in press). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.
- Bates, D. 2005. Fitting linear mixed models in R. *R News*, 5 (1): 27-30.
- Biship, C.M. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Bresnan, J., A. Cueni, T. Nikitina and R.H. Baayen 2007. Predicting the Dative Alternation. In Bouma, G, I. Kraemer and J. Zwarts (eds.), *Cognitive Foundations of Interpretation:* 69-94. Amsterdam: Royal Netherlands Academy of Science.
- De Marneffe, M-C, S. Grimm, U.C. Priva, S. Lestrade, G. Ozbek, T. Schnoebelen, S. Kirby, M. Becker, V. Fong and J. Bresnan 2007. A Statistical Model of Grammatical Choices in Childrens' Productions of Dative Sentences. Presented at FAVS 2007, York, UK.
- Godfrey, J., E. Holliman and J. McDaniel 1992. Switchboard: Telephone speech corpus for research and development. *Proceedings of ICASSP-92*, San Francisco: 517-20.
- Greenbaum, Sidney (ed.) 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.
- Gries, S. Th. 2003. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1: 1-27.
- Gries, S. Th. and A. Stefanowitsch 2004. Extending Collostructional Analysis: A Corpus-based Perspective on ‘Alternations’. *International Journal of Corpus Linguistics* 9: 97-129.