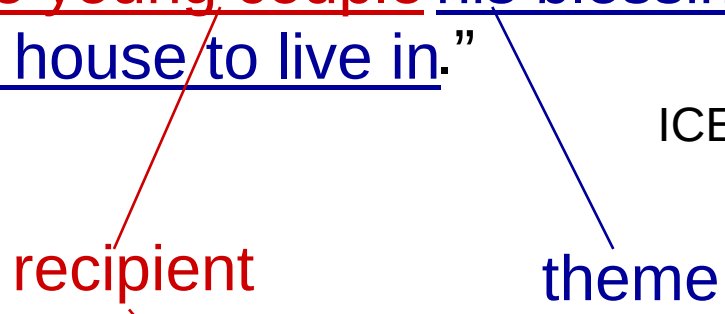
Using the ICE-GB corpus to model the English dative alternation

Daphne Theijssen
PhD student
Department of Linguistics
Radboud University Nijmegen
d.theijssen@let.ru.nl

Examples of dative alternation

- “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

Examples of dative alternation

- “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.” ?

Examples of dative alternation

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

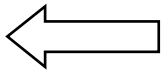
- “<It's really> it used to be given as fourteenth-century <redding> wedding rings and nowadays blokes give it to girlfriends”

ICE-GB S1A-047_216:1:B

Examples of dative alternation

- “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

Question

Can we predict the dative alternation?

This presentation

- Related work by Bresnan et al. 2007
- Research goals
  - Apply existing model to more varied data (ICE-GB)
  - Extend with syntactic variables
- Experimental setup
- Goal 1: Varied written and spoken text
- Goal 2: Extending the model
- Concluding remarks
- Questions

Related work by Bresnan et al. (2007)

- 2360 instances from Switchboard (Godfrey et al. 1992)
- Linear regression modelling
  - Variables taken from previous literature
  - Predicted 95.0% of the data correctly (5.0% uneplained)
- Added written data (financial texts)
  - 905 instances from Wall Street Journal (Penn Treebank)
  - 93.4% predicted correctly
- Added child language (De Marneffe et al. 2007)
  - 530 instances from CHILDES database
  - 95.7% predicted correctly

Research goals

1. Applying Bresnan et al.'s GLM (2007) to a corpus showing more variation in text genre

2. Extending the model with syntactic features

Experimental setup: Data

- Syntactically annotated ICE-GB corpus (Greenbaum 1996)
- spoken texts
  - dialogues (private and public)
  - monologues (unscripted and scripted)
- written texts
  - non-printed (student writing and letters)
  - printed (academic, popular, reportage, instructional, persuasive and creative)
- Find cases with Perl script

Experimental setup: Data

- Excluded (following Bresnan et al.):
  - **preposition other than *to***

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

    S1A-013_118:1:A
  - **passivized object as subject**

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

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

    S2B-038_4:1:A
  - **heavy NP shift**

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

Experimental setup: Data

- Also excluded:
  - **coordinated verbs or verb phrases**

    e.g. “However, anyone caught importing or supplying large quantities of the drug to others will invariably be prosecuted.”

    W2B-020_47:1
  - **phrasal and particle verbs**

    e.g. “I 'll send you out that” S1B-074_46:1:B
  - **all cases with verbs with only NP-NP *or* NP-PP**

    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

- Result: 919 cases

Experimental setup: LMER

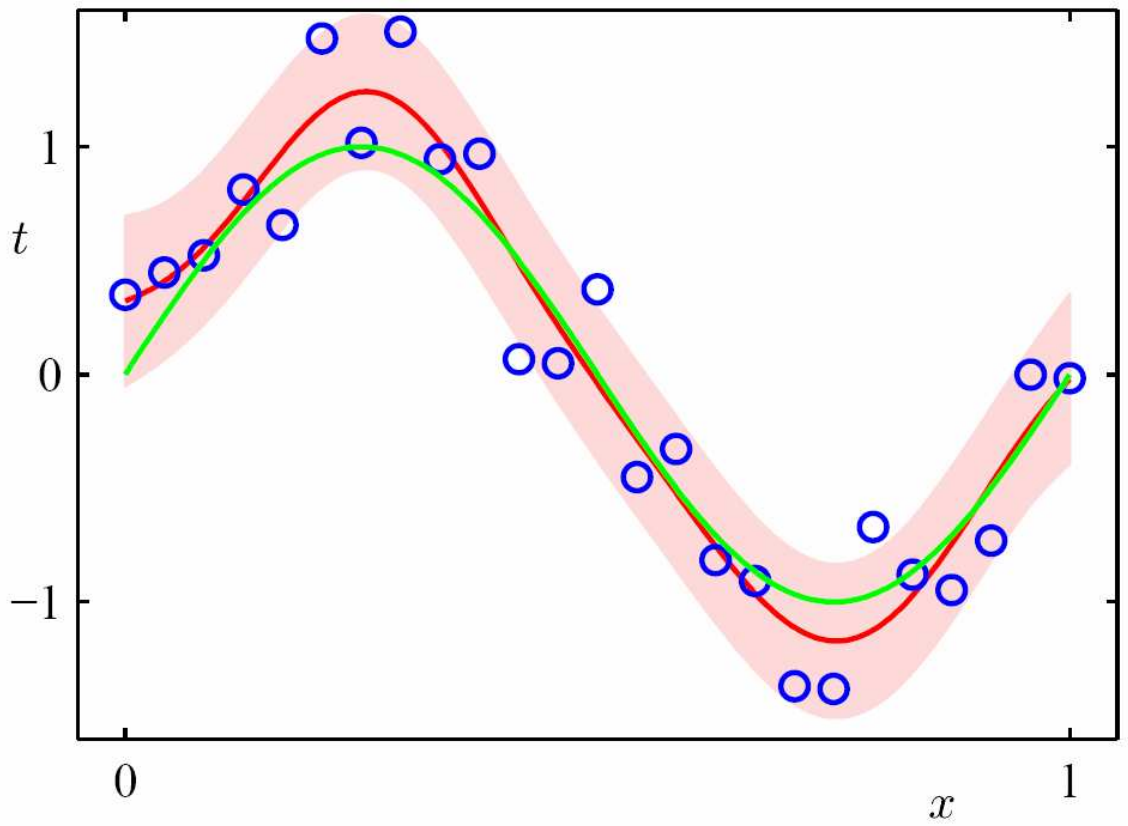
- Linear Mixed-Effect Modelling (Bates 2005)
  - Fixed effexts: variables
  - Random effect: verb sense

- Verb sense => assume lexical bias

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

- Analyzing the model
  - Use coefficients to determine which variables show significant effects in the dative alternation model
  - Evaluate the model fit (% of correctly predicted cases)

Experimental setup: LMER



Source: Biship (2006)

Suggestion for futher reading: Baayen (in press)

Goal 1: Variables

- Pronominality of recipient + theme (pronominal, non-pronominal)
- Definiteness of recipient + theme (definite, indefinite)
- Animacy of recipient + *theme* (animate, inanimate)
- Person of recipient + *theme* (local, non-local)
- Number of recipient + theme (singular, plural)
- Concreteness of *recipient* + theme (concrete, inconcrete)
- *Discourse accessibility of recipient + theme (given, new)*
- Length difference between the theme and the recipient (log scale)
- Semantic verb class (abstact, communication, transfer of possession, future transfer of possession, prevention of transfer)
- *Structural parallelism (yes, no)*

Goal 1: Results

Classification table for SWB and ICE

		SWB: predicted			*ICE: predicted*		
		NP-NP	NP-PP	% correct	NP-NP	NP-PP	% correct
observed	NP-NP	1808	51	97.3%	673	34	95.2%
	NP-PP	79	422	84.2%	51	161	75.9%
			overall:	**94.5%**		*overall:*	**90.8%**

% correct from always guessing NP-NP: 80.0% (SWB) and 78.8% (ICE)

Goal 1: Results

Significant effects in SWB

variable + value	*direction*	*coefficient*	*z-value*	*significance*	*level*
Pronominality of theme	NP-PP	2,34	7,89	2,96E-15	***
Inanimacy of recipient	NP-PP	1,67	3,55	3,79E-04	***
Indefiniteness of recipient	NP-PP	1,32	4,13	3,63E-05	***
Non-local person of recipient	NP-PP	0,54	1,99	4,66E-02	*
Singular number of theme	NP-NP	-0,81	-3,19	1,42E-03	**
Length difference (log)	NP-NP	-1,58	-9,12	2,00E-16	***
Pronominality of recipient	NP-NP	-2,28	-7,78	7,18E-15	***
Indefiniteness of theme	NP-NP	-2,37	-9,43	2,00E-16	***

Goal 1: Results

Significant effects in ICE

variable + value	*Direction*	*coefficient*	*z-value*	*significance*	*level*
Non-local person of recipient	NP-PP	1,52	4,02	5,86E-05	***
Indefiniteness of recipient	NP-PP	1,33	3,36	7,83E-04	***
Indefiniteness of theme	NP-NP	-0,99	-3,19	1,43E-03	**
Pronominality of recipient	NP-NP	-1,05	-3,03	2,43E-03	**
Length difference (log)	NP-NP	-1,73	-8,51	2,00E-16	***

Goal 2: Extending the model

- Clause properties
  - Mode (declarative, interrogative, imperative)
  - Word order (unmarked, fronting)
  - Type of dependent clause (clausal, phrasal)
  - Importance of clausal dependent clause (adjunct, complement)
- Intervening adverbials

  e.g. “Ukraine lacks oil, but much Soviet oil comes from the Transcaucasian republics, now also aspiring to independence, which could try to bypass Moscow by selling oil directly to Ukrainian nationalists.”

  ICE-GB W2C-008_20:1

  - Length in words
  - Length in characters

Goal 2: Results

Classification table for ICE (with syntax)

		ICE: predicted with syntax added		
		NP-NP	NP-PP	% correct
observed	NP-NP	674	33	95.3%
	NP-PP	41	171	80.7%
			overall:	**91.9%**

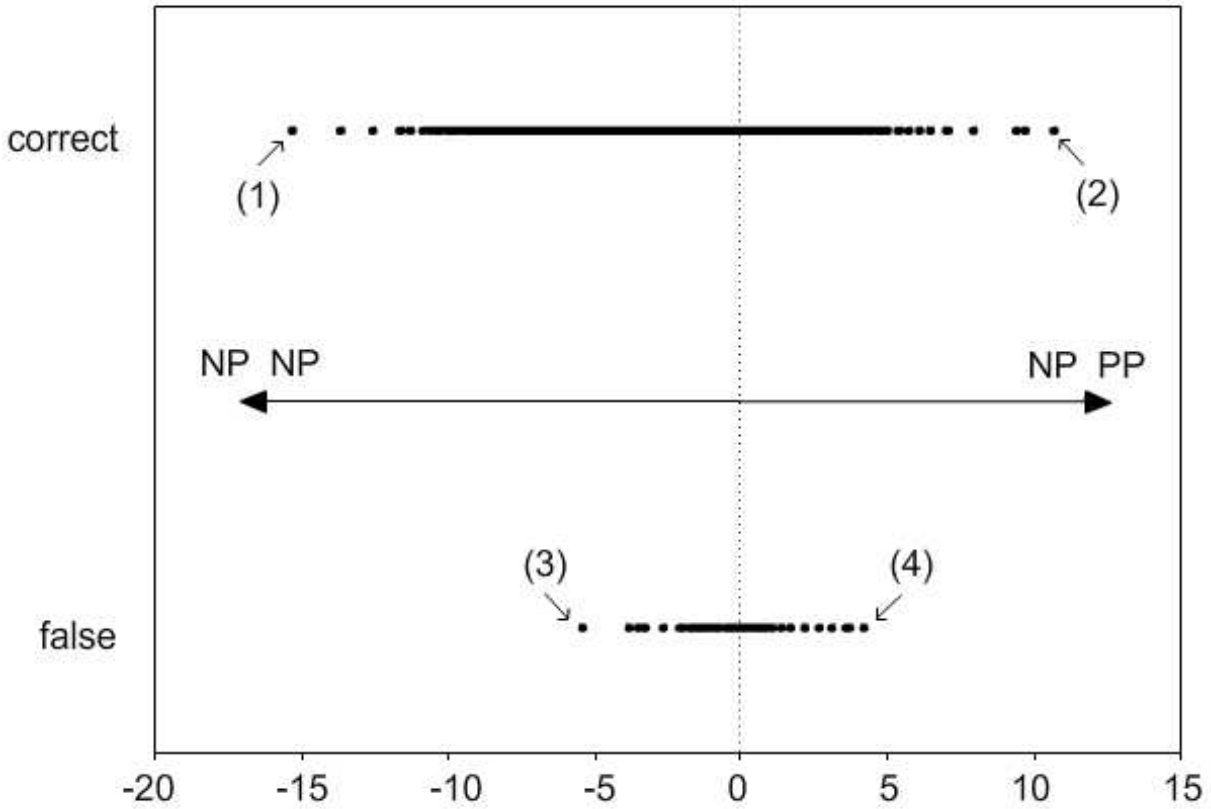
% correct from always guessing NP-NP: 78.8%

Goal 2: Results

Significant effects in ICE (with syntax)

variable + value	*direction*	*coefficient*	*z-value*	*significance*	*level*
Unmarked word order	NP-PP	2,22	3,23	1,25E-03	**
Non-local person of recipient	NP-PP	1,60	3,91	9,40E-05	***
Indefiniteness of recipient	NP-PP	1,56	3,72	2,02E-04	***
Pronominality of theme	NP-PP	1,18	2,57	1,02E-02	*
Pronominality of recipient	NP-NP	-1,11	-3,03	2,44E-03	**
Indefiniteness of theme	NP-NP	-1,18	-3,62	3,00E-04	***
Length difference (log)	NP-NP	-1,90	-8,71	2,00E-16	***

Goal 2: Error analysis



Graph design based on Gries (2003)

Goal 2: Error analysis

Cases that are classified correctly:

(1) You have given me you and you have restored to me myself.
(ICE-GB W1B-006_16:1)

(2) And secondly I obviously can't do justice in sus in such a short time <,> to the exposition of the ways in which this theory differed from other views at the time <,,>
(ICE-GB S2B-049_5:1:A)

Goal 2: Error analysis

Cases that are classified incorrectly:

(3) But why on earth should <,> why on earth should Mr Neil make that comment unless Mr <,> uh Slipper had given the appearance to him uh of uh ignorance of the extradition treaty

(ICE-GB S2A-064_82:2:A)

(4) So I think uh Perez de Cuellar has probably been prevailed on to uh to to come out with some kind of platitude that will uh give all these reporters who were sitting around here all day waiting for something to happen something to report

(ICE-GB S2B-010_86:1:B)

Concluding remarks

- Proportion of correctly predicted constructions for ICE was lower (90.8%) than that for SWB (94.5%): text type affects performance (or fit) of the model?

  Future: text type as additional variable (provided that the data is not too sparse)
- Possible other causes for the lower prediction accuracies
  - annotation differences
  - ICE-GB corpus is British English, Switchboard is American English
  - certain variables had to be ignored (only mutual variables included)

  Future (completed variable set): establish benefit of syntactic variables again and apply SWB model (including its coefficients) to ICE and vice versa
- word order has significant effect in ICE and split objects are difficult to model

  Future: ask ourselves whether we want to model according to traditional variants (NP-NP and NP-PP), or the ordering of theme and recipient.

References

- 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.

Questions?

Text Genre

correct		total	description
100.00%	12	12	written non-printed student writing (essays, exam scripts)
97.78%	44	45	written printed creative (novels)
94.76%	199	210	spoken dialogues private (conversations, phonecalls)
94.49%	120	127	spoken monologues unscripted (commentaries, legal presentations)
93.75%	60	64	spoken monologues scripted (e.g. broadcast news)
91.30%	42	46	written printed popular (various fields)
90.06%	163	181	spoken dialogues public (e.g. class lessons, parliamentary debates)
89.08%	106	119	written non-printed letters (social letters, business letters)
88.57%	31	35	written printed reportage (press reports)
87.10%	27	31	written printed academic (various fields)
85.71%	30	35	written printed instructional (administrative writing, skills/hobbies)
78.57%	11	14	written printed persuasive (editorials)