Modelling the English dative alternation in varied written and spoken text

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

The English dative alternation

Spoken dialogue:

Can you give us an example? vs.
Can you give an example to us?

I 'm reading it Treasure Island at the moment to my son. vs.
At the moment, I'm reading my son it, Treasure Island.

Dictionary on popular phrases:

… from that I coined the term 'lager lout' to give it more meaning. vs.
… from that I coined the term 'lager lout' to give more meaning to it.

The species was identified in 1988, the name clearly owing much to alliteration. vs.
The species was identified in 1988, the name clearly owing alliteration much.

The English dative alternation

- Syntactic approach

  e.g. Quirk et al. (1972): “Indirect objects are typically **animate**.”

- Semantic approach

  e.g. Gries and Stefanowitsch (2004): “… the ditransitive should prefer verbs of **direct face-to-face transfer**, while the *to*-dative should prefer verbs of **transfer over distance**.”

- Discourse approach

  e.g. Collins (1995): “… strong likelihood that a receiver NP will be **informationally given** and an entity NP **informationally new** in the indirect object construction”

Aim

Our aim is to develop a model that integrates existing theories about the dative alternation.

Two steps:

1. apply existing model (Bresnan et al. 2007) to data with more diversity in genre
2. extend it with new variables

This presentation

- Experimental setup
- Step 1: Existing model on data with diversity in genres
- Step 2: Extending the model with new variables
- Summary
- Future research

Experimental setup: Existing models

- Logistic mixed-effect regression (LMER) modelling in Bresnan et al. (2007)
  - 2360 instances from Switchboard (Godfrey et al. 1992)
  - Variables taken from existing literature
  - Predicted 95.0% of the data correctly (5.0% unexplained)

- 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

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)

Extract 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 <,>
  - passivized object as subject
    e.g. “Dido 's pride has been dealt a severe blow .”
  - clausal object
    e.g. “so doctors will tell you that they've only just discovered this idea”
  - heavy NP shift
    e.g. “lending to the houses and pedestrians a faintly unreal or even theatrical quality”

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.”
  - phrasal and particle verbs

    e.g. “I'll send you out that”
- Only include verbs present with both constructions (919):

  e.g. “who bears no resemblance at all to Cathy now”

Experimental setup: Data

- SWB: 2360 cases in telephone conversations
  => Bresnan et al. (2007)'s data from Switchboard (Godfrey et al. 1992)

- ICE: 919 cases in varied spoken and written text
  => our data from the ICE-GB corpus (Greenbaum 1996)

Experimental setup: LMER

- Linear Mixed-Effect Modelling (Bates 2005)
  - Fixed effects: 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

Experimental setup: LMER

- p = probability that observed case is NP-PP

- $p = \frac{e^{g(x)}}{1 + e^{g(x)}}$ (logistic function)

- $g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

  $\beta_0 + \ldots + \beta_k$ = coefficients

  $X_1 + \ldots + X_k$ = variables

Step 1: Variables

	recipient	theme
– Pronominality	x	x
– Definiteness	x	x
– Animacy	x	*(x)*
– Person	x	*(x)*
– Number	x	x
– Concreteness	*(x)*	x
– *(Discourse accessibility)*	*(x)*	*(x)*

- Length difference between the theme and the recipient
- Semantic verb class
- *(Structural parallelism)*

Step 1: Results

	SWB	ICE
Majority baseline	80.0%	78.8%
% correctly predicted	94.7%	90.8%
Somers' C	98.0%	95.5%

Step 1: Results

Significant effects with their coefficients β

	Variable + value (X)	*β SWB*	*β ICE*
NP-PP	theme = pronominal	2.3	-
	recipient = inanimate	1.7	-
	recipient = indefinite	1.3	1.3
	recipient = non-local	0.5	1.5
	theme = singular	-0.8	-
	length difference (th - rec)	-1.6	-1.2
	recipient = pronominal	-2.3	-1.0
	theme = indefinite	-2.4	-1.0
	future transfer of possession	-	-2.9
NP-NP	prevention of possession	-5.9	-

Step 2: Extending the model

Syntactic variables

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

- Intervening adverbials

  e.g. “to bypass Moscow by selling oil directly to Ukrainian nationalists.”
  - *Length in words*
  - *Length in characters*

Step 2: Extending the model

Collostructional analysis (Gries and Stefanowitsch 2004)

Fisher exact test on the distribution of verb *A* and all verbs but *A* to determine expected values of NP-NP and NP-PP for *A.*

Establish significant bias for verb senses in our data

- *Collostructional strength*

  bias towards NP-PP: $1 - p_{fisher}$

  bias towards NP-NP: $-1 * (1 - p_{fisher})$

– a value near 1 indicates the verb has a strong preference for NP-PP
– a value near 0 indicates the verb can be freely applied in both constructions
– a value near -1 indicates the verb has a strong preference for NP-NP

Step 2: Results

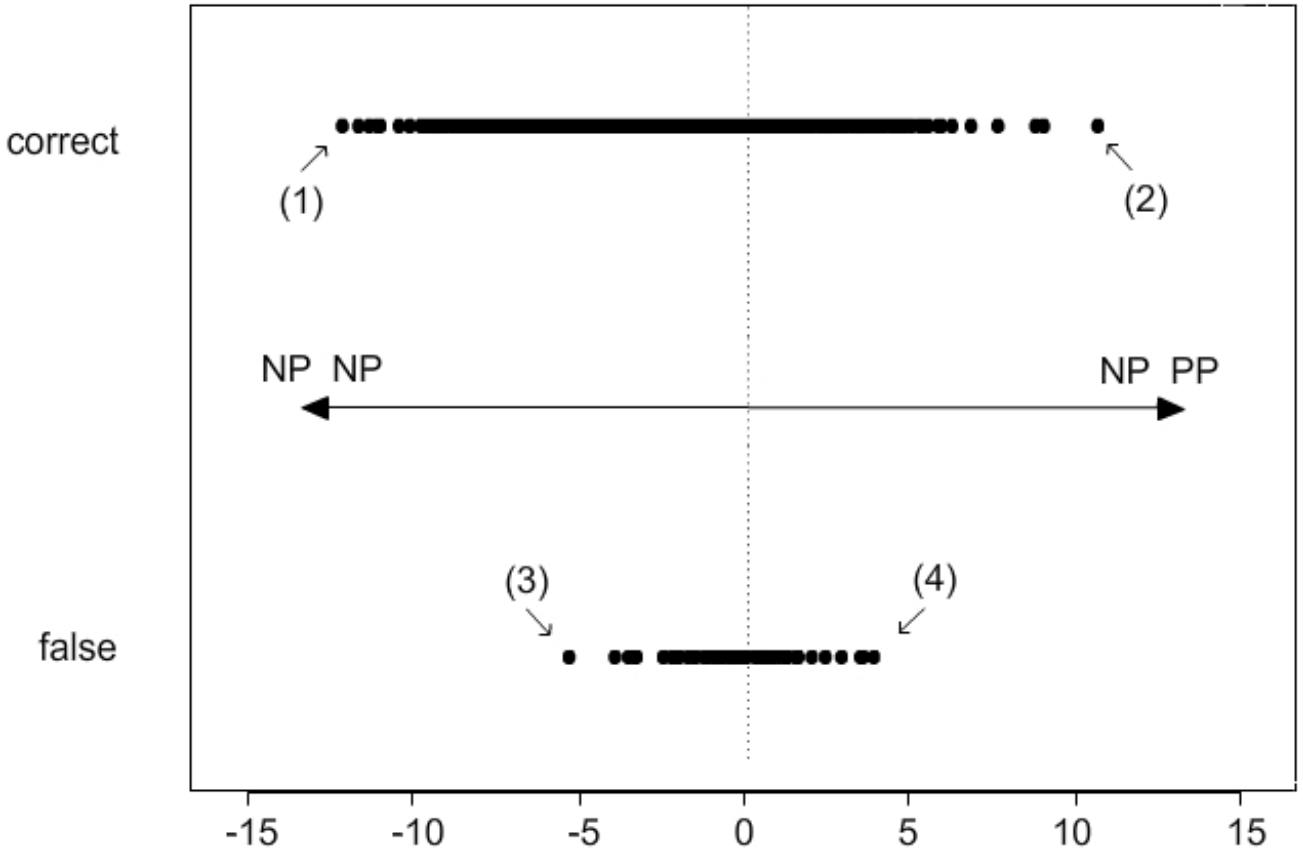
	ICE original	ICE extended
Majority baseline	78.8%	78.8%
% correctly predicted	90.8%	91.6%
Somers' C	95.5%	95.9%

Step 2: Results

Significant effects in ICE with their coefficients β

	Variable + value (X)	*β org.*	*β ext.*
NP-PP	unmarked word order	x	2.0
	collostructional strength	x	2.0
	recipient = non-local	1.5	1.5
	recipient = indefinite	1.3	1.4
	theme = pronominal	-	1.1
	recipient = pronominal	-1.0	-1.1
	theme = indefinite	-1.0	-1.2
	length difference (th - rec)	-1.2	-1.3
NP-NP	future transfer of possession	-2.9	-

Step 2: Result analysis



Graph design based on Gries (2003)

Step 2: Result analysis

Cases that are classified correctly:

(1) … my merely telling you two facts <,,> that Johnny went to Sam 's party and Sam blew out the candles ...

(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 <,,>

Cases that are classified incorrectly:

(3) … unless Mr <,> uh Slipper had given the appearance to him uh of uh ignorance of the extradition treaty

(4) … what independent dance or dance <,> of this nature offers the other <,> activities which would normally be associated with the d disabled people …

Summary

Step 1: Existing model on data with diversity in genres

- Significant effects in ICE showed same patterns as in SWB

- Proportion of correctly predicted constructions for ICE lower (90.8%) than for SWB (94.5%)

  Possible causes:

  - Diversity in genre affects performance (or fit) of the model
  - annotation differences
  - ICE-GB corpus is British English, Switchboard is American English
  - certain variables had to be ignored

Summary

Step 2: Extending the model with new variables

- Unmarked word order and collostructional strength significant effects with tendency towards NP-PP
- Proportion of correctly predicted constructions higher (91.6%) than original model (90.8%)

Future research

- Complete variable set
  => establish benefit of new variables again

- Apply SWB model with its coefficients to ICE and vice versa

- Include genre as a separate variable

- Word order has significant effect, split objects are difficult to model
  => ask ourselves: should we model the variants NP-NP and NP-PP, or theme-recipient and recipient-theme?

Thank you!

This presentation will be available on http://lands.let.ru.nl/~daphne after the conference.

References

- Bates, D. 2005. Fitting linear mixed models in R. *R News*, 5 (1): 27-30.
- 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.
- Collins, P. 1995. The indirect object construction in English: an informational approach. *Linguistics* 33: 35-49.
- 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 Children's' 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.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik 1972. *A Grammar of Contemporary English*. London: Longman.

Text Genre

accuracy	correct	total	description
100.0%	12	12	written non-printed student writing (essays, exam scripts)
97.8%	44	45	written printed creative (novels)
96.1%	122	127	spoken monologues unscripted (e.g. commentaries, legal presentations)
93.8%	197	210	spoken dialogues private (conversations, phonecalls)
93.8%	60	64	spoken monologues scripted (e.g. broadcast news)
89.9%	107	119	written non-printed letters (social letters, business letters)
89.5%	162	181	spoken dialogues public (e.g. class lessons, parliamentary debates)
89.1%	41	46	written printed popular (various fields)
88.6%	31	35	written printed reportage (press reports)
87.1%	27	31	written printed academic (various fields)
80.0%	28	35	written printed instructional (administrative writing, skills/hobies)
78.6%	11	14	written printed persuasive (editorials)

Confusion matrices

		SWB: predicted			*ICE: predicted*		
		NP-NP	NP-PP	% correct	NP-NP	NP-PP	% correct
observed	NP-NP	1809	50	97.3%	673	34	95.2%
	NP-PP	75	426	85.0%	51	161	75.9%
			overall:	**94.7%**		*overall:*	**90.8%**
			Somers' C:	98.0%		*Somers' C:*	95.5%

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

		ICE original: predicted			*ICE extended: predicted*		
		NP-NP	NP-PP	% correct	NP-NP	NP-PP	% correct
observed	NP-NP	673	34	95.2%	675	32	95.5%
	NP-PP	51	161	75.9%	45	167	78.8%
			overall:	**90.8%**		*overall:*	**91.6%**
			Somers' C:	95.5%		*Somers' C:*	95.9%

% correct from always guessing NP-NP: 78.8%

Effects: SWB

	Estimate	Odds ratio	Std.Error	z	value	Pr(>|z|)
(Intercept)	1.3569	3.88	0.8615	1.575	0.115237	
PronomOfRecp	-2.2672	0.10	0.291	-7.792	6.58E-15	***
PronomOfThemep	2.2981	9.96	0.2948	7.795	6.46E-15	***
DefinOfRecin	1.3373	3.81	0.316	4.233	2.31E-05	***
DefinOfThemein	-2.4102	0.09	0.2505	-9.62	2.00E-16	***
AnimacyOfRecin	1.6821	5.38	0.4718	3.566	0.000363	***
PersonOfRecnon-local	0.5328	1.70	0.2696	1.976	0.048102	*
NumberOfRecsingular	0.2158	1.24	0.2276	0.948	0.343044	
NumberOfThemesingular	-0.7977	0.45	0.2531	-3.152	0.001623	**
ConcreteOfThemen	0.4367	1.55	0.2957	1.477	0.139652	
LengthDiffWords	-1.5952	0.20	0.1737	-9.185	2.00E-16	***
SemanticClassc	-1.1782	0.31	1.1453	-1.029	0.303595	
SemanticClassf	-0.7943	0.45	1.3628	-0.583	0.559992	
SemanticClassp	-5.9263	0.00	2.9317	-2.021	0.04323	*
SemanticClasst	0.8914	2.44	0.9959	0.895	0.37074	

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Effects: ICE original

	Estimate	Odds ratio	Std.Error	z	value	Pr(>|z|)
(Intercept)	-0.04256	0.96	0.92618	-0.046	0.96335	
PronomOfRecp	-1.02864	0.36	0.34246	-3.004	0.002667	**
PronomOfThemep	0.51167	1.67	0.39153	1.307	0.191265	
DefinOfRecin	1.31558	3.73	0.39664	3.317	0.00091	***
DefinOfThemein	-1.04606	0.35	0.30405	-3.44	0.000581	***
AnimacyOfRecin	0.51706	1.68	0.37211	1.39	0.164672	
PersonOfRecnon-local	1.49715	4.47	0.37269	4.017	5.89E-05	***
NumberOfRecsingular	0.28003	1.32	0.31025	0.903	0.366742	
NumberOfThemesingular	0.19975	1.22	0.36341	0.55	0.582547	
ConcreteOfThemein	-0.73044	0.48	0.47179	-1.548	0.121565	
LengthDiffWords	-1.19449	0.30	0.14104	-8.469	2.00E-16	***
SemanticClassc	0.15477	1.17	0.82728	0.187	0.851599	
SemanticClassf	-2.93811	0.05	1.33088	-2.208	0.027269	*
SemanticClasst	0.54148	1.72	0.85003	0.637	0.524112	

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Effects: ICE extended

	Estimate	Odds ratio	Std.Error	z	value	Pr(>|z|)
(Intercept)	-2.6714	0.07	1.405	-1.901	0.057253	.
PronomOfRecp	-1.0805	0.34	0.3564	-3.032	0.002431	**
PronomOfThemep	1.0667	2.91	0.4435	2.405	0.016166	*
DefinOfRecin	1.434	4.20	0.4159	3.448	0.000565	***
DefinOfThemein	-1.2233	0.29	0.32	-3.823	0.000132	***
AnimacyOfRecin	0.4948	1.64	0.3842	1.288	0.19782	
PersonOfRecnon-local	1.5158	4.55	0.3977	3.812	0.000138	***
NumberOfRecsingular	0.1842	1.20	0.3248	0.567	0.570724	
NumberOfThemesingular	0.4273	1.53	0.3809	1.122	0.261983	
ConcreteOfThemein	-0.6024	0.55	0.4694	-1.283	0.19934	
LengthDiffWords	-1.2584	0.28	0.1473	-8.543	2.00E-16	***
SemanticClassc	0.5063	1.66	0.6484	0.781	0.434896	
SemanticClassf	-1.9267	0.15	1.1688	-1.649	0.099248	.
SemanticClasst	-0.2671	0.77	0.6934	-0.385	0.700131	
ClauseModeim	0.5336	1.71	0.5728	0.932	0.351564	
ClauseModein	-1.2732	0.28	0.7976	-1.596	0.110435	
ClauseWordOrderu	2.0041	7.42	0.6726	2.98	0.002884	**
DepClauseTypena	0.8945	2.45	1.0971	0.815	0.414852	
DepClauseTyper	0.7603	2.14	1.043	0.729	0.466019	
DepClauseTypes	-0.2091	0.81	1.0267	-0.204	0.838596	
ClDepClauseImportancec	0.2149	1.24	0.4702	0.457	0.647733	
ClDepClauseImportancena	-0.8363	0.43	0.5699	-1.468	0.142238	
LengthAdverbWords	0.9055	2.47	0.8645	1.047	0.294886	
LengthAdverbChars	-0.1138	0.89	0.2136	-0.533	0.594244	
CollostrStrength	1.9521	7.04	0.361	5.408	6.38E-08	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Collostructional Analysis

- significant bias ($p_{fisher} < 0.05$):
  - 11 of 38 verb senses (28.9%)
  - 494 of 919 cases (53.8%)

- Predict construction in these 494 cases:
  - Majority baseline: 81.1%
  - Collostructional Analysis: 85.6%