Making choices

Methodological issues in corpus studies of syntactic alternations

Daphne Theijssen

PhD student
Department of Linguistics
Radboud University Nijmegen
d.theijssen@let.ru.nl
—
Fulbright visiting scholar
Department of Linguistics
Stanford University

November 12, 2010

Making choices...

Making choices...

Overview

- Issues in collecting corpus data
- Issues in coding corpus data
- Issues in modelling corpus data
- Conclusions

Issues in collecting data

Overview

- corpus limitations
- infrequent alternations
- automatic approaches
- what is a suitable instance?

Issues in collecting data

Overview

- corpus limitations
- infrequent alternations
- automatic approaches
- what is a suitable instance?

Issues in collecting data

Overview

- corpus limitations
- infrequent alternations
- automatic approaches
- what is a suitable instance?

Issues in collecting data

Overview

- corpus limitations
- infrequent alternations
- automatic approaches
- what is a suitable instance?

Issues in collecting data

Overview

- corpus limitations
- infrequent alternations
- automatic approaches
- what is a suitable instance?

Issues in collecting data

Example: Infrequent alternations

Benefactive alternation

Daphne Theijssen, Hans van Halteren, Karin Fikkers, Frederike Groothoff, Lian van Hoof, Eva van de Sande, Jorieke Tiems, Véronique Verhagen and Patrick van der Zande (2009). A regression model for the English benefactive alternation. Barbara Plank, Erik Tjong Kim Sang and Tim Van de Cruys (eds.), *Computational Linguistics in the Netherlands 2009*, pp. 115-130.

Issues in collecting data

Example: Infrequent alternations

Research questions:

1. Do the same factors play a role in the dative and the benefactive alternation, with the same strength?
2. How does the benefactive alternation in adult language relate to that in child language?

Corpora (nr of words in written language):

- adult: ICE-GB (423.500) and SUSANNE (130.000)
- child: LCCPW (78.500) and LUCY (30.000)

Issues in collecting data

Example: Infrequent alternations

Research questions:

1. Do the same factors play a role in the dative and the benefactive alternation, with the same strength?
2. How does the benefactive alternation in adult language relate to that in child language?

Corpora (nr of words in written language):

- adult: ICE-GB (423.500) and SUSANNE (130.000)
- child: LCCPW (78.500) and LUCY (30.000)

Issues in collecting data

Example: Infrequent alternations

After automatic extraction: 2,277 candidates

- Adult: 972
- Child: 1,305

After manual filtering: 143 cases

- Adult: 107, of which 70.1% (75) NP-PP
- Child: 36, of which 66.7% (24) NP-PP

Issues in collecting data

Example: Infrequent alternations

After automatic extraction: 2,277 candidates

- Adult: 972
- Child: 1,305

After manual filtering: 143 cases

- Adult: 107, of which 70.1% (75) NP-PP
- Child: 36, of which 66.7% (24) NP-PP

Issues in collecting data

Example: Infrequent alternations

New research questions

1. Can we predict the benefactive in written adult English based on features for the dative alternation?
2. Can we also predict the benefactive in English writing by 6-12 year-olds with the model we find?

Issues in collecting data

Example: Infrequent alternations

Adult model

- majority baseline: 70.1%
- prediction train=test: 86.9%
- prediction 10-fold cv: 79.6%

Variable	Coefficient	
Intercept	8.1	**
Length Difference (th-ben)	-2.3	**
Theme = singular	-2.6	**
Semantic Verb Class = obtain	-2.8	*
Theme = not given	-3.0	*

Issues in collecting data

Example: Infrequent alternations

Adult model

- majority baseline: 70.1%
- prediction train=test: 86.9%
- prediction 10-fold cv: 79.6%

Variable	Coefficient	
Intercept	8.1	**
Length Difference (th-ben)	-2.3	**
Theme = singular	-2.6	**
Semantic Verb Class = obtain	-2.8	*
Theme = not given	-3.0	*

Issues in collecting data

Example: Infrequent alternations

Adult model applied to child data

- majority baseline: 66.7%
- prediction: 80.6%

Issues in data coding

Overview

- selecting factors of interest
- inter-annotator agreement
- definitions / coding

Issues in data coding

Overview

- selecting factors of interest
- inter-annotator agreement
- definitions / coding

Issues in data coding

Overview

- selecting factors of interest
- inter-annotator agreement
- definitions / coding

Issues in data coding

Overview

- selecting factors of interest
- inter-annotator agreement
- definitions / coding

Issues in collecting data

Example: IA agreement

Benefactive and dative alternation

Daphne Theijssen, Hans van Halteren, Karin Fikkers, Frederike Groothoff, Lian van Hoof, Eva van de Sande, Jorieke Tiems, Véronique Verhagen and Patrick van der Zande (2009). A regression model for the English benefactive alternation. Barbara Plank, Erik Tjong Kim Sang and Tim Van de Cruys (eds.), *Computational Linguistics in the Netherlands 2009*, pp. 115-130.

Daphne Theijssen, Lou Boves, Hans van Halteren en Nelleke Oostdijk (under review). The computer as annotator: Automatically detecting and annotating syntactic constructions. Submitted to *Language Resources and Evaluation*.

Daphne Theijssen, Joan Bresnan and Marilyn Ford (in preparation).

Issues in collecting data

Example: IA agreement

Data

- Benefactive: 143 instances, coded by at least two annotators
- Dative US: 30 instances from Switchboard, coded by two annotators
- Dative UK: 40 instances from ICE-GB, coded by two annotators

Issues in data coding

Example: IA agreement

Feature	Benef.	Dativ.US	Dativ.UK
Person of recipient	0.86	0.91	1.00
Pronominality of recipient	0.80	1.00	0.95
Definiteness of theme	**0.78**	0.93	1.00
Number of recipient	0.92	1.00	**0.77**
Number of theme	**0.71**	1.00	0.88
Definiteness of recipient	**0.74**	1.00	**0.78**
Pronominality of theme	**0.39**	1.00	0.84
Concreteness of theme	**0.55**	0.86	**0.75**
Givenness of recipient	**0.55**	**0.60**	0.95
Animacy of recipient	0.87	**0.53**	**0.63**
Givenness of theme	**0.32**	**0.66**	0.80

Issues in data coding

Example: Definitions / coding

Concreteness of dative themes

Daphne Theijssen, Hans van Halteren and Lou Boves (2010). More or less concrete? Comparing approaches to establish the concreteness of nouns. Unpublished manuscript.

Issues in data coding

Example: Definitions / coding

Definitions of concreteness (Spreen and Schulz 1966)

1. 'specificity' (originated in cognitive and neuroscience):
   abstract nouns are general, generic, not specific
2. 'physical perceivability' (more common in linguistics, e.g. Lyons 1977):
   abstract nouns lack sense experience

"Strictly speaking, what is abstract is not the nouns themselves, but what they denote" (Schmid 2000)

Issues in data coding

Example: Definitions / coding

Definitions of concreteness (Spreen and Schulz 1966)

1. 'specificity' (originated in cognitive and neuroscience):
   abstract nouns are general, generic, not specific
2. 'physical perceivability' (more common in linguistics, e.g. Lyons 1977):
   abstract nouns lack sense experience

"Strictly speaking, what is abstract is not the nouns themselves, but what they denote" (Schmid 2000)

Issues in data coding

Example: Definitions / coding

Levels of coding

- noun type, sense or token
- range of values: 0-1, 100-700, etc.
- measurement scale: binary, nominal, ordinal, or interval-level
- way of assigning value: manually, completely automatically, or semi-automatically

Issues in data coding

Example: Definitions / coding

Levels of coding

- noun type, sense or token
- range of values: 0-1, 100-700, etc.
- measurement scale: binary, nominal, ordinal, or interval-level
- way of assigning value: manually, completely automatically, or semi-automatically

Issues in data coding

Example: Definitions / coding

Levels of coding

- noun type, sense or token
- range of values: 0-1, 100-700, etc.
- measurement scale: binary, nominal, ordinal, or interval-level
- way of assigning value: manually, completely automatically, or semi-automatically

Issues in data coding

Example: Definitions / coding

Levels of coding

- noun type, sense or token
- range of values: 0-1, 100-700, etc.
- measurement scale: binary, nominal, ordinal, or interval-level
- way of assigning value: manually, completely automatically, or semi-automatically

Issues in data coding

Example: Definitions / coding

Research questions

- In what way do the actual annotations of the concreteness of nouns change when we use various definitions or different implementations of some definition?
- In what way do the conclusions in a syntactic study change when using various approaches to annotate the concreteness of nouns?

Issues in data coding

Example: Definitions / coding

Six types of concreteness:

	definition	*noun*	*range*	*scale*	*assignment*
MRC	phys. perc.	type	100-700	interval	automatic
SYNTAX	phys. perc.	token	-1-1	interval	automatic
STXOBJ	phys. perc.	token	-1-1	interval	automatic
WNHIER	specificity	sense	0-16	ordinal	semi-automatic
WNPHYS	phys. perc.	sense	0-1	nominal	semi-automatic
PROTO	phys. perc.	token	0-1	nominal	manual

Issues in data coding

Example: Definitions / coding

Data

- 619 instances of the dative alternation from the ICE-GB corpus
- found with automatic parser, manually checked

Missing data

- WNHIER and WNPHYS: 527 items
- MRC: 428 items

Issues in data coding

Example: Definitions / coding

Data

- 619 instances of the dative alternation from the ICE-GB corpus
- found with automatic parser, manually checked

Missing data

- WNHIER and WNPHYS: 527 items
- MRC: 428 items

Issues in data coding

Example: Definitions / coding

Spearman rank correlations

	MRC	Proto	Syntax	StxObj	WNhier
Proto	0.25				
Syntax	0.40	0.30			
StxObj	0.26	0.27	**0.51**		
WNhier	-0.20	0.20	-0.02	0.12	
WNphys	-0.07	**0.60**	0.38	0.35	0.22

Issues in data coding

Example: Definitions / coding

- models with 12 predictors plus 1 type of concreteness each
- coefficients of concreteness types in model:

Type	Conc	
MRC	0.37	·
Proto	1.81	***
Syntax	-1.61	
StxObj	3.66	
WNhier	-0.16	
WNphys	0.56	*

Issues in data coding

Example: Definitions / coding

- models with 12 predictors plus 1 type of concreteness each
- coefficients of concreteness types in model:

Type	Conc	
MRC	0.37	·
PROTO	1.81	***
SYNTAX	-1.61	
STXOBJ	3.66	
WNHIER	-0.16	
WNPHYS	0.56	*

Issues in modeling data

Overview

- collinearity
- modeling technique
- variable selection
- model interpretation
- evaluation measures

Issues in modeling data

Overview

- collinearity
- modeling technique
- variable selection
- model interpretation
- evaluation measures

Issues in modeling data

Overview

- collinearity
- modeling technique
- variable selection
- model interpretation
- evaluation measures

Issues in modeling data

Overview

- collinearity
- modeling technique
- variable selection
- model interpretation
- evaluation measures

Issues in modeling data

Overview

- collinearity
- modeling technique
- variable selection
- model interpretation
- evaluation measures

Issues in modeling data

Overview

- collinearity
- modeling technique
- variable selection
- model interpretation
- evaluation measures

Issues in modelling data

Example: Variable selection

Variable selection in logistic regression: dative alternation

Daphne Theijssen (2010). Variable selection in Logistic Regression: The British English dative alternation. Thomas Icard and Reinhard Muskens (eds.), *Interfaces: Explorations in Logic, Language and Computation*. Series: Lecture Notes in Computer Science (subseries: Lecture Notes in Artificial Intelligence), volume 6211, Springer.

Issues in modeling data

Example: Variable selection

Approaches in variable selection

- all (significant) features
- stepwise forward selection
- stepwise backward selection
- (trying all combinations)
- (random forests)

Types of models

- mixed model (with verb sense as random effect)
- simple model (without a random effect)

Issues in modeling data

Example: Variable selection

Approaches in variable selection

- all (significant) features
- stepwise forward selection
- stepwise backward selection
- (trying all combinations)
- (random forests)

Types of models

- mixed model (with verb sense as random effect)
- simple model (without a random effect)

Issues in modeling data

Example: Variable selection

Research question:

- Is it justified to report only one 'optimal' regression model, if models can be built in several different ways?

Data

- 930 instances of the dative alternation from the ICE-GB corpus
- manually checked and coded

Issues in modeling data

Example: Variable selection

Research question:

- Is it justified to report only one 'optimal' regression model, if models can be built in several different ways?

Data

- 930 instances of the dative alternation from the ICE-GB corpus
- manually checked and coded

Issues in modeling data

Example: Variable selection

Mixed models

Effect	1. significant		2. forward		3. backward	
length diff	-2.50	***	-2.44	***	-2.39	***
rec=anim	-1.01	*				
rec=given					-1.44	***
rec=giv, S			-0.94	*		
rec=giv, W			-1.74	***		
rec=local	-2.53	***	-1.82	***	-1.78	***
th=pron, W	-1.79	*				
(intercept)	2.05	***	2.32	***	2.38	***
th=def	1.78	***				
th=giv			2.34	***	2.33	***
th=pron	2.19	***				

Issues in modeling data

Simple models

Effect	1. significant		2. forward		3. backward	
length diff	-1.73	***			-2.00	***
length diff, S			-2.35	***		
length diff, W			-1.71	***		
rec=def			-1.01	**	-1.15	***
rec=giv, W			-0.66	*		
rec=local	-1.22	***	-0.94	**	-1.15	**
rec=pron	-1.35	***	-0.88	**	-1.25	***
verb=abs, W					-0.99	*
verb=tr, S					-1.04	*
verb=tr, W					-1.32	*
(intercept)			0.82	**	1.56	**
th=conc	1.33	***	1.48	***	1.63	***
th=def			1.58	***	1.16	***
th=giv	1.48	***			0.98	**

Issues in modeling data

Example: Variable selection

Mixed models

			train=test		*10-fold cv*
selection	pred	baseline	AUC	accuracy	aver. accuracy
1. significant	6	0.723	0.979	0.935	0.819
2. forward	4	0.723	0.979	0.932	0.827
3. backward	4	0.723	0.978	0.928	0.833

Simple models

			train=test		*10-fold cv*
selection	pred	baseline	AUC	accuracy	aver. accuracy
1. significant	6	0.723	0.938	0.878	0.872
2. forward	7	0.723	0.943	0.878	0.876
3. backward	8	0.723	0.946	0.882	0.876

Issues in modeling data

Example: Variable selection

Mixed models

			train=test		*10-fold cv*
selection	pred	baseline	AUC	accuracy	aver. accuracy
1. significant	6	0.723	0.979	0.935	0.819
2. forward	4	0.723	0.979	0.932	0.827
3. backward	4	0.723	0.978	0.928	0.833

Simple models

			train=test		*10-fold cv*
selection	pred	baseline	AUC	accuracy	aver. accuracy
1. significant	6	0.723	0.938	0.878	0.872
2. forward	7	0.723	0.943	0.878	0.876
3. backward	8	0.723	0.946	0.882	0.876

Conclusions

Issues in data collection

- beware of data sparseness when studying infrequent alternations
- make an estimate of the data set size on a corpus sample
- use random forests?

Conclusions

Issues in data collection

- beware of data sparseness when studying infrequent alternations
- make an estimate of the data set size on a corpus sample
- use random forests?

Conclusions

Issues in data collection

- beware of data sparseness when studying infrequent alternations
- make an estimate of the data set size on a corpus sample
- use random forests?

Conclusions

Issues in data collection

- beware of data sparseness when studying infrequent alternations
- make an estimate of the data set size on a corpus sample
- use random forests?

Conclusions

Issues in data collection

- beware of data sparseness when studying infrequent alternations
- make an estimate of the data set size on a corpus sample
- use random forests?

Conclusions

Issues in data coding

- always establish inter-annotator agreement
- be explicit about the definitions and coding strategies
- take into account that different definitions/strategies influence your models
- weigh importance of model improvement, replicability and IA agreement

Conclusions

Issues in data coding

- always establish inter-annotator agreement
- be explicit about the definitions and coding strategies
- take into account that different definitions/strategies influence your models
- weigh importance of model improvement, replicability and IA agreement

Conclusions

Issues in data coding

- always establish inter-annotator agreement
- be explicit about the definitions and coding strategies
- take into account that different definitions/strategies influence your models
- weigh importance of model improvement, replicability and IA agreement

Conclusions

Issues in data coding

- always establish inter-annotator agreement
- be explicit about the definitions and coding strategies
- take into account that different definitions/strategies influence your models
- weigh importance of model improvement, replicability and IA agreement

Conclusions

Issues in data coding

- always establish inter-annotator agreement
- be explicit about the definitions and coding strategies
- take into account that different definitions/strategies influence your models
- weigh importance of model improvement, replicability and IA agreement

Conclusions

Issues in modeling data

- never assume that there is one 'optimal' model
- always mention which variable selection method you used
- do not draw strong conclusions if models differ widely
- assess whether your models suffer from overfitting

Conclusions

Issues in modeling data

- never assume that there is one 'optimal' model
- always mention which variable selection method you used
- do not draw strong conclusions if models differ widely
- assess whether your models suffer from overfitting

Conclusions

Issues in modeling data

- never assume that there is one 'optimal' model
- always mention which variable selection method you used
- do not draw strong conclusions if models differ widely
- assess whether your models suffer from overfitting

Conclusions

Issues in modeling data

- never assume that there is one 'optimal' model
- always mention which variable selection method you used
- do not draw strong conclusions if models differ widely
- assess whether your models suffer from overfitting

Conclusions

- Research about making choices in syntax involves making choices in researching
- The research choices influence the conclusions about the syntactic choices

Conclusions

- Research about making choices in syntax involves making choices in researching
- The research choices influence the conclusions about the syntactic choices

Thank you!

http://lands.let.ru.nl/~daphne