Automatically extending linguistically enriched data

Daphne Theijssen

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

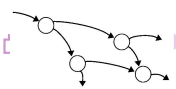
January 2009

About the project

Dative alternation

Dictionary on popular phrases:

- ... from that I coined the term lager lout to give it more meaning.
  - *OR:* ... 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.
  - *OR:* The species was identified in 1988, the name clearly owing alliteration much.

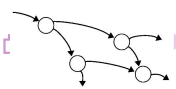


About the project

Dative alternation

Dictionary on popular phrases:

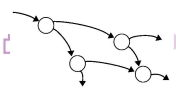
- ... from that I coined the term lager lout to give it more meaning.
  - *OR:* ... from that I coined the term lager lout to give more meaning to it.
- The species was identified in 1988, the name clearly owing much to alliteration.
  - *OR:* The species was identified in 1988, the name clearly owing alliteration much.



About the project

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"



About the project

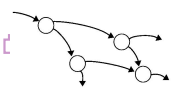
Aim

- Evaluating techniques that allow the integration of existing theories about syntactic variability
  - Logistic Regression
  - Maximum Entropy
  - Bayesian Networks
- British English dative alternation
- *There is no data like more data*

About the project

Aim

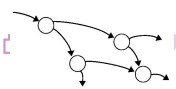
- Evaluating techniques that allow the integration of existing theories about syntactic variability
  - Logistic Regression
  - Maximum Entropy
  - Bayesian Networks
- British English dative alternation
- *There is no data like more data*



About the project

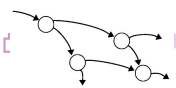
Aim

- Evaluating techniques that allow the integration of existing theories about syntactic variability
  - Logistic Regression
  - Maximum Entropy
  - Bayesian Networks
- British English dative alternation
- *There is no data like more data*



Overview

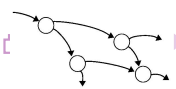
- Background
- Finding new cases
- Automatic enrichment
- Conclusion



Background

Previous research

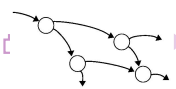
- Extracted 790 instances from ICE-GB corpus
- Manual annotation for the following features:
  - Animacy of recipient
  - Concreteness of theme
  - Definiteness of recipient and theme
  - Discourse givenness of recipient and theme
  - Length difference between the theme and the recipient
  - Number of recipient and theme
  - Person of recipient
  - Pronominality of recipient and theme
  - Semantic verb class
  - Structural parallellism
  - *Verb sense*



Background

Previous research

- Regression models (following Bresnan et al. 2007)
  - Prediction accuracy lower than 94% in Bresnan et al. (2007): 84.7% for Spoken, 88.1% for Written
  - Fewer significant effects than in Bresnan et al. (2007)
  - Written British English is best predicted by new model (not by a Spoken model)
- Data sets are too small for project aim (496 for spoken, 294 for written)
- There is no time to manually create and annotate such a set



Background

Previous research

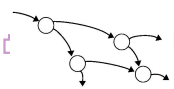
- Regression models (following Bresnan et al. 2007)
  - Prediction accuracy lower than 94% in Bresnan et al. (2007): 84.7% for Spoken, 88.1% for Written
  - Fewer significant effects than in Bresnan et al. (2007)
  - Written British English is best predicted by new model (not by a Spoken model)
- Data sets are too small for project aim (496 for spoken, 294 for written)
- There is no time to manually create and annotate such a set

Background

Research goal

Automatically extending the data set in two steps:

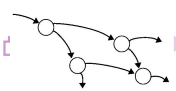
1. Finding new instances
2. Automatically enriching the data (annotation for features)



Finding new cases

Method

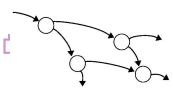
- Extract all sentences with ditransitives in **written** BNC
- Parse sentences with Connexor Machinese parser and keep sentences with dative construction
- Automatically filter dative sentences
- Apply same method to manual (ICE-GB) set for evaluation



Finding new cases

Method

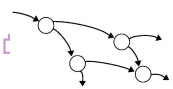
- Develop set of ditransitive verbs alternating with 'to'
  - Extracted 547 ditransitives from the following resources:
    - Johan Bos - Universita di Roma "La Sapienza"
    - College of LifeLong Learning Hong Kong: English Grammar Guide
    - ICE-GB corpus
    - VerbNet
    - TOSCA lexicon (1999)
    - Bresnan et al. (2007)
  - Kept 49 ditranstives:
    - that occur in at least 2 resources
    - verb lemma occurring >1000 times in written BNC
    - recognized as datives by Connexor parser
    - that occur in Bresnan's or my data or were approved by 4 linguists



Finding new cases

Method

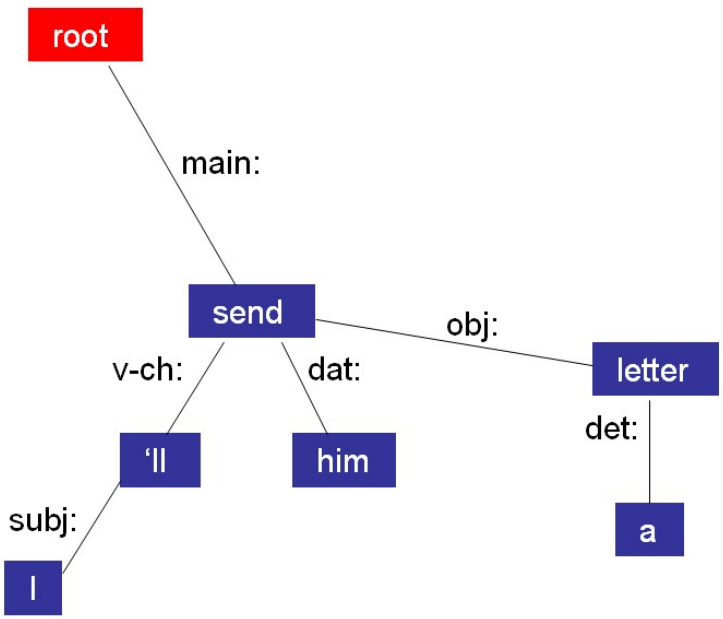
- Develop set of ditransitive verbs alternating with 'to'
  - Extracted 547 ditransitives from the following resources:
    - Johan Bos - Universita di Roma "La Sapienza"
    - College of LifeLong Learning Hong Kong: English Grammar Guide
    - ICE-GB corpus
    - VerbNet
    - TOSCA lexicon (1999)
    - Bresnan et al. (2007)
  - Kept 49 ditranstives:
    - that occur in at least 2 resources
    - verb lemma occurring >1000 times in written BNC
    - recognized as datives by Connexor parser
    - that occur in Bresnan's or my data or were approved by 4 linguists



Finding new cases

Method

Dative construction in Connexor parses:

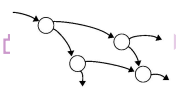


Keep instances with 'dat' and 'obj'

Finding new cases

Method

- Automatically exclude instances with:
  - a clausal object
  - passive voice
  - reversed order, e.g. ‘gave it him’
  - a phrasal verb
  - fronting
  - an adjectival theme, e.g. ‘made her drunk’
  - a fixed expression, e.g. ‘give birth to’
  - an empty object



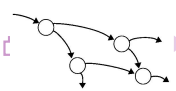
Finding new cases

Results

ICE-GB corpus: written

- Manual set: 294 cases
- Automatic set: 443 found (211 cases)

Correct:	157	(53%)
Missing/additional words in object(s):	54	(18%)
Different object(s):	2	(1%)
Missing:	81	(28%)
Additionally found:	230	



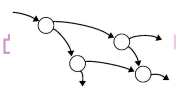
Finding new cases

Results

ICE-GB corpus: written

- Manual set: 294 cases
- Automatic set: 443 found (211 cases)

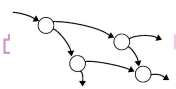
Correct:	157	(53%)
Missing/additional words in object(s):	54	(18%)
Different object(s):	2	(1%)
Missing:	81	(28%)
Additionally found:	230	



Finding new cases

Discussion

- Missing cases need to be inspected (complex structures?)
- BNC: written
  - Total number found: 31,417
  - Presumably $>$ 10,000 relevant cases: too many to check manually
- Possible ways of dealing with this problem:
  - determine number of cases desired (per verb sense?)
  - randomly select cases and manually check them until desired number is reached



Finding new cases

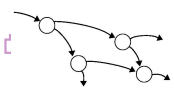
Discussion

- Missing cases need to be inspected (complex structures?)
- BNC: written
  - Total number found: 31,417
  - Presumably $>$ 10,000 relevant cases: too many to check manually
- Possible ways of dealing with this problem:
  - determine number of cases desired (per verb sense?)
  - randomly select cases and manually check them until desired number is reached

Finding new cases

Discussion

- Missing cases need to be inspected (complex structures?)
- BNC: written
  - Total number found: 31,417
  - Presumably $>$ 10,000 relevant cases: too many to check manually
- Possible ways of dealing with this problem:
  - determine number of cases desired (per verb sense?)
  - randomly select cases and manually check them until desired number is reached



Automatic enrichment

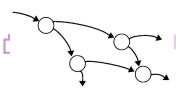
Method

- Animacy of recipient (inanimate)
  - head is person or noun in Wordnet => animate
  - head is in list of company names => animate
- Concreteness of theme (inconcrete)
  - lemma of head is animal, artifact, body, food, person or plant in WordNet => concrete
  - (also) in other category in WordNet (e.g. feeling) => abstract
- Definiteness of recipient and theme (indefinite)
  - head has (attribute with) POS tag 'definite article', 'demonstrative pronoun', 'interrogative/relative pronoun', 'possessive pronoun' => definite
  - head is *each other* or has POS tag 'reflexive pronoun', 'personal pronoun' or 'proper noun' => definite

Automatic enrichment

Method

- Animacy of recipient (inanimate)
  - head is person or noun in Wordnet => animate
  - head is in list of company names => animate
- Concreteness of theme (inconcrete)
  - lemma of head is animal, artifact, body, food, person or plant in WordNet => concrete
  - (also) in other category in WordNet (e.g. feeling) => abstract
- Definiteness of recipient and theme (indefinite)
  - head has (attribute with) POS tag 'definite article', 'demonstrative pronoun', 'interrogative/relative pronoun', 'possessive pronoun' => definite
  - head is *each other* or has POS tag 'reflexive pronoun', 'personal pronoun' or 'proper noun' => definite



Automatic enrichment

Method

- Animacy of recipient (inanimate)
  - head is person or noun in Wordnet => animate
  - head is in list of company names => animate
- Concreteness of theme (inconcrete)
  - lemma of head is animal, artifact, body, food, person or plant in WordNet => concrete
  - (also) in other category in WordNet (e.g. feeling) => abstract
- Definiteness of recipient and theme (indefinite)
  - head has (attribute with) POS tag 'definite article', 'demonstrative pronoun', 'interrogative/relative pronoun', 'possessive pronoun' => definite
  - head is *each other* or has POS tag 'reflexive pronoun', 'personal pronoun' or 'proper noun' => definite

Automatic enrichment

Method

- Discourse givenness of recipient and theme (nongiven)
  - indefinite => nongiven
  - head has POS tag personal pronoun => given
  - head is preceded by POS tag demonstrative pronoun => given
  - (synonym of) head present $\leq$ 20 main verbs before => given
- Length difference between the theme and the recipient
  - automatic: ln (nr words in theme) ln (nr words in recipient)
- Number of recipient and theme (singular)
  - plural (in Connexor parse) => plural
- Person of recipient (nonlocal)
  - *I*, *me*, *my*, *mine*, *myself*, *you*, *your*, *yours*, *yourself*, *yourselves*, *we*, *us*, *our*, *ours*, *ourselves* => local

Automatic enrichment

Method

- Discourse givenness of recipient and theme (nongiven)
  - indefinite => nongiven
  - head has POS tag personal pronoun => given
  - head is preceded by POS tag demonstrative pronoun => given
  - (synonym of) head present $\leq$ 20 main verbs before => given
- Length difference between the theme and the recipient
  - automatic: ln (nr words in theme) ln (nr words in recipient)
- Number of recipient and theme (singular)
  - plural (in Connexor parse) => plural
- Person of recipient (nonlocal)
  - *I*, *me*, *my*, *mine*, *myself*, *you*, *your*, *yours*, *yourself*, *yourselves*, *we*, *us*, *our*, *ours*, *ourselves* => local

Automatic enrichment

Method

- Discourse givenness of recipient and theme (nongiven)
  - indefinite => nongiven
  - head has POS tag personal pronoun => given
  - head is preceded by POS tag demonstrative pronoun => given
  - (synonym of) head present $\leq$ 20 main verbs before => given
- Length difference between the theme and the recipient
  - automatic: ln (nr words in theme) ln (nr words in recipient)
- Number of recipient and theme (singular)
  - plural (in Connexor parse) => plural
- Person of recipient (nonlocal)
  - *I*, *me*, *my*, *mine*, *myself*, *you*, *your*, *yours*, *yourself*, *yourselves*, *we*, *us*, *our*, *ours*, *ourselves* => local

Automatic enrichment

Method

- Discourse givenness of recipient and theme (nongiven)
  - indefinite => nongiven
  - head has POS tag personal pronoun => given
  - head is preceded by POS tag demonstrative pronoun => given
  - (synonym of) head present $\leq$ 20 main verbs before => given
- Length difference between the theme and the recipient
  - automatic: ln (nr words in theme) ln (nr words in recipient)
- Number of recipient and theme (singular)
  - plural (in Connexor parse) => plural
- Person of recipient (nonlocal)
  - *I*, *me*, *my*, *mine*, *myself*, *you*, *your*, *yours*, *yourself*, *yourselves*, *we*, *us*, *our*, *ours*, *ourselves* => local

Automatic enrichment

Method

- Pronominality of recipient and theme (nonpronominal)
  - head has POS tag ‘possessive pronoun’, ‘wh-determiner’, ‘existential there’, ‘indefinite pronoun’, ‘personal pronoun’, ‘wh-pronoun’ or ‘reflexive pronoun’ => pronominal
- Semantic verb class (verb specific default)
  - theme is abstract => abstract
  - lemma of theme head is communication in WordNet => communication
- Structural parallellism (no)
  - preceding construction (if present) is the same => yes
- Verb sense
  - verb + semantic verb class

Automatic enrichment

Method

- Pronominality of recipient and theme (nonpronominal)
  - head has POS tag 'possessive pronoun', 'wh-determiner', 'existential there', 'indefinite pronoun', 'personal pronoun', 'wh-pronoun' or 'reflexive pronoun' => pronominal
- Semantic verb class (verb specific default)
  - theme is abstract => abstract
  - lemma of theme head is communication in WordNet => communication
- Structural parallellism (no)
  - preceding construction (if present) is the same => yes
- Verb sense
  - verb + semantic verb class

Automatic enrichment

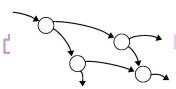
Method

- Pronominality of recipient and theme (nonpronominal)
  - head has POS tag 'possessive pronoun', 'wh-determiner', 'existential there', 'indefinite pronoun', 'personal pronoun', 'wh-pronoun' or 'reflexive pronoun' => pronominal
- Semantic verb class (verb specific default)
  - theme is abstract => abstract
  - lemma of theme head is communication in WordNet => communication
- Structural parallellism (no)
  - preceding construction (if present) is the same => yes
- Verb sense
  - verb + semantic verb class

Automatic enrichment

Method

- Pronominality of recipient and theme (nonpronominal)
  - head has POS tag 'possessive pronoun', 'wh-determiner', 'existential there', 'indefinite pronoun', 'personal pronoun', 'wh-pronoun' or 'reflexive pronoun' => pronominal
- Semantic verb class (verb specific default)
  - theme is abstract => abstract
  - lemma of theme head is communication in WordNet => communication
- Structural parallellism (no)
  - preceding construction (if present) is the same => yes
- Verb sense
  - verb + semantic verb class

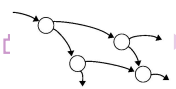


Automatic enrichment

Results

211 cases in both manual and automatic set

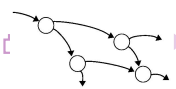
feature	*kappa*	*feature*	*kappa*
Person of recipient	1.00	Verb sense	0.46
Definiteness of theme	0.99	Concreteness of theme	0.34
Pronominality of recipient	0.98	Semantic verb class	0.28
Definiteness of recipient	0.94		
Number of recipient	0.90		
Number of theme	0.90		
Pronominality of theme	0.86		
Length Difference	0.74		
Givenness of theme	0.73		
Animacy of recipient	0.71		
Givenness of recipient	0.71		
Structure parallellism	0.62		



Automatic enrichment

Discussion

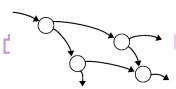
- Problematic features depend on each other:
  - semantic verb class uses concreteness of theme (kappa 0.34)
  - verb sense uses semantic verb class (kappa 0.28)
- Possible ways for improvement:
  - do error analysis of concreteness of theme
  - use verb sense disambiguation tool (suggestions?)
  - look at number of errors in subparts of feature values (e.g. humans, animals, companies in 'animate')



Automatic enrichment

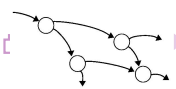
Discussion

- Problematic features depend on each other:
  - semantic verb class uses concreteness of theme (kappa 0.34)
  - verb sense uses semantic verb class (kappa 0.28)
- Possible ways for improvement:
  - do error analysis of concreteness of theme
  - use verb sense disambiguation tool (suggestions?)
  - look at number of errors in subparts of feature values (e.g. humans, animals, companies in 'animate')



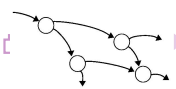
Conclusion

- Our aim is to evaluate techniques for modelling syntactic variability (currently: the British English dative alternation)
- There is a need for more data but no time for the manual creation and annotation of such a set.
- We developed an approach to automatically extend our linguistically enriched data set:
  - Finding new cases
    - ±75% (211) of manual set (294) found automatically
    - ±50% (211) of sentences found (443) is relevant
    - potentially >10,000 new instances from the BNC
  - Automatic enrichment
    - 12 of 15 features give acceptable kappas (>0.6) for the 211 cases



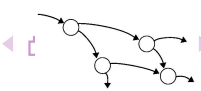
Conclusion

- Our aim is to evaluate techniques for modelling syntactic variability (currently: the British English dative alternation)
- There is a need for more data but no time for the manual creation and annotation of such a set.
- We developed an approach to automatically extend our linguistically enriched data set:
  - Finding new cases
    - ±75% (211) of manual set (294) found automatically
    - ±50% (211) of sentences found (443) is relevant
    - potentially >10,000 new instances from the BNC
  - Automatic enrichment
    - 12 of 15 features give acceptable kappas (>0.6) for the 211 cases



Conclusion

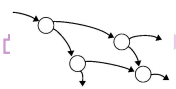
- Our aim is to evaluate techniques for modelling syntactic variability (currently: the British English dative alternation)
- There is a need for more data but no time for the manual creation and annotation of such a set.
- We developed an approach to automatically extend our linguistically enriched data set:
  - Finding new cases
    - $\pm$75% (211) of manual set (294) found automatically
    - $\pm$50% (211) of sentences found (443) is relevant
    - potentially >10,000 new instances from the BNC
  - Automatic enrichment
    - 12 of 15 features give acceptable kappas (>0.6) for the 211 cases



Conclusion

Near future:

- Study cases not found automatically
- Perform error analysis of problematic features and improve algorithm (where possible)
- Compare regression models found for manual and automatic set (211 cases)
- Apply method to written BNC and develop procedure for manual checking



Thank you!