# Tackling data insufficiency

## Automatically extending a (richly annotated) data set

Daphne Theijssen

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

*Corpus Linguistics Conference*
*Liverpool, U.K.*
*July 2009*

# Introduction

## Using corpora

# Introduction

## Dative alternation

The evil queen gives
Snowwhite
the poisonous apple.

The evil queen gives
the poisonous apple
to Snowwhite.

# Introduction

## Dative alternation

- Syntactic approach (e.g. Quirk et al. 1972)
- Semantic approach (e.g. Gries and Stefanowitsch 2004)
- Discourse approach (e.g. Collins 1995)

Aim of the project:

- Integrate approaches with different techniques:
  - Logistic Regression (e.g. Bresnan et al. 2007)
  - Maximum Entropy
  - Bayesian Networks
- Evaluate suitability for British English dative alternation

# Introduction

## Dative alternation

- Syntactic approach (e.g. Quirk et al. 1972)
- Semantic approach (e.g. Gries and Stefanowitsch 2004)
- Discourse approach (e.g. Collins 1995)

Aim of the project:

- Integrate approaches with different techniques:
  - Logistic Regression (e.g. Bresnan et al. 2007)
  - Maximum Entropy
  - Bayesian Networks
- Evaluate suitability for British English dative alternation

# Introduction

## Dative alternation

### Ditransitive constructions

Animacy
Definiteness
Discourse givenness
Number
Person
Pronominality

Concreteness
Definiteness
Discourse givenness
Number
Pronominality

Semantic verb class

# Introduction

## Dative alternation

Ditransitive constructions

Animacy
Definiteness
Discourse givenness
Number
Person
Pronominality

Concreteness
Definiteness
Discourse givenness
Number
Pronominality

Semantic verb class

# Introduction

## Dative alternation

Ditransitive constructions

Animacy
Definiteness
Discourse givenness
Number
Person
Pronominality

Concreteness
Definiteness
Discourse givenness
Number
Pronominality

Semantic verb class

# Introduction

## Data insufficiency

# Manual approach

- Employ corpus with syntactic annotations: ICE-GB
- Extracted 915 instances
- Manual annotation for the features needed

Problems:

- Data sets are too small for project aim (550 for spoken, 365 for written)
- No time to manually create and annotate another set
- Annotations are difficult to reproduce

# Manual approach

- Employ corpus with syntactic annotations: ICE-GB
- Extracted 915 instances
- Manual annotation for the features needed

Problems:

- Data sets are too small for project aim (550 for spoken, 365 for written)
- No time to manually create and annotate another set
- Annotations are difficult to reproduce

# Manual approach

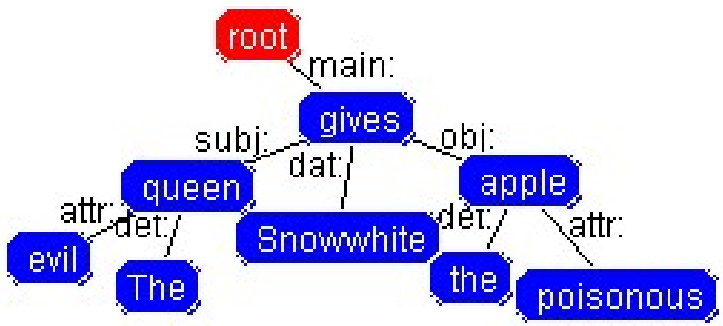
Proposed solution

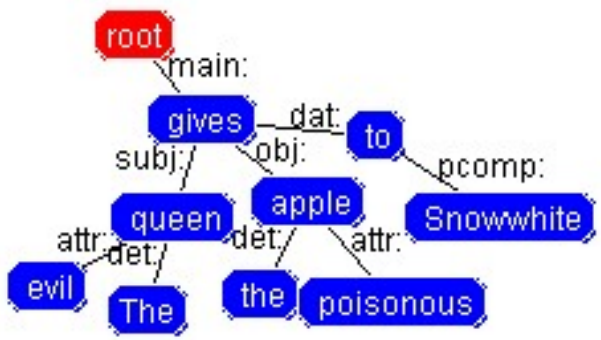
- Employ large corpus with minimal annotations: BNC
- Automatically extract instances
- Automatically enrich the data

# Automatic approach

## Finding cases

- Extract all sentences with ditransitives in **written** BNC
- Parse sentences with Connexor Machinese parser and keep sentences with dative construction




- Automatically filter the sentences

# Automatic approach

## Finding cases

- Result: $>$ 30000 sentences found in BNC
- How many are real cases?
- Apply same method to written part of ICE-GB
  - Precision: 58.4% (251/430)
  - Recall: 68.8% (251/365)
- Develop (semi-)automatic approach to check relevance
- Inspect instances not found

# Automatic approach

## Finding cases

- Result: $>$ 30000 sentences found in BNC
- How many are real cases?
- Apply same method to written part of ICE-GB
  - Precision: 58.4% (251/430)
  - Recall: 68.8% (251/365)
- Develop (semi-)automatic approach to check relevance
- Inspect instances not found

# Automatic approach

## Enriching the data

- Animacy of recipient (inanimate)
  - head is person or animal 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) => inconcrete
- 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 approach

## Enriching the data

- Animacy of recipient (inanimate)
  - head is person or animal 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) => inconcrete
- 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 approach

## Enriching the data

- Animacy of recipient (inanimate)
  - head is person or animal 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) => inconcrete
- 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 approach

## Enriching the data

- 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
- 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 approach

## Enriching the data

- 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
- 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 approach

## Enriching the data

- 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
- 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 approach

## Enriching the data

- 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 inconcrete => abstract
  - lemma of theme head is communication in WordNet => communication

# Automatic approach

## Enriching the data

- 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 inconcrete => abstract
  - lemma of theme head is communication in WordNet => communication

# Automatic approach

## Enriching the data

251 cases in both manual and automatic set

| *feature* | *accuracy* | *majority* |
|---|---|---|
| Person of recipient | 100% | 57% |
| Pronominality of recipient | 99% | 65% |
| Definiteness of recipient | 99% | 90% |
| Pronominality of theme | 98% | 91% |
| Definiteness of theme | 98% | 71% |
| Number of theme | 98% | 86% |
| Number of recipient | 96% | 71% |
| Givenness of theme | 94% | 89% |
| Givenness of recipient | 90% | 75% |
| Animacy of recipient | 89% | 82% |
| Concreteness of theme | 78% | 79% |
| Semantic verb class | 54% | 53% |

# Automatic approach

## Enriching the data

- Improve problematic features
  - Concreteness of theme
  - Semantic verb class
- New approach using contextual features
- For instance for *apple* (head of *the poisonous apple*):
  - lemma of focus word: *apple*
  - upper/lower case, symbols in focus word: *L*
  - POS tags of focus word: *N_NOM_SG*
  - relation with mother: *obj*
  - relation with mother + its lemma: *obj;give*
  - relation with mother + its POS tags: *obj;V_PRES_SG3*
  - relation with daughter: *det*, *attr*
  - relation with daughter + its lemma: *det;the*, *attr;poisonous*
  - relation with daughter + its POS tags: *det;DET*, *attr;A_ABS*

# Automatic approach

## Enriching the data

- Improve problematic features
  - Concreteness of theme
  - Semantic verb class
- New approach using contextual features
- For instance for *apple* (head of *the poisonous apple*):
  - lemma of focus word: *apple*
  - upper/lower case, symbols in focus word: *L*
  - POS tags of focus word: *N_NOM_SG*
  - relation with mother: *obj*
  - relation with mother + its lemma: *obj;give*
  - relation with mother + its POS tags: *obj;V_PRES_SG3*
  - relation with daughter: *det*, *attr*
  - relation with daughter + its lemma: *det;the*, *attr;poisonous*
  - relation with daughter + its POS tags: *det;DET*, *attr;A_ABS*

# Automatic approach

## Enriching the data

Marker-based algorithm:

- Find markers
  - Look at all cases having contextual feature X
  - Check whether at least 3/4 of have class Y
  - If so: feature X is a marker for class Y
- Classify cases
  - Concreteness of theme:
    - if marker for concrete: concrete
    - else: inconcrete
  - Semantic verb class
    - if marker for communication: communication
    - if marker for transfer of possession: transfer of possession
    - if marker for abstract: abstract
    - else: transfer of possession

# Automatic approach

## Enriching the data

Marker-based algorithm:

- Find markers
  - Look at all cases having contextual feature X
  - Check whether at least 3/4 of have class Y
  - If so: feature X is a marker for class Y
- Classify cases
  - Concreteness of theme:
    - if marker for concrete: concrete
    - else: inconcrete
  - Semantic verb class
    - if marker for communication: communication
    - if marker for transfer of possession: transfer of possession
    - if marker for abstract: abstract
    - else: transfer of possession

# Automatic approach

## Enriching the data

Results for 251 instances (leave-one-out)

| *feature* | *majority* | *accuracy old* | *accuracy new* |
|---|---|---|---|
| Concreteness of theme | 79% | 78% | 88% |
| Semantic verb class | 53% | 54% | 72% |

# Conclusion

- Need for richly annotated data but no time for manual creation
- Automatic approach:
  - Finding instances
    - Precision: 58.4% (251/430)
    - Recall: 68.8% (251/365)
    - >30,000 sentences extracted from the BNC
  - Enriching the data
    - accuracy of feature values 72% to 100%

# Conclusion

- Need for richly annotated data but no time for manual creation
- Automatic approach:
  - Finding instances
    - Precision: 58.4% (251/430)
    - Recall: 68.8% (251/365)
    - >30,000 sentences extracted from the BNC
  - Enriching the data
    - accuracy of feature values 72% to 100%

# Conclusion

- Need for richly annotated data but no time for manual creation
- Automatic approach:
  - Finding instances
    - Precision: 58.4% (251/430)
    - Recall: 68.8% (251/365)
    - >30,000 sentences extracted from the BNC
  - Enriching the data
    - accuracy of feature values 72% to 100%

# Conclusion

Near future:

- Study cases not found automatically
- Compare regression models found for manual and automatic set (251 cases)
- Apply method to written BNC and develop procedure for manual checking

Thank you!

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