Semantic similarity in non-native English: the case of *may* and *can* in French-English interlanguage

Sandra C. Deshors
University of Sussex
*Re-thinking synonymy*, Helsinki 28\textsuperscript{th} - 30\textsuperscript{th} October 2010
Introduction and overview

- Acquiring a new language:
  - cognitively demanding:
    - requires to identify large amounts of co-occurrence data
    - are probabilistic rather than absolute
    - hard to discern and learn
  - semantics is a particularly challenging domain:
    - in native language and more so in foreign language learning
    - semantic differences are not explicitly noticeable; their co-occurrence probabilistic patterns are particularly hard to discern
Introduction and overview

- co-occurrence patterns of *may* and *can* in native/learner Engl.:
  - how to investigate them quantitatively
  - in a way that is cognitively-grounded?
- Behavioral Profile approach (Gries & Divjak 2009)
  - highly compatible with psycholinguistic perspectives
  - involves a fine-grained annotation of corpus data and their statistical analysis
  - successfully applied in studies on synonymy, antonymy, polysemy (Divjak & Gries 2008, Gries & Otani 2010)
  - both within native languages (L1) or across L1s (Divjak & Gries 2006, Berez & Gries 2010)
  - has received experimental support (Divjak & Gries 2008)
  - so far the applicability of BP to L1 and L2 has not been tested
Setting the stage: what is problematic about the modals?

- *May* and *can* have fueled much theoretical debate about their semantic relations:
  - distinction between the senses of individual forms (cf. Leech 1969, Coates 1983)
  - distinction between the two lexical forms (cf. Leech 1969, Collins 2009)
  - interference of grammatical context (i.e. grammatical components: voice, aspect, clause type) with the interpretation of the forms
Setting the stage: previous corpus-based work on the modals

- **Native English:**
  - **Hermerén (1978):**
    - morphosyntactic categories (voice, grammatical person, type of main verb, etc.) highly influence the semantics of the modals
    - Hermerén's argument requires a powerful and versatile method to be empirically/quantitatively validated
  - **Gabrielatos & Sarmento (2006):**
    - recognise that the modals' distribution varies as a function of their syntactic contexts
    - no cognitively-motivated theoretical framework to (i) interpret the data meaningfully, and (ii) further explore the findings
  - **Collins (2009):**
    - form-based investigation: does not fully exploit linguistic contexts
    - statistical approach: frequency tables, no statistical analysis
Setting the stage: previous corpus-based work on the modals

- Learner English:
  - Aijmer (2002)
    - quantitative form-based approach to compare frequency of key modal words in native Engl. and Swedish-English IL
    - identifies the need to approach the modals functionally but her statistical approach (form-based freq. counts) is limited
  - Neff et al. (2003)
    - same approach as Aijmer (2002); compares of raw frequencies of modals' occurrences in 5 IL varieties
    - no account of the forms' contextual features; results are not illuminating
Setting the stage: previous corpus-based work on the modals

- Contrastive approaches:
  - Salkie (2004)
    - investigates the semantic relations between native *may/can* and French *pouvoir*; occurrences of *may/can* and their French translations (100 randomly extracted occurrences)
    - although Salkie offers a more analytical approach, he uses a rather small sample
Setting the stage: characteristics of the present study

- Methodological considerations

  - ideal methodological approach:
    - can integrate many different levels of linguistic analysis
    - involves large corpus data samples
    - aims at more than description
    - explores similarities and differences of L1 uses of *may* and *can*
    - explores L2 uses of *may* and *can* (Fr-Engl.IL)
    - explores how the same concept is used by learners in their L1 (*pouvoir*)
Setting the stage: characteristics of the present study

- Methodological considerations
  - Behavioral Profile approach (BP) (Gries & Divjak 2009) [cf. Gries (to appear) for an overview of Behavioral Profiles in corpus-based lexical semantics]
    - relies on the parallelism between the distributional and functional planes
    - explores how meanings and functions of lexical and syntactic elements are correlated with their distribution(s) of formal elements within their contexts (Gries & Divjak 2009)
    - allows for the statistical treatment of semantic and morpho-syntactic components (both main effects and interactions)
Setting the stage: characteristics of the present study

- Methodological considerations
  - BP approach: 4 methodological steps
    1. retrieve all instances of *may*, *can* and *pouvoir* from a corpus in the form of a concordance,
    2. analyse each match (semi) manually according to semantic and morpho-syntactic properties (ID tag) (annotation table)
    3. convert the resulting data points into a co-occurrence table that summarises the behaviour of each single modal form in relation to individual ID tag levels (expressed as co-occurrence %)
    4. evaluate the table statistically
Behavioural Profile vectors for *can*IL, *cannative*, *may*IL, *maynative* and *pouvoir* for all syntactic predictors

<table>
<thead>
<tr>
<th>ID tag</th>
<th>ID tag level</th>
<th>canIL</th>
<th>cannative</th>
<th>mayil</th>
<th>maynative</th>
<th>pouvoir</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>affirmative</td>
<td>0.7620</td>
<td>0.8109</td>
<td>0.9399</td>
<td>0.8798</td>
<td>0.7547</td>
</tr>
<tr>
<td>Negation</td>
<td>negative</td>
<td>0.2380</td>
<td>0.1891</td>
<td>0.0601</td>
<td>0.1202</td>
<td>0.2415</td>
</tr>
<tr>
<td>Negation</td>
<td>NA</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0038</td>
</tr>
<tr>
<td>SentenceType</td>
<td>declarative</td>
<td>0.9558</td>
<td>0.9690</td>
<td>0.9945</td>
<td>1.0000</td>
<td>0.9925</td>
</tr>
<tr>
<td>SentenceType</td>
<td>interrogative</td>
<td>0.0442</td>
<td>0.0310</td>
<td>0.0055</td>
<td>0.0000</td>
<td>0.0038</td>
</tr>
<tr>
<td>SentenceType</td>
<td>NA</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0038</td>
</tr>
<tr>
<td>ClType</td>
<td>coordinate</td>
<td>0.0992</td>
<td>0.1225</td>
<td>0.1366</td>
<td>0.1395</td>
<td>0.1132</td>
</tr>
<tr>
<td>ClType</td>
<td>main</td>
<td>0.5760</td>
<td>0.4516</td>
<td>0.5984</td>
<td>0.4764</td>
<td>0.5925</td>
</tr>
<tr>
<td>ClType</td>
<td>subordinate</td>
<td>0.3248</td>
<td>0.4259</td>
<td>0.2650</td>
<td>0.3820</td>
<td>0.2943</td>
</tr>
<tr>
<td>ClType</td>
<td>NA</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0021</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Data and methods: retrieval and annotations

• Data:
  - extracted from 3 untagged corpora:
    - International Corpus of Learner English
    - Louvain Corpus of Native English Essays
    - Corpus de Dissertations Françaises
  - (total) 652,386 words
  - written data, academic essays (500 words), 3rd/4th year university students

• Retrieval:
  - 3710 occurrences: may/can
    native/learner Eng. and French pouvoir
    - extracted and imported into a spreadsheet software (with R)
    - annotated for 22 morpho-syntactic and semantic variables
## Data and methods: retrieval and annotations

<table>
<thead>
<tr>
<th>Type</th>
<th>Variable</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>Corpus</td>
<td>native, interlanguage, French</td>
</tr>
<tr>
<td></td>
<td>GramAcc (acceptability)</td>
<td>yes, no</td>
</tr>
<tr>
<td>syntactic</td>
<td>Neg (negation)</td>
<td>affirmative, negated</td>
</tr>
<tr>
<td></td>
<td>SentType (sentence type)</td>
<td>declarative, interrogative</td>
</tr>
<tr>
<td></td>
<td>CItType (clause type)</td>
<td>main, coordinate, subordinate</td>
</tr>
<tr>
<td>morphological</td>
<td>Form</td>
<td><em>can, may, pouvoir</em> (and their negated forms)</td>
</tr>
<tr>
<td></td>
<td>SubjMorph: subject morphology</td>
<td>adj, adv, common noun, proper noun, relative pronoun, date, noun phrase, etc.</td>
</tr>
<tr>
<td></td>
<td>SubjPerson: subject person</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td></td>
<td>SubjNumber: subject number</td>
<td>singular, plural</td>
</tr>
<tr>
<td></td>
<td>Voice</td>
<td>active, passive</td>
</tr>
<tr>
<td></td>
<td>Aspect</td>
<td>perfect, perfective, progressive</td>
</tr>
<tr>
<td></td>
<td>Mood</td>
<td>indicative, subjunctive</td>
</tr>
<tr>
<td></td>
<td>SubjRefNumber: subject referent number</td>
<td>singular, plural</td>
</tr>
<tr>
<td>semantic</td>
<td>Senses (MF)</td>
<td>epistemic, deontic, dynamic</td>
</tr>
<tr>
<td></td>
<td>SpeakPresence (MF)</td>
<td>weak, medium, strong</td>
</tr>
<tr>
<td></td>
<td>Use (SV)</td>
<td>accomplishment, achievement, process, state</td>
</tr>
<tr>
<td></td>
<td>VerbSemantics (SV)</td>
<td>abstract, general action, action incurring transformation, action incurring movement, perception, etc.</td>
</tr>
<tr>
<td></td>
<td>RefAnim: subject referent animacy (RFT)</td>
<td>animate, inanimate</td>
</tr>
<tr>
<td></td>
<td>AnimType: subject referent animacy type (RFT)</td>
<td>animate, floral, object, place/time, mental/emotional, etc.</td>
</tr>
</tbody>
</table>
Data and method: the multifactorial statistical analysis

1. Hierarchical Cluster Analysis (HCA)
   - to assess similarities of the 5 forms in all sub-corpora
   - on the basis of
     - all variables
     - only syntactic
     - only morphological
     - only semantic

2. logistic regression
   - to measure the contribution of indep. variables to *may/can*
     - their main effect on Form (i.e. *may* and *can*)
     - their interaction with Corpus (as an additional independent variable)
Results: cluster analysis

- cluster analysis
  - similarities btw 5 forms on the basis of all ind. variables
  - long vertical line: cluster amalgamated early:
    - high intracluster similarity
    - low intercluster similarity
  - 2 cans yield highest degree of similarity
  - pouvoir and cans: highest dissimilarity

Fig. 1    Dendrogram for all independent variables (can/mayil = interlanguage can/may)
Results: cluster analysis

- cluster analysis
  - results from Fig. 1 are not replicated when morphosyntactic and semantic variables are treated separately
  - sem: IL can and native can are very similar; pouvoir is more similar to can than may is
  - morph/syn: clear Engl/Fr divide; IL may is too different from native may to be grouped together.
Results: cluster analysis

- BP snakeplot
  - pairwise differences between % of IL *may* and native *may*
  - learners deviate from natives:
    - underused *may* in subordinates and negated clauses
    - learners disprefer *may* in more complex grammatical environments

Fig. 3  Snakeplot for the most extreme differences between syntactic ID tags of *may*
**Results: logistic regression**

- 16 sign. variables; 6 interactions
- 99% classification accuracy
- correlation between the observed forms – *may* vs. *can* – and predicted probabilities is very high: $R^2=0.955$

**Overview of the results of the final GLM model**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>24.9 (1) ***</td>
<td>AnimType</td>
<td>98.2 (11) ***</td>
</tr>
<tr>
<td>GramAcc</td>
<td>13.8 (1) ***</td>
<td>Voice</td>
<td>55 (1) ***</td>
</tr>
<tr>
<td>Use</td>
<td>67.9 (1) ***</td>
<td>SentType</td>
<td>47.2 (1) ***</td>
</tr>
<tr>
<td>Elliptic</td>
<td>100 (2) ***</td>
<td>Negation</td>
<td>87.2 (1) ***</td>
</tr>
<tr>
<td>CIType</td>
<td>10.9 (1) ***</td>
<td>SpeakPresence</td>
<td>29905.9 (2) ***</td>
</tr>
<tr>
<td>VerbType</td>
<td>97.4 (2) ***</td>
<td>Corpus : CIType</td>
<td>60 (2) ***</td>
</tr>
<tr>
<td>VerbSemantics</td>
<td>384.9 (6) ***</td>
<td>Corpus : VerbSem</td>
<td>32.2 (6) ***</td>
</tr>
<tr>
<td>SubjPerson</td>
<td>26.6 (2) ***</td>
<td>Corpus : SubjNumb</td>
<td>37.4 (1) ***</td>
</tr>
<tr>
<td>SubjNumber</td>
<td>1.3 (1) ns</td>
<td>Corpus : RefAnim</td>
<td>122.2 (1) ***</td>
</tr>
<tr>
<td>SubjMorph</td>
<td>49.1 (4) ***</td>
<td>Corpus : AnimType</td>
<td>118.2 (11) ***</td>
</tr>
<tr>
<td>RefAnim</td>
<td>59.2 (1) ***</td>
<td>Corpus : Negation</td>
<td>12 (1) ***</td>
</tr>
</tbody>
</table>
Results: logistic regression

- logistic regression: Corpus:ClType
  - In IL *can* is more strongly preferred over *may* in subordinate clauses than it is in native English

---

**Fig. 4** Bar plots of relative frequencies of Corpus:ClType
Results: logistic regression

- logistic regression: Corpus:Neg
  - both native spk and learners prefer *can* in negated clauses
  - BUT learners do so more strongly
  - negated clauses are more complex and preferred with the more frequent modal

Fig. 4 Bar plots of relative frequencies of Corpus:Neg
Results: logistic regression

- logistic regression: Corpus:SubjNumb
  - native spk use *can* more often with singular subjects
  - learners use *can* more often with plural subjects

Fig. 5 Bar plots of relative frequencies of Corpus:SubjNumb
Results: logistic regression

- logistic regression: Corpus:RefAnim
  - native speakers' choices of *may* and *can* do not vary much
  - learners prefer *can* more strongly with animate subjects

Fig. 6 Bar plots of relative frequencies of Corpus:RefAnim
Results: logistic regression

- logistic regression: Corpus:VerbSem
  - learners and native speakers differ most strongly with semantic more abstract verbs (achieve, cause, deprive or lead to) and temporal verbs (end up, spend or begin)
  - learners prefer can with abstract verbs more strongly than natives
  - learners prefer may with temporal verbs more strongly than natives
  - learners prefer may with communication verbs and can with action-transf. verbs

Fig. 6 Bar plots of relative frequencies of Corpus:VerbSem
Interim summary

- BP approach and the regression:
  - clusters:
    - *can/may* (in both lang. varieties) and *pouvoir* relate to each other differently whether they are investigated semantically, syntactically or morphologically
    - semantically, English *can* is more similar to French *pouvoir* than English *may*
  - regression: learners choose more frequent *can* over *may* in more complex grammatical environments (*negation, subordinate clauses, abstract lexical verbs, plural subjects*)
Discussion

• regression results are in line with Rodenburg's (1996) complexity principle
  – speakers tend to prefer "more explicit grammatical alternatives (...) in cognitively more complex environments" (p. 149)
  – Rodenburg's study focuses on:
    • native English
    • syntactic environments (discontinuous constructions, heavy subject expressions or passive constructions, subordinate clauses)

• implications of the study for the complexity principle:
  – applies to L2 as well as L1
  – applies to semantics (VerbSem, RefAnim, AnimTyp) and morphology (SubjNumb, Neg) environments
  – grammatical contexts present processing constraints that influence learners' lexical choices
Concluding remarks

- learners have built up a mental category for *may* and *can* that is internally rather coherent
- ! interactions show that 6 cues are weighted incorrectly by learners

- BP approach has proved successful:
  - has reached beyond quantitative data description into Cognitive Linguistics and psycholinguistics
  - the overall results can testify to the strength of the categories under study
  - the regression (with its interactions) pinpoints where the categories of the learner are still substantially different from the native speaker

- overall, the regression results show that learners' "non-nativeness" manifests itself at all linguistic levels simultaneously
- this is the first study proposing this kind of approach to learner-language, more rigorous testing involving more IL varieties is necessary
References


Gries, Stefan Th. 2010. Behavioural Profiles 1.01. A program for R 2.7.1 and higher.


Thank you!

this presentation is based on Deshors & Gries (to appear)
http://sites.google.com/site/sandracdeshors/home
http://tinyurl.com/stgries
Excerpt of an annotation table including selected variables

<table>
<thead>
<tr>
<th>Case</th>
<th>Match</th>
<th>Corpus</th>
<th>CIType</th>
<th>Use</th>
<th>VerbSemantics</th>
<th>Neg</th>
<th>RefAnim</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>may</td>
<td>native</td>
<td>coordinate</td>
<td>process</td>
<td>ment/cog/emotiona l</td>
<td>affirmative</td>
<td>animate</td>
</tr>
<tr>
<td>133</td>
<td>may</td>
<td>native</td>
<td>main</td>
<td>state</td>
<td>copula</td>
<td>affirmative</td>
<td>inanimate</td>
</tr>
<tr>
<td>186</td>
<td>may</td>
<td>native</td>
<td>main</td>
<td>process</td>
<td>ment/cog/emotiona l</td>
<td>negative</td>
<td>animate</td>
</tr>
<tr>
<td>188</td>
<td>can</td>
<td>il</td>
<td>coordinate</td>
<td>process</td>
<td>ment/cog/emotiona l</td>
<td>affirmative</td>
<td>animate</td>
</tr>
<tr>
<td>287</td>
<td>cannot</td>
<td>il</td>
<td>subordinate</td>
<td>state</td>
<td>abstract</td>
<td>negative</td>
<td>inanimate</td>
</tr>
<tr>
<td>354</td>
<td>peut</td>
<td>fr</td>
<td>main</td>
<td>process</td>
<td>ment/cog/emotiona l</td>
<td>negative</td>
<td>animate</td>
</tr>
<tr>
<td>364</td>
<td>peuven t</td>
<td>fr</td>
<td>subordinate</td>
<td>process</td>
<td>abstract</td>
<td>negative</td>
<td>inanimate</td>
</tr>
</tbody>
</table>