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 28th -30th 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-occurence %)
    4. evaluate the table statistically

Behavioural Profile vectors for *can*IL, *can*native, *may*IL, *may*native and *pouvoir* for all syntactic predictors

ID tag	ID tag level	*can*IL	*can*native	*may*il	*may*native	*pouvoir*
Negation	affirmative	0.7620	0.8109	0.9399	0.8798	0.7547
Negation	negative	0.2380	0.1891	0.0601	0.1202	0.2415
Negation	NA	0.0000	0.0000	0.0000	0.0000	0.0038
SentceType	declarative	0.9558	0.9690	0.9945	1.0000	0.9925
SentceType	interrogative	0.0442	0.0310	0.0055	0.0000	0.0038
SentceType	NA	0.0000	0.0000	0.0000	0.0000	0.0038
ClType	coordinate	0.0992	0.1225	0.1366	0.1395	0.1132
ClType	main	0.5760	0.4516	0.5984	0.4764	0.5925
ClType	subordinate	0.3248	0.4259	0.2650	0.3820	0.2943
ClType	NA	0.0000	0.0000	0.0000	0.0021	0.0000

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

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

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

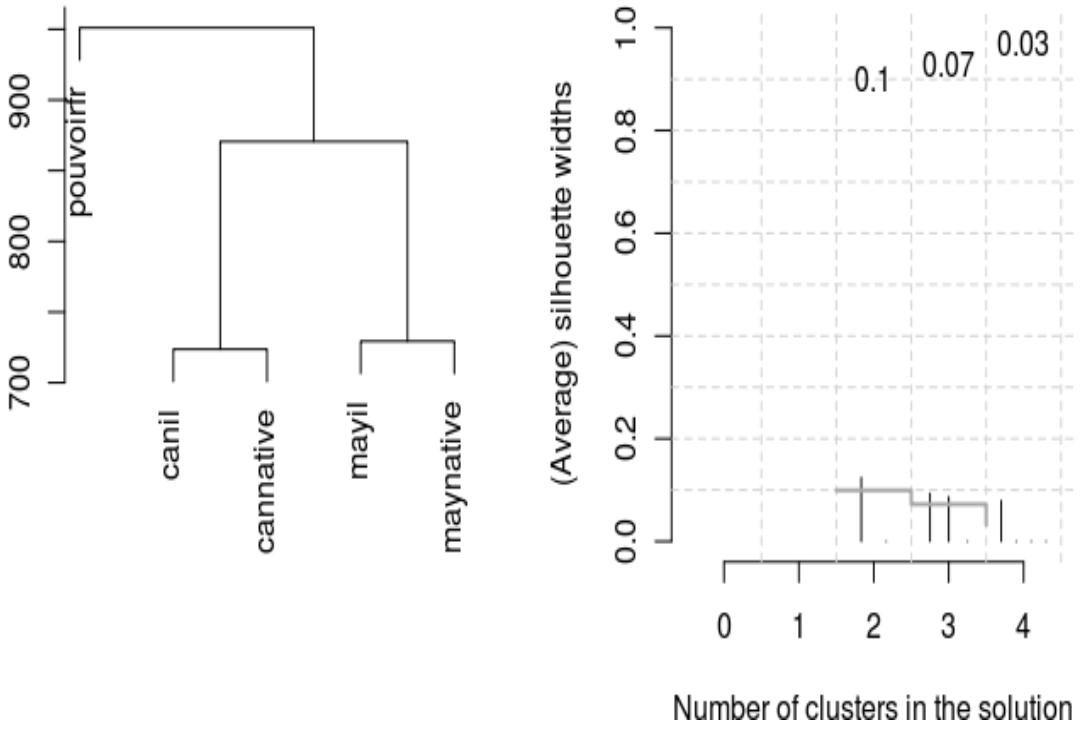


Fig. 1 Dendrogram for **all independent variables** (*can/may*il = interlanguage *can/may*)

- 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

Results: cluster analysis

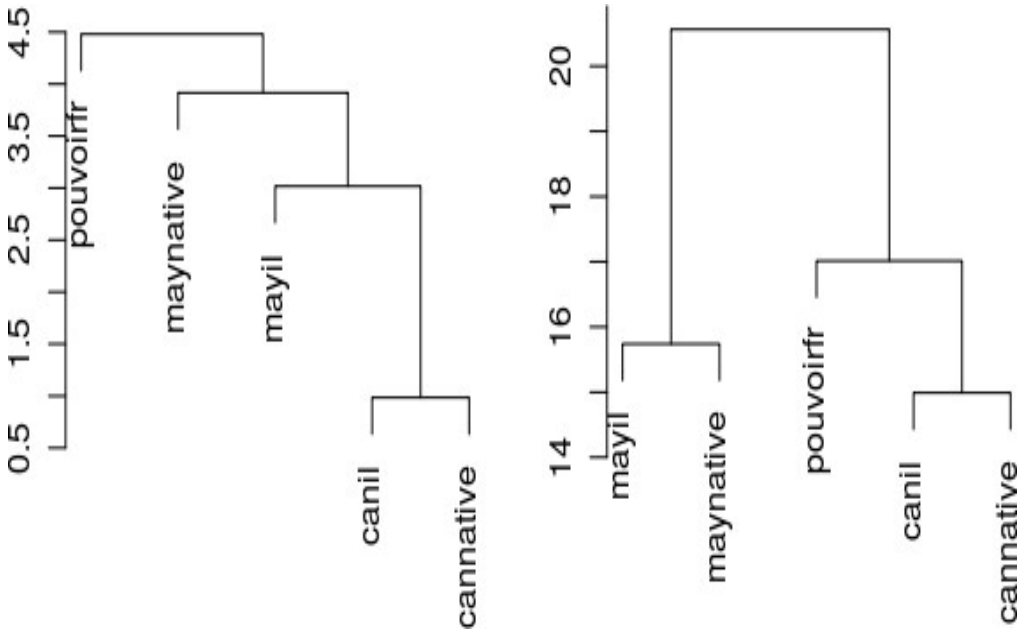


Fig. 2 Dendrogram for all **morphosyntactic** variables (**left**) and all **semantic** variables (**right**)

- 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

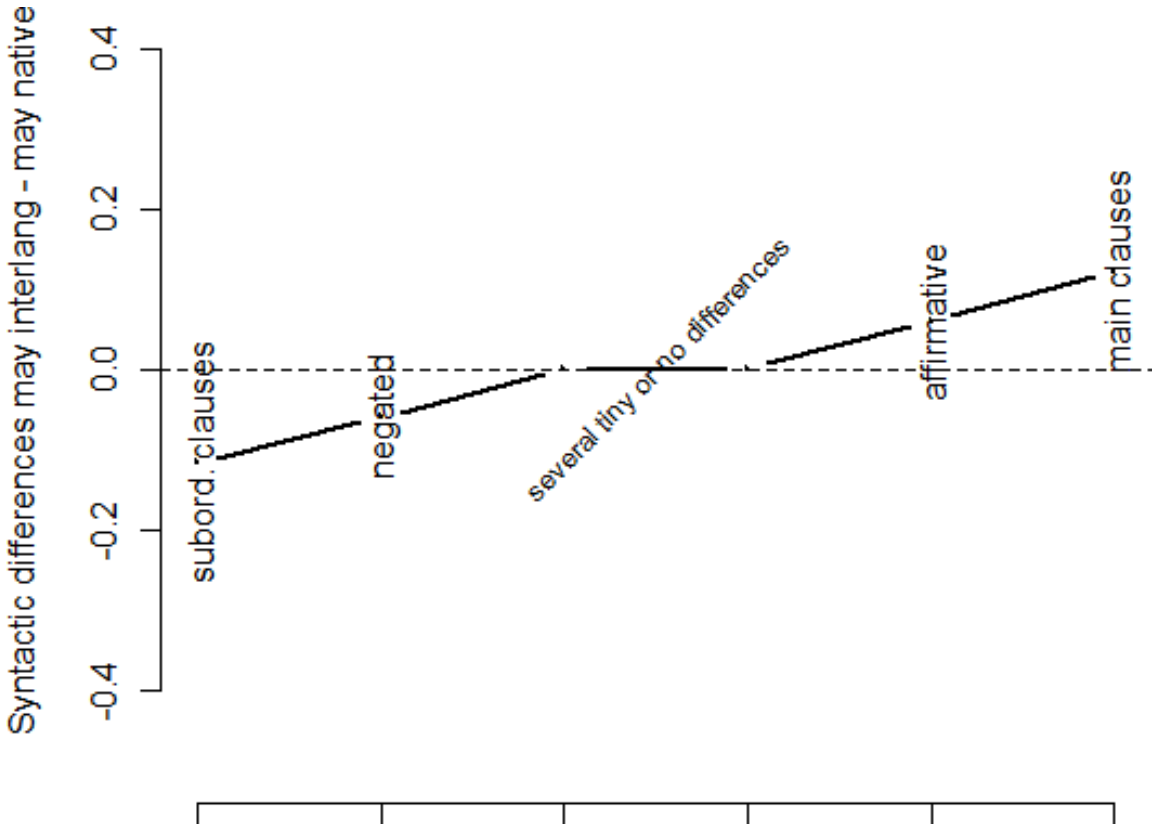


Fig. 3 Snakeplot for the most extreme differences between syntactic ID tags of *may*

- 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

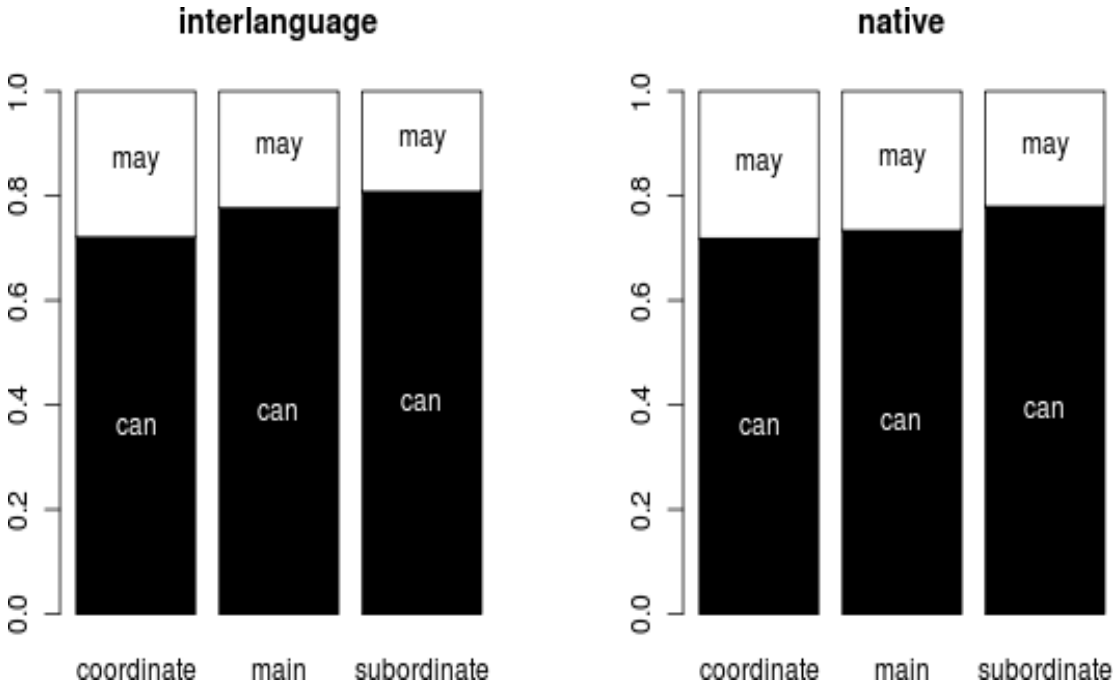
Results: logistic regression

- 16 sign. variables; 6 interactionsR
- 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

Predictor	Chi-square (*df*): sign.		Predictor	Chi-square (*df*): sign.
Corpus	24.9 (1) ***		**AnimType**	98.2 (11) ***
GramAcc	13.8 (1) ***		**Voice**	55 (1) ***
Use	67.9 (1) ***		**SentType**	47.2 (1) ***
Elliptic	100 (2) ***		**Negation**	87.2 (1) ***
ClType	10.9 (1) ***		**SpeakPresence**	29905.9 (2) ***
VerbType	97.4 (2) ***		**Corpus : ClType**	60 (2) ***
VerbSemantics	384.9 (6) ***		**Corpus : VerbSem**	32.2 (6) ***
SubjPerson	26.6 (2) ***		**Corpus : SubjNumb**	37.4 (1) ***
SubjNumber	1.3 (1) ns		**Corpus : RefAnim**	122.2 (1) ***
SubjMorph	49.1 (4) ***		**Corpus : AnimType**	118.2 (11) ***
RefAnim	59.2 (1) ***		**Corpus : Negation**	12 (1) ***

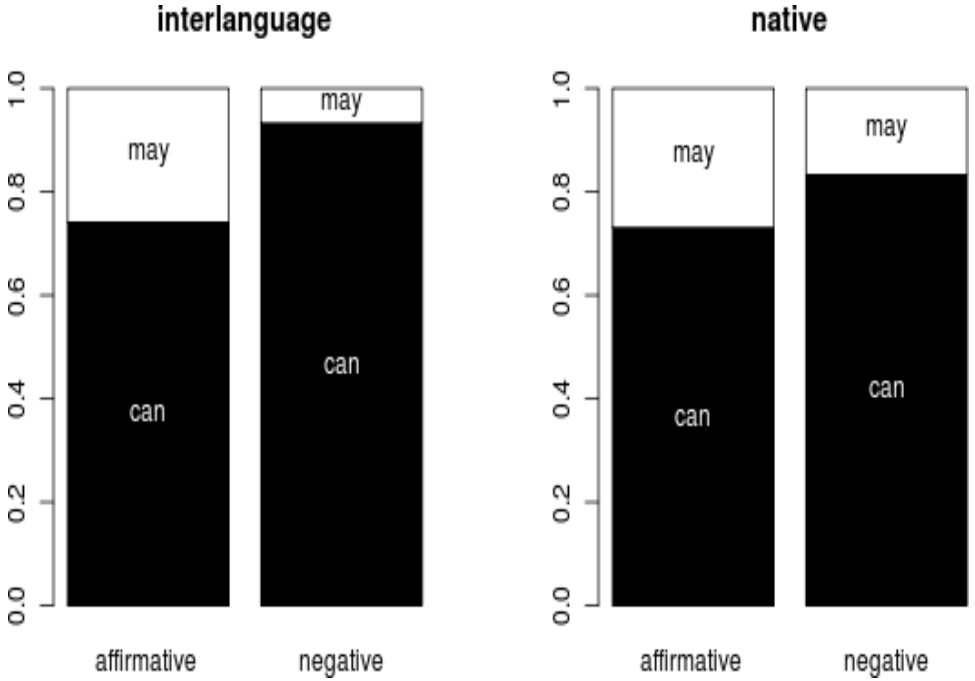
Results: logistic regression



- 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

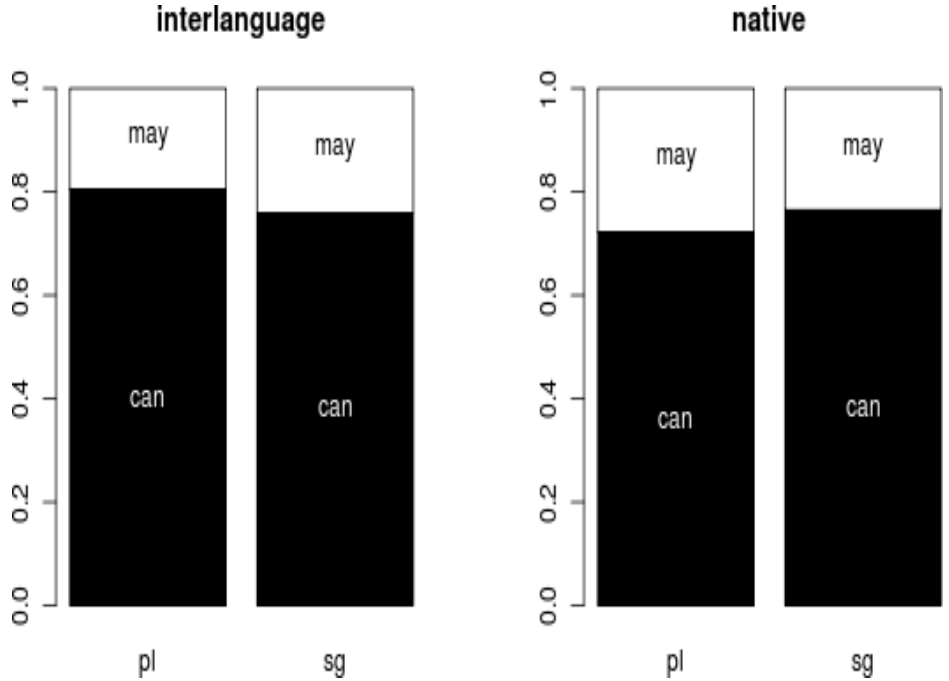
Results: logistic regression



- 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

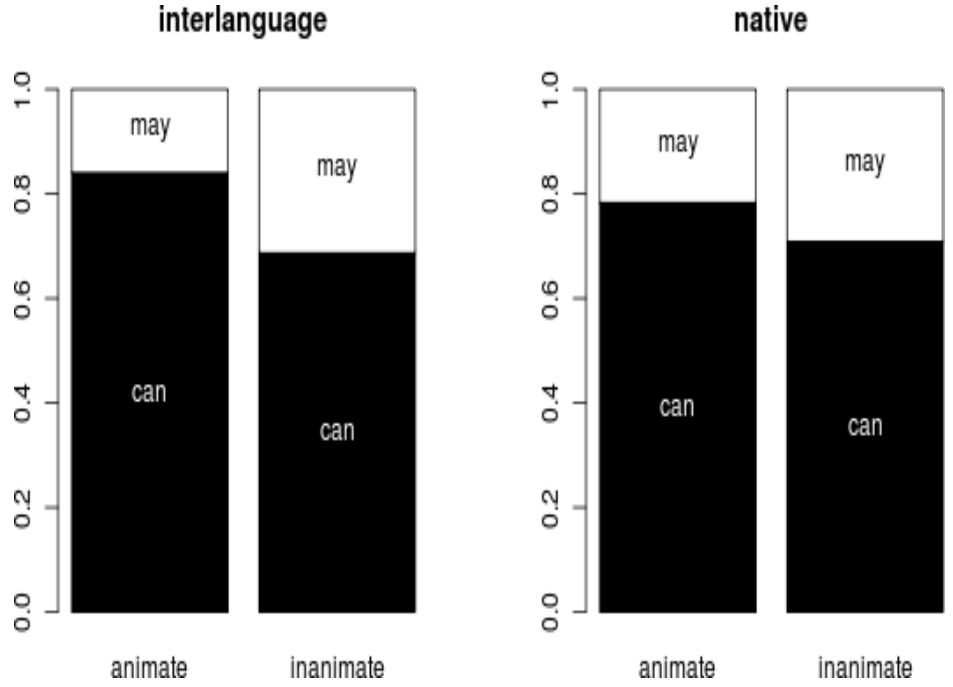
Results: logistic regression



- 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

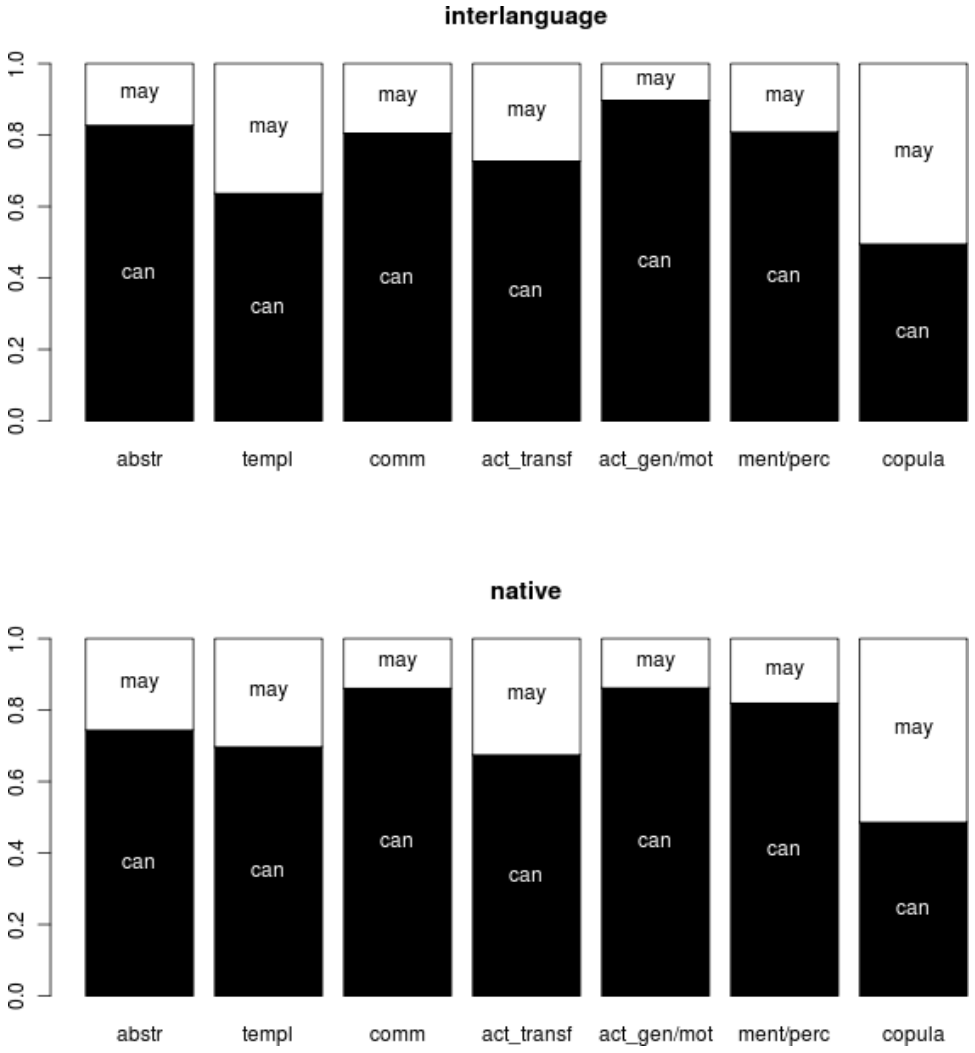


Fig. 6 Bar plots of relative frequencies of Corpus:VerbSem

- 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

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

Aijmer, Karin. 2002. Modality in advanced Swedish learners' written interlanguage. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, ed. by Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson, 55-76. John Enjambement

Bates, Elizabeth & Brian MacWhinney. 1982. Functionalist approaches to grammar. In *Language acquisition: the state of the art*, ed. by Eric Wanner & Lila R. Gleitman, 173-218. Cambridge University Press.

Bates, Elizabeth & Brian MacWhinney. 1989. Functionalism and the competition model. In *The cross-linguistic study of sentence processing*, ed. by Brian MacWhinney & Elizabeth Bates, 3-73. Cambridge University Press.

Coates, Jennifer. 1983. *The semantics of the modal auxiliaries.* London: Croom Helm.

Collins, Peter. 2009. *Modals and quasi modals in English*. Amsterdam: Rodopi.

Divjak, Dagmar S. & Stefan Th. Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23-60.

Divjak, Dagmar S. & Stefan Th. Gries. 2008. Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon* 3(2). 188-213.

Gabrielatos, Costas & Simone Sarmento. 2006. Central modals in an aviation corpus: frequency and distribution. *Letras de Hoje* 41(2). 215-240.

Gries, Stefan Th. To appear. Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon*.

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

Gries, Stefan Th. & Dagmar S. Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In *New directions in cognitive linguistics*, ed. by Vyvyan Evans & Stephanie S. Pourcel, 57-75. Amsterdam, Philadelphia: John Benjamins.

Gries, Stefan Th. & Naoki Otani. 2010. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34. 121-150.

Hermerén, Lars. 1978. *On modality in English: a study of the semantics of the modals*. Lund: LiberLäromedel/Gleerups.

Leech, Geoffrey. 1969. *Towards a semantic description of English*. Indiana University Press.

Leech, Geoffrey. 2004. *Meaning and the English Verb.*Longman.

MacWhinney, Brian. 2004. A unified model of language acquisition. URL ‹http://psyling.psy.cmu.edu/papers/CM-general/unified.pdf›, accessed 18 June 2010.

Neff JoAnne, Emma Dafouz, Honesto Herrera, Francisco Martínez, & Juan Pedro Rica. 2003. Contrasting the use of learner corpora: the use of modal and reporting verbs in the expression of writer stance. In *Extending the scope of corpus-based research. New applications. New challenges*, ed. by Sylviane Granger & Stéphanie Petch-Tyson, 211-230. Amsterdam: Rodopi.

Palmer, Frank. 1979. *Modality and the English modals*. London, New York: Longman.

R Development Core Ream. 2010. R: A Language and Environment for Statistical Computing. Foundation for statistical computing. Vienna, Austria. ‹ http://www.R-project.org ›.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2):149-182.

Salkie, Raphael. 2000. Corpus linguistics. In *A brief guide to research in French language and linguistics*, ed. by Anthony Lodge, 44-52. AFLS Cahiers 6.HS.

Salkie, Raphael. 2004. Towards a non-unitary analysis of modality. In *Contrates: mélanges offerts à Jacqueline Guillemin-Flescher*, ed. by Lucie Gournay & Jean-Marie Merle, 169-182. Paris: Ophrys.

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

Case	Match	Corpus	ClType	Use	VerbSemantics	Neg	RefAnim
5	may	native	coordinate	process	ment/cog/emotional	affirmative	animate
133	may	native	main	state	copula	affirmative	inanimate
1760	may	native	main	process	ment/cog/emotional	negative	animate
1886	can	il	coordinate	process	ment/cog/emotional	affirmative	animate
2876	cannot	il	subordinate	state	abstract	negative	inanimate
3540	peut	fr	main	process	ment/cog/emotional	negative	animate
3645	peuvent	fr	subordinate	process	abstract	negative	inanimate