

# Towards Predicting New Words from Newer Words: Lexical Borrowings in French

Paula Chesley  
ches0045@umn.edu

Linguistics Program  
University of Minnesota  
USA

American Association for Corpus Linguistics  
Brigham Young University  
13 March 2008

- ▶ Can we predict new words?
  - ▶ Of the possible neologisms, which are actually adopted into a language?
  - ▶ Test theories of lexical productivity
- ▶ Ways of creating new words (neologistic strategies)
  - ▶ Make acronyms, clippings, blends, or borrow
- ▶ Zoom in: lexical borrowings in French
  - ▶ Borrowing is a very productive neologistic strategy in French
  - ▶ Of interest to linguists and non-linguists alike
    - ▶ 1975 Bas-Lauriol Law and 1994 Toubon Law regulate the use of foreign words in public domains in France
    - ▶ Lots of conscious language planning

# Predicting new lexical borrowings in French

- ▶ Definition of lexical borrowing
  - ▶ The (approximate) form and meaning are copied from donor to recipient language
  - ▶ The borrowing does not yet exist in a French dictionary
- ▶ Get a corpus from an early date
  - ▶ T1 corpus – articles from *Le Monde* (Abeillé et al. 2003)
  - ▶ 21,560 parsed, POS-labeled sentences, gathered from 1989-1993
- ▶ Find the new borrowings in it
- ▶ Look up these borrowings in a corpus from a later date, preferably controlling for register
  - ▶ T2 corpus – online archives of *Le Figaro*, 1996-2006
- ▶ Use frequency in T2 corpus as a proxy for integration into the lexicon
- ▶ Find relevant predictor variables

# How to identify lexical borrowings in T1 corpus?

- ▶ Corpus features manual verification of automatic POS labeling, including foreign words
- ▶ Problem
  - ▶ A lot of foreign words are not indicated as such (e.g. *popiwek*)
  - ▶ And a lot of tagged “foreign” words are highly frequent in current French and exist in a French dictionary (e.g. *week-end*)
- ▶ Solution
  - ▶ Automatic search in corpus for non-native, low frequency letters or letter combinations, like “k”, “ö”, “qi”, etc.
  - ▶ Examine context around words with these letters since borrowed words might occur in clusters
  - ▶ Add any new letter combinations found in neighboring words to list of letter combinations queried (93 total)
- ▶ Exclude proper names, products, etc.

# Data: how many borrowings in the T1 corpus?

- ▶ 280 borrowed tokens, 138 borrowed types
  - ▶ Lemma *deutschemark* was extremely frequent (49 occurrences)
- ▶ Sensitivity of method + initial corpus labellings
  - ▶ Random sample of 431 sentences from T1 corpus (4% of total)
  - ▶ 9/14 borrowings found
  - ▶ Substantially more robust at detecting borrowings than the corpus labellings alone (3/14)
- ▶ Estimate how many borrowings are in the corpus
  - ▶ 95% score confidence interval for number of borrowings in corpus: between 334 and 722
  - ▶ 95% score confidence interval for sensitivity of method: between 38.76% and 83.66%

# Using frequency as a proxy for lexical integration

- ▶ Borrowings in T2 corpus have a bimodal distribution
- ▶ Baayen and Lieber (1997): bimodal distribution of Dutch words with a particular prefix shows a difference between frequent, well-entrenched lexical items and infrequent nonce formations using the prefix
- ▶ Lexical borrowings are composed of two types of borrowings: infrequent nonce borrowings and frequent, productive borrowings

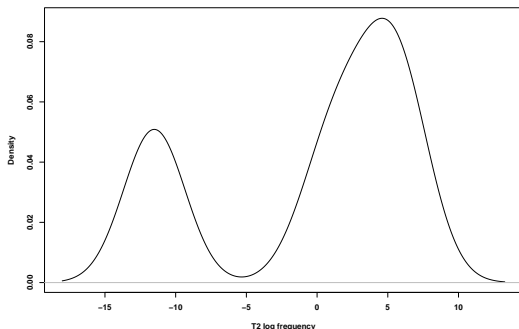


Figure: Estimated probability density function for borrowings at T2.

# Distinguishing nonce borrowings from productive borrowings

- ▶ Frequency at T2 shows whether the borrowings are productive or nonce

Nonce borrowings			Productive borrowings		
Borrowing	T1 Freq	T2 Freq	Borrowing	T1 Freq	T2 Freq
popiwek	1	0	lobbying	3	865
taref	1	0	come-back	1	333
the	1	0	hedge funds	3	368
huasipongo	2	0	perestroika	8	210
ejido	9	0	success story	2	382
classless society	1	0	running	1	53

**Table:** Examples of nonce and productive borrowings and their respective frequencies in the T1 and T2 corpora.

- ▶ **Sense pattern** of borrowings (**monosemous** vs. **polysemous**)
- ▶ **Cultural context** of borrowings (relationship between language and culture of denotatum: **restricted** vs. **unrestricted**)
  - ▶ restricted: e.g., *perestroika* when talking about Russia
  - ▶ unrestricted: e.g., *perestroika* when talking about China
- ▶ Hypothesis: **polysemous** senses and **unrestricted** cultural contexts can refer to more things, so they are likely to be more frequent and thus more likely to be integrated

- ▶ **Source language** of the borrowing
  - ▶ Binary: English vs. non-English
  - ▶ Hypothesis: English will correlate with higher frequencies in T2 corpus (English borrowings will be more integrated)
- ▶ **Number of syllables** of the borrowing
  - ▶ Fewer syllables hypothesized to correlate with higher T2 frequencies

## Results: *sense pattern*

- ▶ Significant correlation between borrowings with polysemous senses and higher T2 frequencies
- ▶ Polysemous ex.: *flash*: in corpus, *flash memory*; elsewhere in French, *flashy* (33 occurrences in T2 corpus)
- ▶ Monosemous ex.: *minifundio*: in corpus, small Latin American farm; same elsewhere in French (0 occurrences in T2 corpus)



Figure: Minifundios.

# Results: *language* and *number of syllables*

- ▶ Language
  - ▶ Borrowings from English significantly more likely to be frequent in T2 corpus
- ▶ Number of syllables
  - ▶ Significant correlation between borrowings with fewer syllables and higher T2 frequencies
  - ▶ Fewer syllables: *french doctors*: 25 occurrences in T2 corpus
  - ▶ More syllables: *french travel way of life*: 0 occurrences in T2 corpus
  - ▶ Caveat: average length of extant words not controlled for
  - ▶ Is this the relevant property?

- ▶ No significant correlation between unrestricted cultural contexts and higher T2 frequencies
- ▶ Flaitz 1988: The words *success story*, for example, appear...in an article about the life and career of Lee Iococca...Had the article been focussed on the life and career of François Mitterrand, the anglophone phrase *success story* would have incited much well-deserved criticism (87).
- ▶ Why not significant?
  - ▶ The newspaper talks about foreign events more?
  - ▶ Study doesn't pick up relevant cultural context features?
  - ▶ Interaction of cultural context and another predictor variable?

# Conclusions and future work

- ▶ Corpus-based, diachronic study on lexical borrowings in French
- ▶ Aims to predict new lexical borrowings
- ▶ Presents a semi-automatic way of detecting foreign words in a corpus
- ▶ Uses frequency of borrowings at a later date as a proxy for the degree of integration of a newly borrowed lexical item
- ▶ Examines possible predictor variables: source language, sense pattern, and number of syllables of borrowing look promising
- ▶ Future work
  - ▶ Number of syllables – the relevant factor?
  - ▶ Is cultural context relevant?
  - ▶ Prediction step

# Further reading



A. Abeillé, L. Clément, and F. Toussnel.

*Building a treebank for French*, pages 165–188.

*Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers, 2003.



R. H. Baayen and R. Lieber.

Word frequency distributions and lexical semantics.

*Computers and the Humanities*, 30:4:281–291, 1997.



R. H. Baayen and A. Renouf.

Chronicling the Times: Productive Lexical Innovations in an English Newspaper.

*Language*, 72:69–96, 1996.



J. Flaitz.

*The Ideology of English: French Perceptions of English as a World Language*.

Mouton de Gruyter, Berlin, 1988.



S.G. Thomason.

*Language Contact*.

# POS breakdown of data

Part of Speech	English	Non-English	Total
Nouns	119	129 (131)	248 (250)
Adjectives	20	2	22
Idiomatic and Multi-word Expressions	3	2	5
Adverb	0	1	1
Determiner	1	0	1
Preposition	1	0	1

**Table:** A breakdown of part of speech tokens according to languages. Figures including the lemma *deutschemark* in parentheses.

# Example borrowings

Borrowing	Frequency	Borrowing	Frequency
cash flow	2	stand - by	3
check - up	1	struggle for life	1
citizen' s charter	1	success story	2
classless society	1	sustainable	1
come - back	1	swaps	1
cross borders	1	teddy	1
debt deflation	1	the	1
deregulation	1	top - down	1
discount	1	trade unions	1
downgrading	1	training groups	1
flint glass	1	Karenztag	2
hedge funds	3	chapka	1
industrial design	1	ejido	9
lease - back	2	glasnost	2
lobbying	3	huasipongo	2

# Examples of written forms queried in the corpus

xxx- = prefix

-xxx = suffix

á	grund-	ö
ä	-gut	-platz
-ag	-haft	-qab
-ah	haupt-	qe
and	-hen	qh
-anh	hoch-	qi
-arm	-huan	qs
auf	í	-sai
aus-	-ial	sch
bank	-iang	sf-
-bar	ich	sh
-betont	-ig	spr-
-chen	ijt	-ss
-chi	-ing	ß
-cional	-ism	-tad