



The Socio-Pragmatics of Anglicisms: A barren wasteland?

Eline Zenner (FWO Flanders),
Dirk Geeraerts & Dirk Spielman



University of Leuven
RU Quantitative Lexicology and Variational Linguistics

Overview

1. Background: the evolution of “quantitative” usage-based anglicism research
2. Overcoming the Bottleneck: presenting DAD
3. Negating the Wasteland: DAD’s theoretical possibilities



Overview

1. Background: the evolution of “quantitative” usage-based anglicism research
2. Overcoming the Bottleneck: presenting DAD
3. Negating the Wasteland: DAD’s theoretical possibilities

Overview

1. Background: the evolution of “quantitative” usage-based anglicism research

Krauss (1958) The increasing use of English words in German. *The German Quarterly* 31, 4, 272-286

Carstensen (1965) *Englische Einflüsse auf die deutsche Sprache nach 1945*. Heidelberg: Carl Winter

Yang (1990) *Anglizismen im Deutschen*. Tübingen: Max Niemeyer Verlag

Fink (1997) *Von Kuh-Look bis Fit for Fun*. Frankfurt am Main: Peter Lang

Onysko (2008) *Anglicisms in German*. Berlin/New York: Walter de Gruyter

“quantitative” usage-based anglicism research

The same pattern comes back in different shapes:

- **Data:** newspaper corpora, leading to listings of anglicisms
- **Main research lines:** → anecdotal enumeration of anglicisms, complemented with corpus examples
→ structuralist (quantified) account of spread & adaptation
- **Side comments:** → lexical variation (why the anglicism?)
→ what about longer stretches of English
→ what determines the life-span of anglicisms?

“quantitative” usage-based anglicism research

The same pattern comes back in different shapes:

- **Data:** newspaper corpora, leading to listings of anglicisms
- Main research lines: → anecdotal enumeration of anglicisms, complemented with corpus examples
→ structuralist (quantified) account of spread & adaptation
- Side comments: → lexical variation (why the anglicism?)
→ what about longer stretches of English
→ what determines the life-span of anglicisms?

“quantitative” usage-based anglicism research

The same pattern comes back in different shapes:

- Data: newspaper corpora, leading to listings of anglicisms
- **Main research lines:** → anecdotal enumeration of anglicisms, complemented with corpus examples
 - structuralist (quantified) account of spread & adaptation
- Side comments:
 - lexical variation (why the anglicism?)
 - what about longer stretches of English
 - what determines the life-span of anglicisms?

Krauss (1958)

“In the entertainment world - movies, music, radio, television - we find a large number. This is especially noticeable in film reviews. **We note such loans as** Show-Business, Come-back, Double, Musical, and Make-up in addition to familiar words like Star, Film, and En-gagement” (p.274)

Fink (1997)

- “- **Allround Fitness Center**
- **Alpin-Freak**: ‘Die Jet-Hose von Colmar ... begeistert jeden Alpin-Freak.’
- **Anti-Müffel-Shirt**: ‘Das Anti-Müffel-Shirt. Wenn das Deo allein zuwenig ist.’
- **Are you ready to touch the world of Sport Freaks?**
- **Babyface-Typ**: ‘Durstige Babyface-Typen sollten stets ihren Ausweis mitnehmen: denn Alkohol gibt es erst ab 21 Jahren’. ” (p.66/67)



“quantitative” usage-based anglicism research

The same pattern comes back in different shapes:

- Data: newspaper corpora, leading to listings of anglicisms
- Main research lines: → anecdotal enumeration of anglicisms, complemented with corpus examples
→ structuralist (quantified) account of spread & adaptation
- Side comments: → lexical variation (why the anglicism?)
→ what about longer stretches of English
→ what determines the life-span of anglicisms?

Carstensen (1965)

“Der Plural geht in den meisten Fällen auf –s aus: *Count-downs, Callgirls, Eggheads* etc. Entlehnungen auf –er schliessen sich an die zahlreichen dt. Wörter auf –er an (wie z.B. *Mieter, Meister*), die nur im Dativ Plural ein *n*, sonst keine Pluralendung haben” (p.67)

Yang (1990)

Jahrgang	1950	1960	1970	1980
Substantive	574	1023	3241	4443
Verben	50	45	168	199
Adjektive	9	43	134	119
Adverbien	0	3	14	5

Onysko (2008)

anglicism	determinant	determinatum	sum (types)
<i>Film</i>	297	162	459
<i>Computer</i>	367	58	425
<i>Internet</i>	354	3	357
<i>Test</i>	136	152	288
<i>Manager</i>	16	261	277



“quantitative” usage-based anglicism research

The same pattern comes back in different shapes:

- Data: newspaper corpora, leading to listings of anglicisms
- Main research lines: → anecdotal enumeration of anglicisms, complemented with corpus examples
→ structuralist (quantified) account of spread & adaptation
- Side comments: → lexical variation (why the anglicism?)
→ what about longer stretches of English
→ what determines the life-span of anglicisms?

“quantitative” usage-based anglicism research

Why do the side comments creep up so often, but aren't they ever dealt with?

1. quantitative accounts do not exploit the possibilities of statistical technique
2. data collections are too small to answer questions that rely heavily on high token frequencies / high TTR's

Carstensen (1965)

“Der Plural geht in den meisten Fällen auf –s aus: *Count-downs, Callgirls, Eggheads* etc. Entlehnungen auf –er schliessen sich an die zahlreichen dt. Wörter auf –er an (wie z.B. *Mieter, Meister*), die nur im Dativ Plural ein *n*, sonst keine Pluralendung haben” (p.67)

Yang (1990)

Jahrgang	1950	1960	1970	1980
Substantive	574	1023	3241	4443
Verben	50	45	168	199
Adjektive	9	43	134	119
Adverbien	0	3	14	5

Onysko (2008)

anglicism	determinant	determinatum	sum (types)
<i>Film</i>	297	162	459
<i>Computer</i>	367	58	425
<i>Internet</i>	354	3	357
<i>Test</i>	136	152	288
<i>Manager</i>	16	261	277



“quantitative” usage-based anglicism research

Why do the side comments creep up so often, but aren't they ever dealt with?

1. quantitative accounts do not exploit the possibilities of statistical technique
2. data collections are too small to answer questions that rely heavily on high token frequencies / high TTR's

“quantitative” usage-based anglicism research

Why do the side comments creep up so often, but aren't they ever dealt with?

1. quantitative accounts do not exploit the possibilities of statistical technique
2. data collections are too small to answer questions that rely heavily on high token frequencies / high TTR's

Compare:

distribution over POS: <i>manager vs. cool</i>	synonym distribution: 160 <i>managers</i> vs. 50 <i>bedrijfsleiders</i>
gender assignment: <i>manager vs. coolness</i>	analysis of use of fixed expressions: <i>as good as it gets</i> in context
↓ type-oriented	↓ token-oriented / TTR

“quantitative” usage-based anglicism research

Why do the side comments creep up so often, but aren't they ever dealt with?

1. quantitative accounts do not exploit the possibilities of statistical technique
2. data collections are too small to answer questions that rely heavily on high token frequencies / high TTR's. **WHY?**

“quantitative” usage-based anglicism research

Why do the side comments creep up so often, but aren't they ever dealt with?

1. quantitative accounts do not exploit the possibilities of statistical technique
2. data collections are too small to answer questions that rely heavily on high token frequencies / high TTR's. **WHY?**

→ Methodological bottleneck in extraction methods


“quantitative” usage-based anglicism research

Methodological bottleneck : manual extraction of anglicisms

	Material	Extraction	Ang.Types	Ang.Tokens	Analyses
Krauss	“since 1956”	manual	?	?	qualitative
Carstensen	“1961-1964”	manual	?	2 per page	qualitative
Yang	24 issues	manual	3646	10 070	descriptive
Fink	10 issues	manual	?	?	descriptive
Onysko	5 million words	man/TTR	16 663	57 591	descriptive

“quantitative” usage-based anglicism research

Methodological bottleneck : manual extraction of anglicisms



	Material	Extraction	Ang.Types	Ang.Tokens	Analyses
Krauss	“issues 1950”		?	?	qualitative
Carstensen				per page	qualitative
Yang	24 issues	manual	3646	10 070	descriptive
Fink	10 issues	manual	?	?	descriptive
Onysko	5 million words	man/TTR	16 663	57 591	descriptive

- minimally exhaustive
+ maximal control

occasional attempts to automation

Pfitsner & Romsdorfer (2003), Farrugia (2005)

→ unevaluated, or evaluated on small sets

Alex (2005, 2008)

→ valuable

→ basic heuristic fully automatic (dictionary look-ups + web-based module)

→ full-fledged automation has led to:

- exclusion of established anglicisms

(*computer, film, manager, job, ...*)

- neglect of precise and complete MWU-identification

- remaining noise

+ maximally exhaustive

- minimal control

?

occasional attempts to automation

Pfitsner & Romsdorfer (2003), Farrugia (2005)

→ unevaluated, or evaluated on small sets

Alex (2005, 2008)

→ valuable

→ basic heuristic fully automatic (dictionary look-ups + web-based module)

→ full-fledged automation has led to:

- exclusion of established anglicisms

(*computer, film, manager, job, ...*)

- neglect of precise and complete MWU-identification

- remaining noise

Due to **minimal control**

occasional attempts to automation

Pfitsner & Romsdorfer (2003), Farrugia (2005)

→ unevaluated, or evaluated on small sets

Alex (2005, 2008)

→ valuable

→ basic heuristic fully automatic (dictionary look-ups + web-based module)

→ full-fledged automation has led to:

- exclusion of established anglicisms

(*computer, film, manager, job, ...*)

- neglect of precise and complete MWU-identification

- remaining noise

Lack of exhaustivity

Summary

Existing anglicism research

- focus on structuralist aspects like the adaptation of anglicisms.
- socio-pragmatic issues are noticed, but neglected
- caused by data-sparseness
- the methodological bottleneck in extraction methods has to be overcome



Maximum control (attained)

Minimum exhaustivity

Existing attempts to automation

- very few & preliminary attempts
- Alex (2005, 2008), but focus on possibilities of language processing, not on workable end-result
- neglect of established loans, noise and MWU's
- how to tailor to the needs of socio-pragmatic analyses?



Maximum exhaustivity (aspired)

Minimum control

Summary

Existing anglicism research

- focus on structuralist aspects like the adaptation of anglicisms.
- socio-pragmatic issues are noticed, but neglected
- caused by data-sparseness
- the methodological bottleneck in extraction methods has to be overcome

Existing attempts to automation

- very few & preliminary attempts
- Alex (2005, 2008), but focus on possibilities of language processing, not on workable end-result
- neglect of established loans, noise and MWU's
- how to tailor to the needs of socio-pragmatic analyses?



maximally exhaustive
maximal control

?

Summary

Existing anglicism research

- focus on structuralist aspects like the adaptation of anglicisms.
- socio-pragmatic issues are noticed, but neglected
- caused by data-sparseness
- the methodological bottleneck in extraction methods has to be overcome

Existing attempts to automation

- very few & preliminary attempts
- Alex (2005, 2008), but focus on possibilities of language processing, not on workable end-result
- neglect of established loans, noise and MWU's
- how to tailor to the needs of socio-pragmatic analyses?



“quantitative” usage-based anglicism research

Database of Anglicisms in Dutch

	Material	Extraction	Ang.Types	Ang.Tokens	Analyses
Krauss	“since 1956”	manual	?	?	qualitative
Carstensen	“1961-1964”	manual	?	2 per page	qualitative
Yang	24 issues	manual	3646	10 070	descriptive
Fink	10 issues	manual	?	?	descriptive
Onysko	5 million words	man/TTR	16 663	57 591	descriptive
DAD	1 billion words	semi-aut	> 100 000	> 50 million	inferential

“quantitative” usage-based anglicism research

Database of Anglicisms in Dutch

	Material	Extraction	Ang.Types	Ang.Tokens	Analyses
Krauss	“since 1956”	manual	?	?	qualitative
Carstensen	“1961-1964”	manual	?	2 per page	qualitative
Yang	24 issues	manual	3646	10 070	descriptive
Fink	10 issues	manual	?	?	descriptive
Onysko	5 million words	man/TTR	16 663	57 591	descriptive
DAD	1 billion words	semi-aut	> 100 000	> 50 million	inferential

mwu's included

Overview

1. Background: the evolution of “quantitative” usage-based anglicism research
2. Overcoming the Bottleneck: presenting DAD
3. Negating the Wasteland: DAD’s theoretical possibilities



DAD: the Database of Anglicisms in Dutch

Material: Parsed Newspapers Corpora

- TwNC: Netherlandic Dutch from 1999 to 2002
 - LeNC: Belgian Dutch from 1999 to 2005
- } > 1 billion words

Goal:

tokenbased database of all English found in the corpus, accounting for:

- Dutch vs. English articles (language guesser)
- single tokens vs. MWU's and quotes (*computer vs. venture capital*)
- common vs. proper (*as good as it gets vs. As Good As It Gets*)

Method

automated where possible (Python) → exhaustivity
complemented with manual coding → control



Method: overview

Type-Based

- Extraction methods

Token-Based

- Getting out English articles
- Common or Proper?
- Identifying English MWU's

No one's perfect: what DAD can't do



Extraction Methods

Goal: a first rough list of all English in the corpora (*computer, whose, venture, capital, you, ...*)

Method:

- Type-based
- *Automated:* list-matching → exhaustivity
- *Manual:* filtering noise → control

Automated: list-matching

Frequencylists for the corpora

- TwNC (POS)
- LeNC (no POS)

bodembedekking
burka
Bush
...
citroen
cd
...
word
you
Youri

Indexlist for English WordNet

- + Quirk (1985)
- + RE's for conjugated forms

ankle
...
burka
bush
...
cd
...
undergarment
word
you



Automated: list-matching

- capitalisation is disregarded
- only full matches are allowed

bodembedekking
burka
Bush
...
citroen
cd
...
word
you
Y ouri

ankle
...
burka
bush
...
cd
...
undergarment
word
you

QML

Automated: list-matching

Result:

> 55 000 types

→ what is going wrong?

> 400 million tokens

bodembedekking
burka
Bush
...
citroen
cd
...
word
you
Y uri

ankle
...
burka
bush
...
cd
...
undergarment
word
you

Automated: list-matching

Result:

> 55 000 types

→ what is going wrong?

> 400 million tokens

bodembedekking
burka
Bush
...
citroen
cd
...
word <i>ik word gezien</i> - 2 million tokens
you
Yuri

ankle
...
burka → not english
bush
...
cd → WSD
...
undergarment
word <i>a wonderful word</i> → English in English context
you

QML

Manually: Filtering out Noise

Task: Classifying the 55 000 types in five groups:

- Definitely English (*computer, acceptance, abundancy*)
- Definitely not English (*burka, auberge, aubergine, avantgarde*)
- Unclear cases (*cd, sport, bandage*)
- English in an English Context (*word, die, alarm, bombardment*)
- Proper nouns (*Clinton, MacBeth*)

Sources:


- Van Dale Groot Woordenboek van de Nederlandse Taal
- Collin's Cobuild
- Oxford English Dictionary
- Wikipedia



Results

	Types	Tokens
Definitely English	30 000	10 million
Definitely not English	15 000	300 million
Unclear cases	4000	6 million
English in an English context	5500	122 million
Proper nouns	1500	4 million

Results

	Types	Tokens
Definitely English	30 000 	10 million
Definitely not English	15 000	300 million
Unclear cases	4000	6 million
English in an English context	5500	122 million
Proper nouns	1500	4 million

Results

	Types	Tokens
Definitely English	30 000	10 million
Definitely not English	For now, these will remain unclear (see remarks)	
Unclear cases	4000	6 million
English in an English context	5500	122 million
Proper nouns	1500	4 million

Results

	Types	Tokens
Definitely English	Top 8 (van, met, of, die, in, over, we, word) accounts for 100 million tokens	
Definitely not English	Only English in English Context, all others can be deleted	
Unclear cases	→ Identifying English Contexts?	
English in an English context	5500	122 million
Proper nouns	1500	4 million

Method: overview

Type-Based

- Extraction methods

Token-Based

- Getting out English articles
- Common or Proper?
- Identifying English MWU's



No one's perfect: what DAD can't do

From Types to Tokens

Type-Based List (after deleting “definitely not English”)

alphabetical

Token-Based List

“chronological”, giving each token that occurs in the type-based list a special code

e.g. *Op zijn computer zocht Clinton informatie over venture capital. Zijn computer werkte niet mee en hij zuchtte: “what a wonderful world”.*



From Types to Tokens

<sentence 00-01>

Op zijn *computer* zocht *Clinton* informatie rond *venture capital*.

<sentence 01-02>

Zijn *computer* werkte niet mee en hij zuchtte: “*what a wonderful world*”.

computer	<FN.S00-01.W02-03>
Clinton	<FN.S00-01.W04-05>
venture	<FN.S00-01.W07-08>
capital	<FN.S00-01.W08-09>
computer	<FN.S01-02.W01-02>
what	<FN.S01-02.W08-09>
a	<FN.S01-02.W09-10>
wonderful	<FN.S01-02.W10-11>
world	<FN.S01-02.W11-12>



Token-Based

1. Getting out English articles
2. Common or Proper?
3. Identifying English MWU's



Token-Based

1. Getting out English articles
 - language guesser
 - indicating for each element whether it is part of an English article
2. Common or Proper?
3. Identifying English MWU's



Token-Based

1. Getting out English articles
2. Common or Proper?
Based on capitalization and position within the sentence, taking punctuation into account
3. Identifying English MWU's



Token-Based

1. Getting out English articles
2. Common or Proper?
3. Identifying English MWU's

string identification

- Alpino Parser (automated)
- tokenbased list (semi-automated)

→ exhaustivity

string-coding:

- language choice (semi-automated)
- type of proper name (manual)

→ control



String Identification

Alpino

- Using the parser's MWU-tagger
- Checking for every token in the tokenbased list if it is part of an MWU
- If so, conflate all lines belonging to the MWU

computer	<FN.S00-01.W02-03>
Clinton	<FN.S00-01.W04-05>
venture	<FN.S00-01.W07-08>
capital	<FN.S00-01.W08-09>
computer	<FN.S01-02.W01-02>
what	<FN.S01-02.W08-09>
a	<FN.S01-02.W09-10>
wonderful	<FN.S01-02.W10-11>
world	<FN.S01-02.W11-12>

computer	<FN.S00-01.W02-03>
Clinton	<FN.S00-01.W04-05>
venture capital	<FN.S00-01.W07-09>
computer	<FN.S01-02.W01-02>
what	<FN.S01-02.W08-09>
a	<FN.S01-02.W09-10>
wonderful	<FN.S01-02.W10-11>
world	<FN.S01-02.W11-12>



String Identification

Complementation: the tokenbased list

- Using the unique codes in the list
- Looking for successions of elements
- Allowing for interruptions by punctuation
- Manual check-up (but learning process)

computer	<FN.S00-01.W02-03>
Clinton	<FN.S00-01.W04-05>
venture capital	<FN.S00-01.W07-09>
computer	<FN.S01-02.W01-02>
what	<FN.S01-02.W08-09>
a	<FN.S01-02.W09-10>
wonderful	<FN.S01-02.W10-11>
world	<FN.S01-02.W11-12>



String-Coding

MWU Language-Coding

- English / hybrid / not English
- based on language codes of the elements
- automatic where possible, manual where needed (learning process)

Proper-Coding (manual for freq ≥ 5)

- title/product
- event/organisation
- person
- location
- disease/nature

computer	<FN.S00-01.W02-03>
Clinton	<FN.S00-01.W04-05>
venture capital	<FN.S00-01.W07-09>
computer	<FN.S01-02.W01-02>
what a wonderful world	<FN.S01-02.W08-12>



DAD

Token	Position	English art	language	type	PROP	PROP-type
computer	<FN.S00-01.W02-03>	NO	English	ST	COM	/
Clinton	<FN.S00-01.W04-05>	NO	English	ST	PROP	person
venture capital	<FN.S00-01.W07-09>	NO	English	MWU	COM	/
computer	<FN.S01-02.W01-02>	NO	English	ST	COM	/
what a wonderful world	<FN.S01-02.W08-12>	NO	English	MWU	COM	/

Method: overview

Type-Based

- Extraction methods

Token-Based

- Getting out English articles
- Common or Proper?
- Identifying English MWU's

No one's perfect: what DAD can't do



What DAD can't do

- **No hybrids** due to the list-matching technique (only full matches)
 - some are found in the MWU-identification process
 - others can be traced easily (e.g. all compounds with *manager*)

- 6 million unclear cases
 - a priori: exoticisms, eponyms, neo-classical forms, ...
 - WSD: context per token?
 - patch-work: using weighting techniques (e.g. *0.5)
 - sampling: disambiguate manually for subsets
 - long-term solution: automatic WSD

What DAD can't do

- No hybrids due to the list-matching technique (only full matches)
 - some are found in the MWU-identification process
 - others can be traced easily (e.g. all compounds with *manager*)
- 6 million unclear cases
 - **a priori**: exoticisms, eponyms, neo-classical forms, ...
 - **WSD**: context per token?
 - patch-work: using weighting techniques (e.g. *0.5)
 - sampling: disambiguate manually for subsets
 - long-term solution: automatic WSD



Overview

1. Background: the evolution of “quantitative” usage-based anglicism research
2. Overcoming the Bottleneck: presenting DAD
3. Negating the Wasteland: DAD’s theoretical possibilities



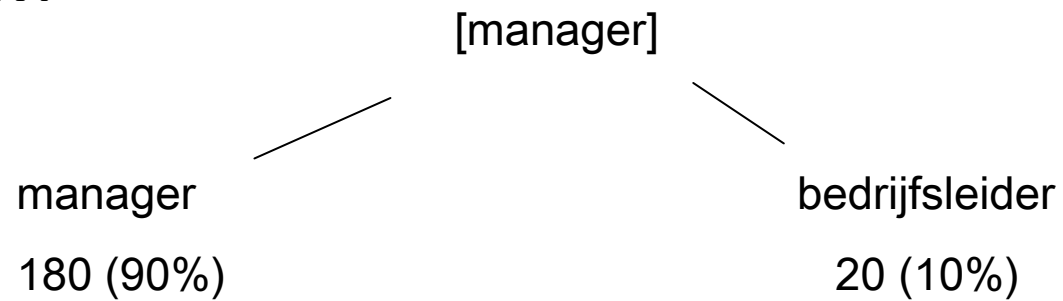
Main Questions

- lexical variation (why the anglicism?)
- what about longer stretches of English?
- what determines the life-span of anglicisms?

Socio-Lexicology

Why use the anglicism?

Profile-based



Variation in the succes of the anglicism?

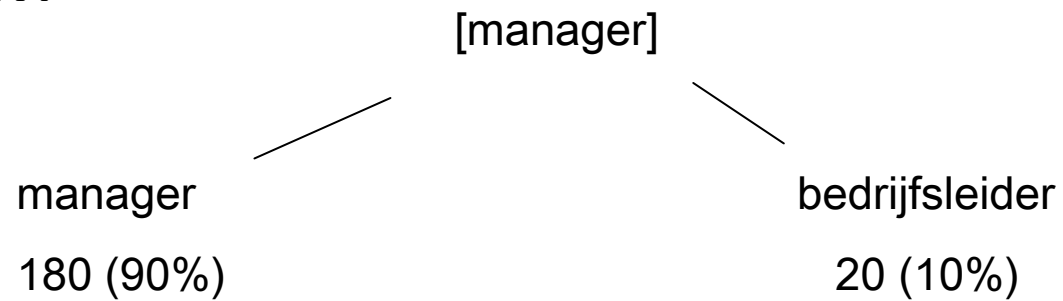
Socio-Lexicology

Manager	86780	Bedrijfsleider	8632
Babysit(ter)	2673	Oppas	7395
Expat	619	Emigrant	2487
Diehard	417	Volhouder	484
Loser	2678	Verliezer	19539

Socio-Lexicology

Why use the anglicism?

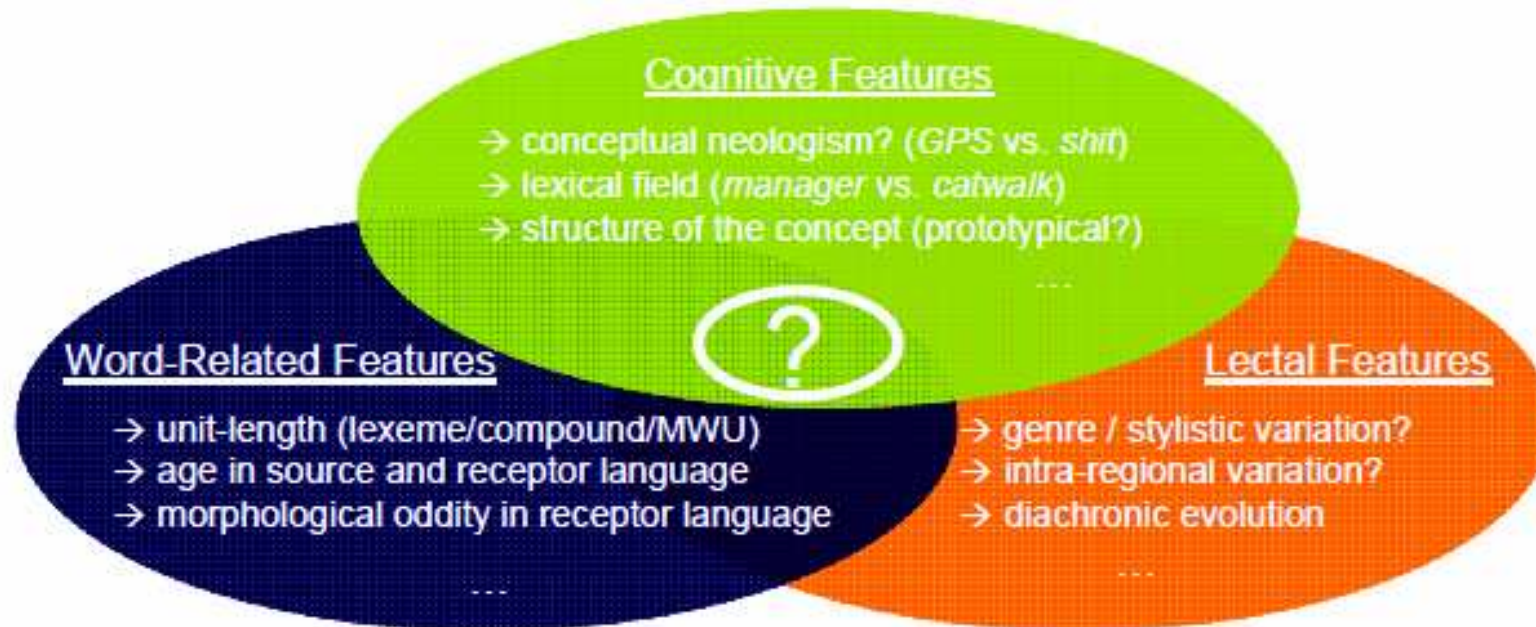
Profile-based



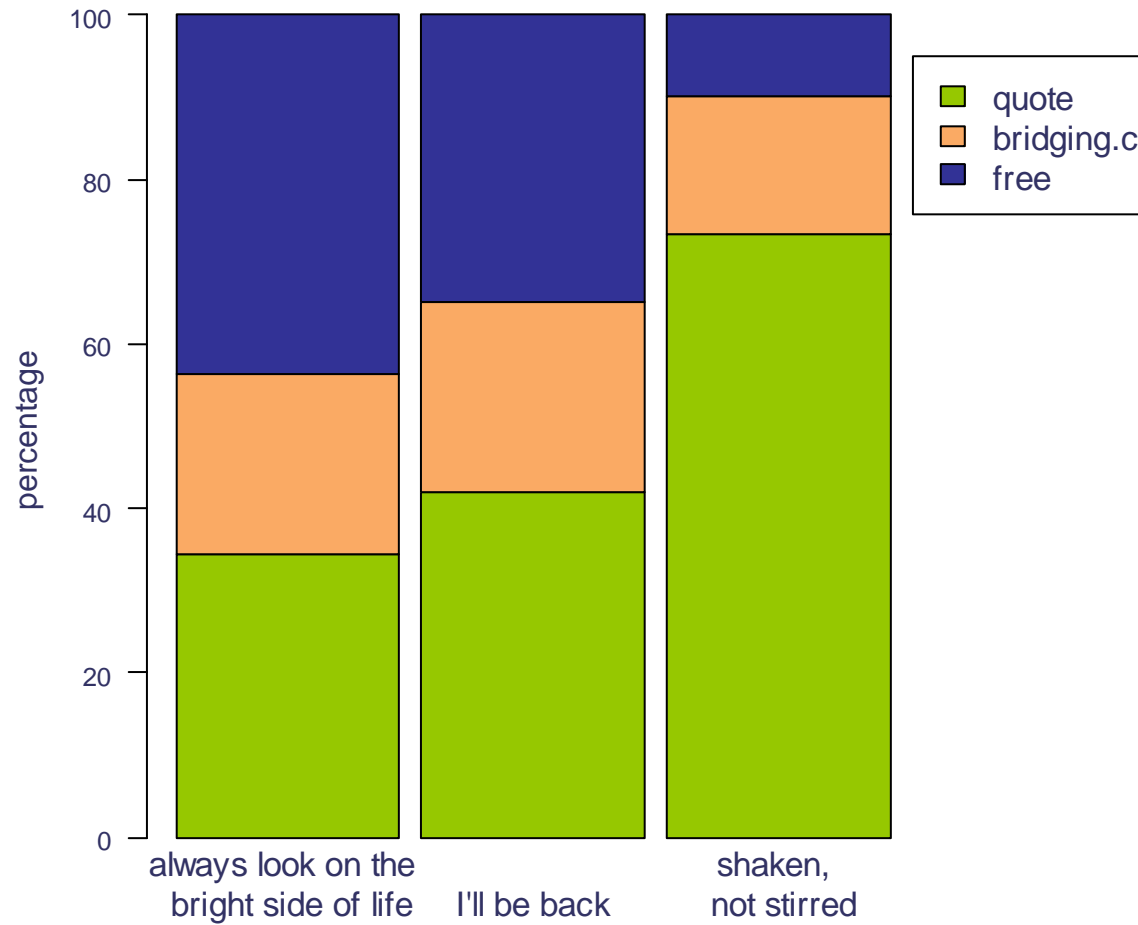
Variation in the succes of the anglicism?

Socio-Lexicology

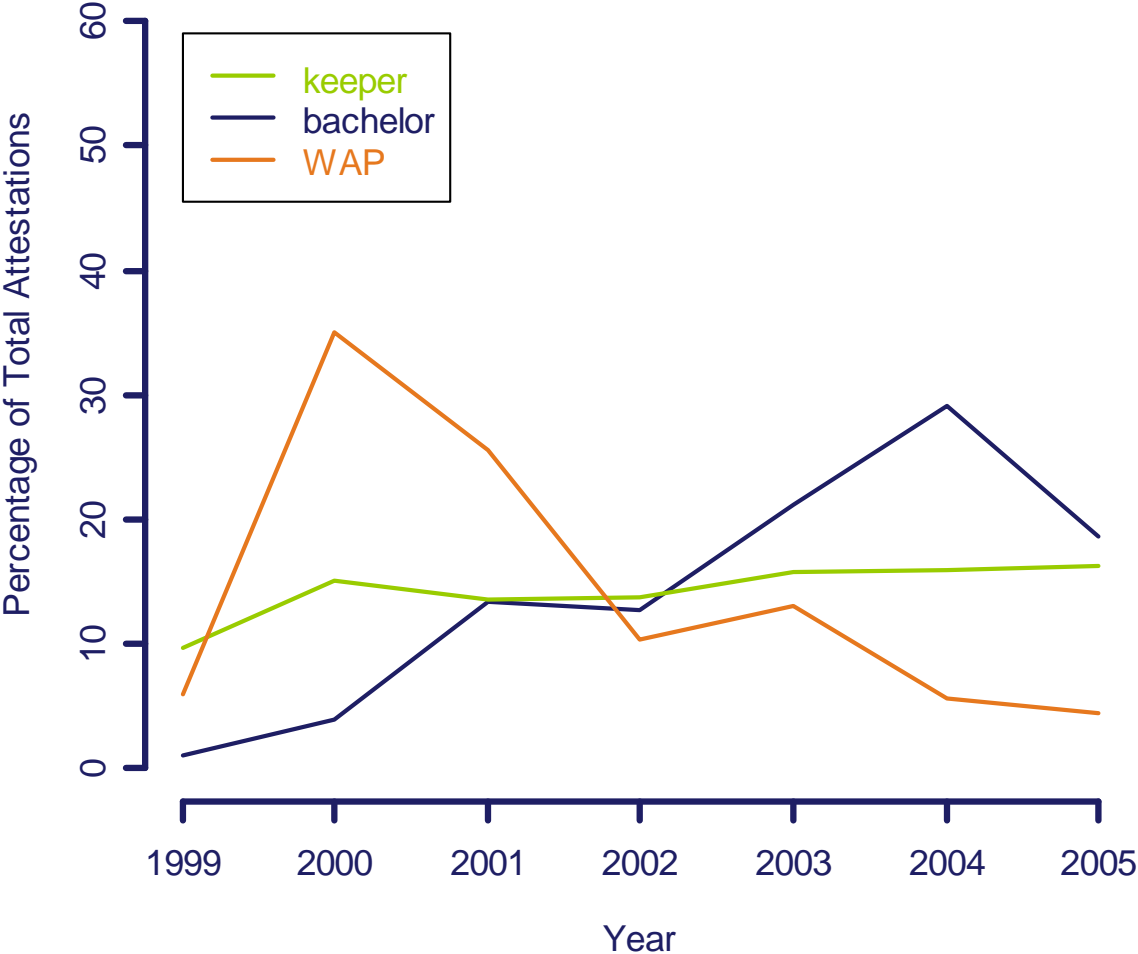
Why use the anglicism?



Phraseology: catchphrases



Survival Analysis





For more information:
<http://www.ling.arts.kuleuven.be/qlv>
eline.zenner@arts.kuleuven.be