

Eline Zenner, Dirk Speelman, Dirk Geeraerts
FWO Flanders/University of Leuven, University of Leuven, University of Leuven
eline.zenner@arts.kuleuven.be, dirk.speelman@arts.kuleuven.be,
dirk.geeraerts@arts.kuleuven.be

A concept-based approach to measuring the success of loanwords

Introduction

This paper presents a multivariate, quantitative, corpus-based study in which we examine, by means of mixed effect regression modeling, the combined effect of a set of structural, conceptual, extra-linguistic and contact-related features affecting the success of 125 English person reference nouns (*manager, babysitter* ...) in Dutch.

Background

THEORETICAL ACCOUNTS ON LEXICAL BORROWING often focus on the demarcation of types of loanwords (e.g. Haugen 1950, Filipovic 1977, Field 2002). Classifications are typically based on (1) objectively defined structural characteristics of the loanword (e.g. morphological adaptation); (2) subjective distinctions serving normative purposes, like the division between the tolerated “necessary” loanwords (i.e. those naming a new concept) and the denounced “luxury” loanwords (i.e. those introduced as alternative lexicalization for an existing concept) (see Onysko & Winter-Froemel *forthcoming* for discussion).

Current QUANTITATIVE CORPUS-BASED ANGLICISM RESEARCH aims at counting the number of English loanwords within the proposed classifications, trying to identify those classes which are most susceptible to foreign influence (e.g. Yang 1990, Nettmann-Multanowska 2003). Typically, the focus is restricted to classifications based on structural characteristics (i.e. the first class presented above). Furthermore, most studies show several methodological shortcomings, which have hitherto remained unaddressed. As such, a fully comprehensive empirical attempt at pin-pointing the locus of foreign influence is still missing.

The *aim of our paper* is twofold. First, we wish to encourage the quantitative perspective in loanword research by presenting ways to clear the recurring methodological hurdles, which we will present below. Second, we wish to open up the theoretical scope of current research by presenting objective operationalizations for the so far neglected distinctions (e.g. luxury vs. necessary loans) and by introducing new conceptual and extra-linguistic features that possibly influence variation in the success of anglicisms.

Methodology: Clearing the hurdles

DATA-COLLECTION: As most studies extract anglicisms manually from their sources, existing corpus-based anglicism research has typically relied on relatively small datasets (e.g. Fink 1997). However, empirical analyses dealing with lexicology and lexical variation require sufficiently large data collections (Armstrong 2001, Geeraerts 2010). For our analyses on variation in the success of 125 English loanwords we rely on a lemmatised and syntactically parsed Dutch newspaper corpus containing over 1.5 billion words. The corpus comprises material from 1999 to 2005 and represents the two national varieties of Dutch (Belgian Dutch and Netherlandic Dutch).

SUCCESS MEASURE: The second hurdle to be cleared is the definition of the success of a loanword. So far, this success has simply been equated with the token count for the borrowed item in a given corpus. However, as topic specificity can easily distort these counts and hence disfigure the results (Speelman *et al.* 2003), a more advanced success measure is required (e.g. Van Hout & Muysken 1994).

We propose a concept-based approach, taking alternative expressions into account. Adapting the profile-based method introduced by Geeraerts *et al.* (1999), we define the success of an anglicism as the corpus frequency of that anglicism, relative to the total frequency of the loanword and its synonyms (see example in Table 1).

TABLE 1. CONCEPT-BASED SUCCESS MEASURE.

Synonymous expressions for the concept BABYSITTER	Token frequencies in the Dutch corpus
babysit(ter)	1031
kinderoppas	456
success rate for <i>babysit(ter)</i> : $1031 / (1031 + 465) = 69.33 \%$	

Our conceptualization of the success of a loanword is thus the extent to which it succeeds in “fighting off” alternative lexicalisations for a given concept. As such, maximal success is the situation in which the loanword is (or has become) the only occurring lexicalization for a given concept.

DATA COLLECTION: For this study we focus on 125 English person reference nouns (*fulltimer, freak...*) used in Dutch. Only lexical items which are clearly identifiable as English loanwords by native speakers of Dutch are included as anglicisms. Denotational synonyms for the loanwords were identified using multiple lexicographical sources, complemented with results from automatic synonym detection based on word space models (see Lin 1998 and Peirsman *et al.* 2007). To avoid unreliable success-rates due to semantic specialization of the items in a profile, we specifically seek out profiles with a high degree of synonymy between the lexical items (avoiding fuzziness and near-synonymy as much as possible). Next, automated methods are introduced to retrieve tokens and remove noise (e.g. proper names, longer stretches of English) from the dataset. Then, we semi-automatically disambiguate remaining polysemous items (e.g. *chicken*). Using the resulting database, we finally measure the success rates for the set of 125 English nouns across eight subsections of the corpus (see example in Table 2). This way, we acquire 1000 success rates (8 per loanword), based on over 30 million tokens of person reference nouns.

OBJECTIVE DEFINITION OF PREDICTORS: As was mentioned above, current quantitative approaches only focus on the influence of structural features on variation in the success of loanwords. If we want to get a finer grip on the actual impact of the different classifications that have been proposed, a wider variety of features should be discussed. In this study, we therefore not only incorporate structural features, but also include (1) operationalizations for the distinction between necessary and luxury loans; (2) extra-linguistic features (e.g. regional variation - see Table 2); (3) conceptual features (e.g. entrenchment of the concept); (4) contact-related features (e.g. “travel time” from donor language to recipient language – Chesley & Baayen 2010).

STATISTICAL ANALYSES: A final step to a comprehensive empirical view on the success of loanwords is to inquire into the combined effect of the different influential features on the success measure. Using a logarithmic transformation of the success rate,

we perform mixed-effect linear regression analyses to study the interplay between the variables.

TABLE 2. SUCCESS-RATES ACROSS SUBSECTIONS OF THE CORPUS.

subcorpus:	success-rate for <i>babysit(ter)</i> in the subcorpus:
Belgian Dutch; 1999-2000; popular newspapers	$348 / (348 + 135) = 0.720$
Belgian Dutch; 1999-2000; qualitative newspapers	$102 / (102 + 48) = 0.680$
Belgian Dutch; 2001-2002; popular newspapers	$339 / (339 + 95) = 0.781$
Belgian Dutch; 2001-2002; qualitative newspapers	$130 / (130 + 74) = 0.637$
Netherlandic Dutch; 1999-2000; popular newspapers	$33 / (33 + 17) = 0.660$
Netherlandic Dutch; 1999-2000; qualitative newspapers	$34 / (34 + 24) = 0.586$
Netherlandic Dutch; 2001-2002; popular newspapers	$19 / (19 + 30) = 0.388$
Netherlandic Dutch; 2001-2002; qualitative newspapers	$26 / (26 + 33) = 0.441$

Results and Conclusion

The results show the overall usefulness of the multivariate approach: features from all distinguished classes (structural, conceptual, extra-linguistic and contact-related features) contribute to the explanation of variation found in the success rates of the 125 anglicisms under scrutiny. Furthermore we find that the most important factor in the model is the distinction between luxury and necessary loans: the status of a loanword at the time of introduction in a language has a long-term effect on its success rate. We discuss two converging interpretations for this effect: (1) the influence of language planning; (2) the effect of conceptual entrenchment. Overall, the results from our empirical study indicate how current structuralist theories of the borrowing process need to be complemented with a variationist, usage-based perspective.

References

- Armstrong, Nigel 2001. Social and stylistic variation in spoken French: a comparative approach. Amsterdam: John Benjamins.
- Chesley, Paula, Baayen, R. Harald 2010. Predicting new words from newer words: Lexical borrowings in French. *Linguistics* 48(6).
- Field, Fredric W. 2002. *Linguistic borrowing in bilingual contexts*. Amsterdam: John Benjamins.
- Filipovic, Rudolf 1977. English words in European mouths and minds. *Folia Linguistica* 11, 195-206.
- Fink, Hermann 1997. Von Kuh-Look bis Fit for Fun: Anglizismen in der heutigen deutschen Allgemein- und Werbesprache. Frankfurt am Main: Peter Lang.
- Geeraerts, Dirk 2010. Lexical variation in space. In *Language and Space I. Theories and Methods*, Auer, Peter, Schmidt, Jürgen Erich (eds.), 820-836. Berlin/New York: Mouton de Gruyter.
- Geeraerts, Dirk, Grondelaers, Stef, Speelman, Dirk 1999. *Convergentie en divergentie in Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut.
- Haugen, Einar 1950. The analysis of linguistic borrowing. *Language* 26(2), 210-231.
- Lin, Dekang 1998. Automatic Retrieval and Clustering of Similar Words. *Proceedings of COLING-ACL98*, Montreal (Canada).

- Nettmann-Multanowska, Kinga 2003. *English Loanwords in Polish and German after 1945: Orthography and Morphology*. Frankfurt am Main: Peter Lang.
- Onysko, Alexander and Winter-Froemel, Esme forthcoming. Necessary loans? Luxury loans? Explaining the pragmatic dimension of borrowing.
- Peirsman, Yves, Heylen Kris and Speelman, Dirk 2007. Finding semantically similar words in Dutch. Co-occurrences versus syntactic contexts. *Proceedings of the CoSMO workshop*, 9-16.
- Speelman, Dirk, Grondelaers, Stef and Geraerts, Dirk 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37, 317-337.
- Van Hout, Roeland and Muysken, Pieter 1994. Modeling lexical borrowability. *Language Variation and Change* 6(1), 39-62.
- Yang, Wenliang 1990. *Anglizismen im Deutschen: am Beispiel des Nachrichtenmagazins der Spiegel*. Tübingen: Niemeyer.