The crosslinguistic intertextuality of loanwords

Vejdemo, Susanne Vandewinkel, Sigi
Montero-Melis, Guillermo

December 20, 2012

1 Purpose and aims

The semantic meaning potential of words is to a large extent governed by their intertextual history of use (see ??). Loanwords are an interesting case, since they have a history in their source language, but lose some of their meaning potential when they are borrowed. There exists as yet no large scale investigation into the precise nature of this semantic loss - thus this project has two ultimate aims: to create a freely available descriptive database with measurable meaning change data, and to contribute to the theoretical knowledge on semantic change in loanword transfer.

1.1 Research Questions

The semantic meaning potential of words is to a large extent governed by their intertextual history of use (see ??). Loanwords are an interesting case, since they have a history in their source language, but lose some of their meaning potential when they are borrowed. We wish to examine the following topics:

1. How much of a word's meaning changes when it is borrowed (e.g. the term 'body guard', borrowed from English into Swedish), i.e how much of its meaning potential and history is lost?

2. After a loan word is established, how does it share the semantic space with already existing, seemingly synonymous native words (e.g. the Swedish term 'livvakt', body guard.)

3. How do X and Y compare across three European languages, i.e. Swedish, Dutch, and Spanish?

4. What factors determine whether a loanword is successfully integrated into a receiver language?

By using both quantitative experimental and corpus methods, as well as qualitative interview methods, we also wish to examine if there are marked discrepancies between the measurements of meaning and the subjective reported opinions about meanings of speakers.

1.2 Hypotheses

1. Borrowings will have the same semantic profiles as their (so stated) native equivalents.

2. If they do not have the same semantic profiles, they will differ in which registers they appear in.

3. If they do not have the same semantic profiles, they will differ in which referents they denote.

4. Compounded borrowings will be more easier integrated if their compound parts are also earlier borrowings - and the semantic profile of the compounded borrowing will be influenced by that of the compound part.

5. Likewise, the semantic profiles of borrowings will align with those of native cognates.

2 Survey of the field

2.1 Theory

2.1.1 Meaning (Wälchli and Cysouw)

As a word is borrowed from one language to another, its meaning changes - the number and kind of referents it represents can grow or shrink, and its register (in which social setting it is appropriate to use) can also vary. This projects seeks to investigate this change - how large is it, and are there recurrent patterns in the kinds of changes that occur to loan words?

This project aims to create both etic and emic definitions of the meaning, and meaning changes, of the loanwords. In an etic definition, the meaning of a word as defined as the set of its uses (**?**), or the set of its situated instances (**?**). From this perspective the meaning of the Swedish word *livvakt* or the Dutch word *lijfwacht* is constituted by all the situated instances of its usage taken together; this may then be contrasted to, for instance, the meaning of the English loanword *bodyguard*. In this, we follow the exemplar semantics research done by Wälchli & Cysouw (2012), and will take a denotational approach to meaning where similarity in form will be assumed to represent similarity in meaning. In an emic definition, by contrast, the commonalities behind the different uses are sought (**?**), what **?** call a "stabilized, institutionalized, and prototypical magnetic center that can be contextually interpreted in constrained ways" and which **?** refers to as the meaning potentials of a word. Seen from this perspective, the meaning of Swedish *livvakt* or Dutch *lijfwacht* is most closely related to that of a central prototypical sense with fuzzy boundaries.

The etic definitions will be investigated through the uses of the words in corpora (see section CL) and by acceptability judgments in the psycholinguistic research tradition (see section AJ). The emic definitions can aprtly be arrived of by a careful analysis of the commonalities in these two experiments, but the subconscious evaluations of the words by speakers will also be measured through psychological Semantic Differential experiments (see section SD).

2.1.2 Loanwords

Loanwords are the most straightforward way of studying languages in contact. They're highly visible, easily borrowable, and are subject to a measure of control by the speaker community (**?**). One main reason for borrowings consists of filling referential gaps – this is also the reason that the most common borrowings are nouns (cf. **?**, p.168).

Besides filling in referential gaps in the host language, the main reasons speakers borrow lexical items into their language have to do with various special effects: euphemisms, a need for trendiness and creativity (cf. **?**); or for humorous effects, expressiveness or group identity (cf. **??**). This entails that borrowings are usually not quite synonymous with the native alternative(s) available in the host language: it seems that connotational distinctions suffice to warrant the incorporation of borrowed vocabulary alongside denotationally-synonymous native items. We expect this to show up clearly in the results of our LSA research: after all, it is commonly accepted that (near-) synonyms need not share the same antonyms (cf. **?**).

It is common for loan words to become structurally integrated in the host language, phonetically as well as grammatically, with native phonemes and e.g. plurality markers substituting for the donor language's. The degree of structural integration into the host language is often a correlate of the level of bilingualism prevalent in the speaker community. Given that speakers of Swedish, Dutch and Spanish may show significant difference in their familiarity with or fluency in English (FOR REFERENCE SEE EMAIL), we expect to see differences here (FLESH OUT). Furthermore, the typological and structural commonalities between our three languages should ensure a low threshold for borrowability and make incorporation feasible, minimizing interference stemming from typological incompatibility. REPHRASE FOR BETTER FLOW.

For the purposes of this project, we will adopt **?**, p.37's definition of borrowings: "the incorporation of foreign features into a group's native language by speakers of that language", since it stresses the role of native speakers as agents in the process (see also **?**, p.12). Operationally speaking, this means that we will not be dealing with hapax legoumena, single-speaker innovations or loanwords that are not commonly accepted by the speaker communities. Also excluded are couplings of loanwords and native alternatives that occur in popular-scientific list drives. Only those loanwords that are in general usage, yet still engender prescriptivist pushback and that enjoy a relatively common native alternative will be considered. If existing lists of such pairs are any guide (cf. **?**) cf. Språkrådet) words that are part of what has been argued to be the core vocabulary of a language (see **?**) do not feature in them; instead, many are centred in the areas of technology, marketing and international relationships.

Clearly the "gap" hypothesis of borrowing is insufficient to explain all or even most of the items on the word lists. By scrutinizing the details of pairs of anglicisms / native alternatives this study will propose precise measurements for various denotational and connotational aspects of their respective meanings: denotational meaning loss/gain; dividing the semantic workload and the carving up of semantic space; ...

2.2 Method

2.2.1 Corpus linguistics

Creating Semantic Profiles
The increase of and access to computational power has made it possible to use large amounts of texts – corpora – to create semantic profiles for words. **?** have shown this for semantic investigations into temperature terms using a method known as Multidimensional Scaling (MDS). **?** has shown how semantic profiles can be made for the Swedish term *helig* using the method Latent Semantic Analysis (LSA).

Both MDS and LSA are based on comparing collocations of a words, that is, co-occurrences with those lexical items a few words of context to the left and right, with a more semantic representation of the semantic relationships of words. Semantic representations are made by measuring how often all the words in a multimillion word text corpus co-occur with all other words in the corpus - the result is an n-dimensional space where words that co-occur often cluster together. See **??** for general overviews of LSA, **?** for a general overview of vector models for semantic processing). This methodology has proven its worth in setting up semantic profiles in ways that allow for quantitative as well as qualitative analyses.

It is important to note here that we will not be normalising our data: all data points will be included in our analysis and none will be discarded as outliers in order to bring out the generalities more strongly (cf. **?**, p.678) and **?**, p.293).

Comparing Semantic Profiles
The semantic profiles we wish to compare can be represented as semantic vectors. Semantic vectors can be subtracted from one another, leaving a semantic difference vector - a measure of the difference between the semantic profiles (cf. **?**, p.19). Quantitatively, all words under investigation can be ranked according to how much they change: we can thus compare an English loanword with a native language equivalent.

We also aim to develop a quantitative measurement of the semantic difference between the English word in the English corpora and the English loandword in the Swedish corpora. In addition, this comparison will also be done more qualitatively, by looking at which other words cluster with the target word.

In order to get at the meaning of loanwords and their nearest synonyms in each language, we will use the theory and method of Latent Semantic Analysis, henceforth LSA (cf. **???**).

Latent Semantic Analysis
LSA applies statistical computations to large corpora in order to build a semantic space. The semantic space is derived by applying Singular Value Decomposition to a large $terms \times documents$ matrix, where the rows contain all the unique words that occur in the corpus, and the columns the documents (i.e. contexts or texts) which form the database.

In this semantic space word meanings are represented as vectors in a high-dimensional space[1] where semantic differences are represented in terms of dis-

[1] The choice of the dimension of the space is determined by the researcher, but typically it involves the 100 first dimensions or so extracted from a singular value decomposition. For

tance: the more different two senses are, the further apart they will be plotted from each other, and vice versa. Words with similar meaning are close to each other in this semantic space. Thus LSA offers a quantitative measure of *how similar* two words actually are, which is exactly what we wish to determine for each language.

Beyond inter-word similarity, LSA also allows to test whether two words are used in similar types of texts across languages. Thus, if we have external information about the texts of our database, e.g. that certain texts are, say, computer technology written by experts, we may see if a certain loanword is more often used in this type of texts or by a different set of users. Subsequently we can test if the same pattern applies in the other languages, as well as in the donor language.

2.2.2 Experimental

We will develop a series of tests in order to test the semantic and pragmatic intuitions of native speakers of different backgrounds. The two key issues here are: a) the use of a valid and reliable test battery, and b) the use of appropriate sampling techniques in order to be able to draw generalizations on a wider population.

Tests

It is useful to broadly differentiate between semantic and pragmatic tests, although they are necessarily related at a deeper level. ¡¡¡¡¡¡ HEAD The former will test the inherent *meaning* of a particular word (section 2.1.1). The latter are concerned with appropriateness of use in certain contexts.

Semantic tests These will include synonymy tests and semantic differential tests. In a synonymy test a participant is asked to rate similarity among words. In addition to pairwise ratings of similarity on a graded scale, we will use arrangement tasks, in which several words are placed on a plane according to their respective similarity (**??**). This will allow direct comparison to the similarity ratings obtained from LSA. Semantic differential tests (**?**) will be the other tool to obtain a characterization of the semantic space occupied by a certain word. The framework of semantic differential also offers an appropriate tool to measure attitudes towards words, which we believe will show to be a determinant factor in the use of loanwords, as put forward in section 2.1.2 on loanwords. Within semantic differential research it is well established that there are three cross-linguistically valid underlying factors that capture the attitude of speakers towards words: evaluation ('good–bad'), potency ('strong–weak'),and activity ('active–passive') (see **?**).

Pragmatic tests In order to determine the appropriateness of using loanwords or their native equivalents in certain contexts, we will use acceptability ratings and production tasks. Here we will directly draw on the database built up during the corpus phase. Our stimuli now will not be single words but text fragments of different length. From a pool of authentic contexts we will manipulate certain texts, replacing a loanword with its native equivalent or vice versa;

the mathematical details see **?**.

in other cases the text will be the same as the original; the third possibility is to leave out the word in order to elicit a word from the participant. All of the factors that might influence ratings will be crossed, and the design will be counterbalanced across participants. The participant's task will then be either to judge acceptability of text fragments (acceptability ratings), or to provide the word that they deem most appropriate in a text were the original item has been left blank (production task).

Sampling technique
Our goal is to ensure representative sampling of the population using web-based surveys. All of the above tests can easily be run through the internet. Moreover this method allows for massive collection of data without losing quality of data, and is rapidly gaining momentum in social research (**?**).

3 Project description

3.1 Phase I: Corpus phase

The goal of the initial phase of the project will be to have produced semantic profiles of the source words in the source language, the source words as borrowings in the target languages and the native equivalences.

The first step will be to create word specific literature overviews, using existing dictionaries and resources from other scholars who might have worked on the words. For Swedish, Svenska Akademien (**?**) and Sprkbanken (**?**) provide such resources; for Dutch, large collections have been published by **?** and **?**, which may be supplemented by large etymological dictionaries such as **?** and **?**.

After this step, the initial list of potential concepts will have been reduced to a more manageable size. We wish to retain concepts that have a clear extension, that are present in all the culture of all three language communities.

The next step will be to create semantic profiles using Latent Semantic Analysis, based on both source language and target language corpora. Two English corpora will be evaluated - the British National Corpus (BNC; 100m words) and the Corpus of contemporary American English (COCA; 450m words), access to which has been generously provided by Mark Davies (REF). For Dutch, both the INL corpus (http://www.inl.nl/) and perhaps the Corpus Gesproken Nederlands (Corpus of spoken Dutch ; http://lands.let.kun.nl/cgn/) will be used. For Swedish, the various corpora available at Språkbanken (http://www.spraak banken.gu.se) will be considered.

3.2 Phase II: Experimental phase

The experimental phase has in itself three broad goals: first, to *validate* the results from the Corpus phase (cf. section 3.1); second, to *explain* what factors determine that a certain loanword get integrated in the language, in the sense of being accepted by the speakers; and finally to *develop a methodology* consisting of different tests that can easily be replicated, making it possible to repeat measurements over time, and thus get an idea of how the studied phenomenon develops over time.

This phase will thus directly build on the results of the previous one, since the design of the experimental stimuli will be determined by the outcome of the corpus phase (cf. section 2.2.2). At a practical level, we will first construct an internet-based platform to run the experiments on-line[2]. Then we will collect the data using stratified sampling techniques to ensure representativity of the sample. Finally the data will be analysed using standard multivariate techniques to analyse semantic and attitude data. Here we expect that the results will validate the findings from the first phase.

It is important to emphasize that this methodology can become a replicable standard, which can be exactly replicated at other moments in time, say after five years. It can also be easily adapted to incorporate new loanwords that find their way into a language. Finally, given the degree of automatisation it will be adaptable to different languages if there exists a comparable-sized data base of texts for that language (note that the internet will provide such a corpus for relatively many of the world's languages).

3.3 Phase III: Compilation phase

The compilation phase of this project will focus on putting together a series of deliverables. These may be divided in three groups.

In the first place, the compilation phase will be concerned with arranging and writing up all the data gathered in the previous steps, and finalizing the product for publication. One of the deliverables of this project is a monograph with extensive appendices, with among others a database accessible online. This database will be available to other researchers and lexicographers, and we believe that this database and the methodology outlined in the monograph will be of use in a variety of theoretically-oriented and practically-oriented projects.

In the second place, the results arrived at will be written up as papers and submitted to high-profile peer-reviewed journals in the field of lexicology, semantics, pragmatics and typology; we project a further three or four deliverables for this project to be realised this way. Targets include *Journal of Pragmatics*, *Journal of Language Contact*, *Journal of Semantics*; as well as lower-profile journals such as the electronic journal *Lexis* and the annual *Språk i Norden.*

Furthermore, we will also submit a number of articles to popular-scientific publications in the language areas under study, notably the Swedish-language *Språktidningen*, the Dutch-language *Onze Taal* and the Spanish-language *Lenguaje y Textos.* Each of these regularly publish pieces on the influence of English, and we feel that they will be open to the findings of our rigorously empirically-based studies. We aim for at least three essays in these magazines.

In the third place, we will host a workshop on quantitative approaches to loanword studies at conferences traditionally centred on corpus linguistics and methodological issues in linguistics; an obvious target is ICAME or LREC. Further targets include business and translator conferences, such as the *Nordic Translation Conference* (the 2013 installment will be held 4-6 April in Norwich, UK) and the Canadian Association for Translation Studies Conference (the next installment, held 3-5 June 2013, is themed *Science in Translation*).

[2] There are different commercial solutions that offer appropriate interfaces to run this kind of surveys, and other studies published in high-profile journals have used this method (e.g., ?).

It is projected that meeting these goals will require approximately one year, but portions of the research will of course be submitted for publication before this phase, as soon as they become available.

4 Significance

The academic significance of this project is clear, and will be dealt with in the two following sections. In addition to this, there are also several matters of practical significance. One is to provide much needed input to the politicians and civil servants drafting language policies set to guard against the encroachment of English in Western Europe. Another is that the outcome of this research project can help dictionary makers and lexicographers. Finally, we believe that a principled execution of our methodology will serve as an illustration of its values and an example to future linguists.

In order to reach this audience, it is a stated goal of the project to not only produce academic output, but also to publish in popular linguistic journals such as *Onze Taal* (the Netherlands) and *Språktidningen* (Sweden), as well as to attend business conferences for translators and to be active in its collaboration with language prescriptivist agencies in both countries.

4.1 Descriptive value

The descriptive value of the project is manifold. One outcome will be the first a freely available database with detailed information on the use of loanwords in target language, source language, and, in addition, the native equivalence of the word in the target language. The published results will also include a very detailed dictionary of all the chosen loanwords, and their transition process. The semantic profiles created by LSA will be freely available, as will be the scripts that underlie their creation.

4.2 Theoretical value

The theoretical value of the project lies in the discussion of the research questions, namely clearer answers to the following:

- What factors predict loanword integration?
- How much of a word's meaning potential is lost in language transfer?
- How is the semantic space shared between loanwords and their native equivalents?

The monograph we intend to accompany the database and the dictionary will clearly outline the methodology used in arriving at our results, making it easy for other researchers to adopt and - or improve our methods.

5 Timeplan

In terms of scheduling, we expect each of the three phases outlined in Section 3 to take up approximately one year. The process of writing up our results and preparing the final documents for publication will take longer than that, but since our results will come in incremental batches, interpreting and compiling them is more of a continuous process running in parallel with the other phases.

References