The crosslinguistic intertextuality of loanwords

Vejdemo, Susanne Vandewinkel, Sigi
Montero-Melis, Guillermo

December 17, 2012

1 Purpose and aims

The semantic meaning potential of words is to a large extent governed by their intertextual history of use (ref Linell, ref Traugott & Dasher). 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 (ref Linell, ref Traugott & Dasher). 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 (Walchli 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 (Koptjevskaja-Tamm 2008), or the set of its situated instances (Evans 2009). 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 Wlchli & 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 (Koptjevskaja-Tamm 2008), what Traugott & Dasher 2002 call a "stabilized, institutionalized, and prototypical magnetic center that can be contextually interpreted in constrained ways" and which Linell (REF) 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 being referential gaps – this is also the reason that the most common borrowings are nouns (cf. Matras 2009:168).

Besides filling in referential gaps in the host language, the main reason speakers borrow lexical items into their language have to do with various special effects: euphemisms, a need for trendiness and creativity (cf. Rebuck 2002); or for humorous effects, expressiveness or group identity (cf. Gottlieb 2006; Wennberg 2010). 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. Miller et al 1990).

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).

We will not be dealing with true hapax legoumena, single-speaker innovations or not readily accepted loanwords; but only those words that are in general usage, yet still engender prescriptivist pushback. None of the words on the list are part of what has been argued to be the core vocabulary of a language (see Swadesh 1952): many are centred in the areas of technology, marketing and international relationships. Since the explicitly prescriptivist word lists include many single-morpheme native alternatives, the "gap" hypothesis of borrowing is clearly insufficient to explain them all. 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.

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. KOPTJEVSKAJATAMM-SAHLGREN has shown this for semantic investigations into temperature terms using a method known as Multidimensional Scaling (MDS). SIGSTRM 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(see REF for an overview of collocations), that is occurrences of the words with 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 DEERWEISTER, DUMAIS for general overviews of LSA, TURNEY-PANTEL http://www.jair.org/media/2934/live-2934-4846-jair.pdf for general overview of vector models for semantic processing).

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.(SIKSTRM SEMANTIC TESTS p19) 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 CITE Deerwester1990, Landauer1999, Dumais2004.

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 × 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 distance: 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.

[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 the mathematical details see CITEDeerwester1990.

2.2.2 Experimental

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. After this step, the initial list of potential concepts will have been reduced to a more manageable size. We wish to retain concept that have a clear extension, that is 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. For Dutch, both the INL corpus (http://www.inl.nl/) and perhaps Corpus Gesproken Nederlands (Corpus of spoken Dutch ; http://lands.let.kun.nl/cgn/) will be used. For Swedish, the various corpora available at Spkbanken (http://www.spraakbanken.gu.se) will be considered.

3.2 Phase II: Experimental phase

3.3 Phase III: Compilation phase

4 Significance

4.1 Descriptive value

- *The most detailed dictionary for (certain) loanwords.*
- *Database*
- *Semantic profiles*

4.2 Theoretical value

- *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?*

5 Timeplan

& one phase per year.

6 Ethical considerations????