A usage-based approach to borrowability

Backus, Albert

Publication date:
2012

Document Version
Peer reviewed version

TILBURG UNIVERSITY
Understanding Society

Tilburg Papers in Culture Studies

Paper 27

A usage-based approach to borrowability

by

Ad Backus

a.m.backus@uvt.nl

March 2012

A usage-based approach to borrowability

Ad Backus, Tilburg University, a.m.backus@uvt.nl

Abstract

Borrowability has been a topic in language contact research since the field began. It has been approached from various angles, and has led to borrowability hierarchies that rank parts of speech according to the ease with which they can be borrowed. Such hierarchies provide a starting point for explanatory efforts: why is it, for example, that nouns are eminently borrowable, and why is inflectional morphology rarely borrowed? Several methodological problems, however, plague the investigation of borrowability. One is the availability of sufficient data. Most hierarchies are based on reported summaries in the literature and relatively small corpora. Since funding agencies will not easily fund the building of large corpora of bilingual speech, it is important to develop additional methods. In fact, psycholinguistic experimentation would be a welcome addition to the field of contact linguistics, as it will allow investigating questions about borrowability that are only beginning to be asked. These questions are driven by the advent of the usage-based approach in linguistics, an approach that has not been applied much to contact data yet, but which is very compatible with how most theorists have accounted for language contact. The paper goes over some of these theoretical issues, and discusses the methodological implications. Most importantly, a usage-based approach to borrowability demands we collect data on loanwords' entrenchment in individual speakers and their conventionalization across speech communities. In doing this, the paper attempts to solidify the links between contact linguistics and cognitive linguistics, thereby contributing to 1) a better understanding of the phenomenon of borrowing; 2) the account of language contact phenomena in a Cognitive Sociolinguistics framework (more specifically a usage-based account of contact-induced change); and 3) a further appreciation of the methodological issues involved in researching borrowing from these perspectives.

Introduction

Borrowability has been a topic in language contact research since the field began. It is clear that languages borrow from each other, but it is much less clear what exactly they will borrow (and what not), or what determines the rate with which they do so. Knowledge about borrowing is important for understanding how and to what degree cultures influence each other, and to what degree languages just follow suit, i.e. whether or not languages are a direct reflection of culture. The degree to which *language* boundaries are permeable may or may not be independent of the degree to which *cultural* boundaries are permeable. Knowing more about this issue means knowing more about the essence of language.

For many languages, there are estimates about the percentages of their vocabularies that consist of borrowed words, but these only tell us so much. They generally provide a cumulative picture of the lexicon of the entire speech community, many of the words in question will not be in general use, or no longer so, and the unit of counting tends to be the lemma, not the actual word form. This means that estimating that the English word stock consists of 70% borrowed material doesn't mean at all that 70% of everyday language use or of the lexical competence of an individual English speaker is of foreign etymology. For that, we need a more direct picture of loanword usage in everyday discourse. This paper attempts to make the case that we don't actually have that picture, mainly because linguistic theory hasn't prompted the right questions for linguists to start looking for it. It will also argue that this situation has changed with the advent of the usage-based approach, and that now that the issue is on the agenda, some methodological hurdles need to be overcome.

The paper will go over the theoretical background to these issues, in an attempt to solidify the links between contact linguistics and cognitive linguistics, thereby hopefully contributing to 1) a better understanding of the phenomenon of borrowing; 2) an account of language contact phenomena in a Cognitive Sociolinguistics framework (more specifically a usage-based account of contact-

induced change); and 3) a further appreciation of the methodological issues involved in researching borrowing from these perspectives. Illustration will come from a Dutch Turkish spoken corpus, collected by the author and associates.

Studying loanwords

While there are many types of language change, this paper focuses on borrowing as one of the more common types. Other types include, for example, the loss of features, monolingual inter-idiolectal borrowing, or deliberate creation. We further simplify the base of discussion here to *lexical* borrowing, i.e. the adoption of loanwords.

Loanwords have been studied in *historical linguistics* and in *contact linguistics*, and this has led to the borrowability hierarchies that rank parts of speech according to the ease with which they can be borrowed (e.g. van Hout & Muysken 1994; Field 2002). Such hierarchies provide a good starting point for explanatory efforts: why is it, for example, that nouns are eminently borrowable, and why is inflectional morphology rarely borrowed? Various theories ask the question in some form or another, for example conceptualizing it as a question of attractiveness (Johanson 2002). The question is not just what is attractive, but especially what causes something to be attractive. Suggested explanations for why nouns seem the most borrowable include the distinction between open and closed classes (words from open classes are more easily borrowed), or between content and function words (content words are more easily borrowed), the degree of syntagmatic freedom (nouns are less tied structurally to other words in the sentence, and can therefore be borrowed more easily), or an underlying dimension of semantic specificity: the more specific the meaning of a word, the more attractive it is for other languages, as there is a good chance it would add to that language's expressive richness (Backus 2001). Nouns tend to have highly specific meanings. Note that this body of work tends to focus on words only; almost no attention is paid to multiword units or constructions. This is understandable given the traditional division between lexicon and syntax as

separate modules, but it also hinders progress, as it keeps the field from looking for commonalities with other kinds of borrowing.

The two research traditions, historical linguistics and contact linguistics, have much to offer each other, but in my estimate they do not always communicate very well. The situation has certainly improved much, though, since the publication of Thomason & Kaufman (1988), and the opportunities for a better theory of contact-induced change once information about historical changes is combined with observations of ongoing change in current contact settings are being actively explored now (cf. Mufwene 2008; Matras 2009). Examination of the lexical stock of languages provides much information about historical loanword layers (e.g. Latin and French words in English), and about the cultural scenarios that can be extrapolated (for example that Germanic people adopted many cultural artefacts from the Latin-speaking Romans). Studies of modern bilingual settings ostensibly show how loanwords come to be: bilingual speakers often codeswitch, and one prominent type of codeswitching is the insertion of foreign words into utterances otherwise framed in the base language. Such inserted words may well be future loanwords, or they might even be established loans already. One interesting difference between the empirical data these two fields make available is that historical loanword layers almost exclusively yield simplex words, while insertional codeswitching data include many other types of insertions besides simple words. There are many attested examples of inserted phrases and collocations, and these have increasingly become the focus of theoretical attention in codeswitching research (e.g. Muysken 2000; Myers-Scotton 2002; Backus 2003). Engaging with this paradox implies a shift away from the exclusive attention on simple words: what happens to these inserted chunks, phrases and expressions diachronically? Why does only a subset of insertions, i.e. simple words, end up as loanwords?

The study of *contact-induced change* could be seen as a third field, with links to both historical and contact linguistics. It shares with historical linguistics a focus on grammatical changes, and with contact linguistics an empirical focus on on-going contact settings. This field enjoys high vitality, and many excellent articles and

monographs have appeared in the last two decades (e.g. Aikhenvald 2002; Heine & Kuteva 2005; Verschik 2008). Generally, these studies make use of the tried and tested linguistic modes of descriptions, featuring a strict separation of lexical and structural issues (and, as mentioned, largely focusing on the latter). Having said that, though, together they have been building an impressive library of contact-induced grammatical changes in a growing range of languages and contact settings, allowing detailed hypotheses about what is typical and what is not in how languages influence each other.

Potentially, these research traditions could be combined into a more comprehensive theory of how languages lexically and structurally influence each other in the various stages of contact situations, from emergent bilingualism to the cessation of language contact (i.e. when one of the languages is no longer present, e.g. because of completed language shift). Probably, the reason why this is not done much has more to do with the sociology of science (different networks, different methods, different publication outlets, and different conferences) than with any principled incompatibility. Certainly, it seems that achieving this combination is not a widely felt need; part of the goal of this paper is to get it on the agenda. I will argue that the usage-based approach currently in the ascendancy in linguistics fuels this sense of urgency.

Solidifying links between contact linguistics and cognitive linguistics

The conjoined fields of contact linguistics, historical linguistics and sociolinguistics face challenges of a theoretical nature if they are to build a comprehensive theory of borrowing. This is all the more true, I will argue, if this is done in a usage-based framework (Barlow & Kemmer 2000 is a good introduction to this approach).

During the rise of usage-based linguistics in the previous twenty years or so, links with the concerns of sociolinguistics have repeatedly been mentioned. In a sense, it seems astounding that the fields have not embraced each other immediately, since a

usage-based approach to mental representation all but calls out for attention to differences between people in their language use, as studied by sociolinguists, while it can provide sociolinguistics with a model of the cognitive organization of language that is much more in line with its central concerns (variation and change) than the long-dominant generative approach was (cf. Kristiansen & Dirven 2008).

Language change has not featured prominently in recent linguistic theorizing. For the strictly synchronic linguistics of the past decades, the goal was to model the stable and invariant components of linguistic knowledge, usually hypothesized as innate knowledge, and then language change seems a relatively superficial concern. Change and variation were seen as interesting at best, or as relevant for the concerns of social science, but not for linguistics. The strict separation between lexicon and syntax has also kept up the apparent irrelevance of at least lexical change, including the adoption and diffusion of loanwords. However, usage-based approaches to linguistic competence do attribute direct theoretical importance to the social and psychological determinants of language use, and to fluctuations in the use of particular linguistic elements. Change is often a matter of 'merely' increasing or decreasing frequency of use, rather than the adoption or complete loss of particular forms.

As for loanwords, in a 'cognitive sociolinguistics' account, the use of foreign units (words, expressions, constructions, patterns) would be seen as raising their degrees of entrenchment in the mental representations of individual speakers, and cumulatively this may ultimately lead to levels that are so high that we can reasonably speak of 'change'. A logical correlate of this is that the disuse of a native equivalent leads to a lesser degree of entrenchment, and perhaps to its ultimate loss from memory. Or, more likely perhaps, we will see the waxing and waning of the entrenchment levels of different aspects of the unit's polysemous uses: particular aspects of its meaning get a boost from contact; others waste away. In this perspective, the explanation of contact-induced language change comes down to two things: explaining the *social* determinants of language use and the way our

cognitive system deals with this, both in terms of synchronic processing and of diachronic storage.

A usage-based approach logically entails that variation and change are essential design features of language. In fact, since it assumes that performance directly influences competence, and holds performance and usage to be largely synonymous, it provides the performance-based linguistic theory sociolinguistics has long called for. That entails, in turn, that a usage-based approach calls for the unification of sociolinguistics and general linguistics: if variation and change are central features of language, linguistic theory needs to account for them in an integrated theory of mental representation.

Combining the two research paradigms may truthfully be innovative in the explanation of language change. Traditionally, externally and internally induced types of change are distinguished, and theories about each tend to be developed separately. Externally induced, or contact-induced, change, is studied within contact linguistics, and is concerned with issues of taxonomic classification (for example distinguishing lexical and structural borrowing), mechanisms (codeswitching, direct borrowing, etc.), and the origins of the change in question (this is trivial for loanwords but not for grammatical interference: it is notoriously difficult to prove beyond doubt that a particular structural feature originated in, or was the result of influence from, the other language, cf. Thomason & Kaufman 1988). The characteristics of the social context that ultimately gave rise to the change tend to be viewed in relatively crude terms, emphasizing global aspects such as dominance relations. Internally induced change, on the other hand, tends to be studied through the methods of variational sociolinguistics, tracking the frequencies with which old and new variants of a particular variable, e.g. the pronunciation of a particular sound, are used by different sections of the population. This paradigm is mainly concerned with measuring the rate of change and any links the change may have with social factors, as these, again, provide a clue to the ultimate reason for the change. Here, there is relatively little interest in matters of the brain: how variation and change are made possible or are constrained by cognitive factors is not a topic

prominent on the sociolinguistic research agenda. Usage-based linguistics provides a way, an incentive even, for these traditions to merge into a unified study of language change.

There are two reasons for this. First, usage is influenced at more concrete levels than the broad-brush community-based factors commonly considered in sociolinguistics and contact linguistics, such as the relative dominance of the languages and the intensity of contact, and speaker-based factors such as age, gender or social class. Though these factors ultimately help explaining any individual's usage, there are still many basic-level factors that determine usage at a more subtle level, such as who one's friends are, what one's hobbies and interests are, and what job one has. While the macro-level factors determine one's *repertoire* in terms of the languages and varieties one masters, it is likely, at least, that the basic-level factors exert considerable influence on one's *inventory* of lexical and constructional forms, particularly on the degree to which they are entrenched in one's idiolect. Second, conceptualizing change as the increase or decrease of the degree of entrenchment of particular form-meaning units takes the concerns of variational sociolinguistics straight into the realm of mental representation. At the very least, incorporating a cognitive component into a social, or performance-based, account of language change provides a more comprehensive model, as the mind is, ultimately, the place where the change is located. Language is, after all, a mental phenomenon (otherwise, there wouldn't be any psycholinguistics or neurolinguistics).

In addition, it might be worth pointing out that since usage-based linguistics conceptualizes language as a set of form-meaning units, it imposes the same bottom-up procedure for describing languages that modern sociolinguistics advocates, particularly the strain that started with Hymes and Gumperz, and that is now alternately known as linguistic anthropology, interactional sociolinguistcs or discourse analysis (cf. Blommaert & Backus 2011). One question these fields broadly engage with is 'what is a language?' Space doesn't permit reviewing this field here, but one finding other fields may well adopt from this tradition is that languages are not always the bounded entities that people generally believe them to be. To be

sure, there are social settings that qualify as 'focused' in terms of Le Page & Tabouret-Keller (1989), and linguistic boundaries are tightly controlled in such settings. In those cases, it is a prominent part of the meaning of any linguistic element, especially words, to which language they belong, but there are other situations ('diffused' ones), in which attributing linguistic adherence plays a much less important role. Diffused settings are conducive to using words and structures from 'other' languages. The limiting case of this freedom may be switching between styles of the same language, or between idiolects.

Perhaps the best diagnostic for whether a speech community is focused or diffused is the degree to which a purist attitude is widely shared. Purism typically targets elements from a foreign language, for a variety of reasons. Foreign languages may stand for, or index, certain norms and values that are deemed alien or incompatible with the norms and values associated with the native language. In addition, foreign words stand out more, and are, therefore, an easier target for purism than, say, words associated with a different register of the same language, or with a different speaker of the same language. A usage-based approach to language change conceptualizes the origin of change as the adoption of a unit from someone else's speech; purism acts as a brake on this process, making it harder to adopt a unit that has as part of its meaning that it belongs to another language (cf. Hill & Hill 1986 for a telling illustration of the pragmatics of using loanwords, specifically of the effects of this kind of purism on people's linguistic awareness). Whether or not this part of its meaning is salient depends on the attitudes of the speaker, and these are informed by the level of purism present in the speech community he is a member of (cf. Aikhenvald 2002). In diffused communities, this part of the meaning is not very salient, and this stimulates building up an inventory that just consists of words from two or more different languages. Loanwords, then, stand out less in diffused communities. Obviously, this account only makes sense if one adopts a definition of 'meaning' that is encyclopedic, including anything from denotational semantics to individual pragmatic associations. Social indexicality is ultimately an aspect of meaning.

Loanwords in cognitive sociolinguistics: towards an account

Loanwords provide what may be the conceptually easiest type of contact-induced change. As their foreign origin is beyond doubt, there will rarely be discussions about the pre-contact presence of the word in the receiving language, an issue that makes suspected cases of contact-induced grammatical change often very hard to prove (Thomason & Kaufman 1988). The pre-contact entrenchment level of a loanword in the speech of individual bilinguals will have been zero. During contact, however, as the change they instantiate is being propagated, entrenchment levels fluctuate somewhere between low and high, depending on whether the individual uses it or not, whether people around him use it or not, and the extent to which it is used. This brings up a thorny methodological problem.

There are two levels at which the question how well a loanword is integrated in the speech community can be investigated, and they are not always kept properly apart. Most of the time, what is meant is *community-based conventionalization*. This is a sociolinguistic notion which refers to the degree to which the loanword has become a conventional lexical choice for the various members of the community. If all members use it, it is fully conventionalized as a normal word in the language. The other level is that of *person-based entrenchment*. This psycholinguistic notion deals with the degree to which a particular speaker knows the word. Theoretically, a loanword may be the conventional choice for one or a few people in the community, so that it is an established loanword for them and a highly entrenched part of their inventory, but never be used by others, so that we couldn't really see it as a conventionalized loanword in the variety spoken in the bilingual community.

A moment's reflection shows that none of this is unique for loanwords: the question how well individual entrenchment and community convention correlate holds for all lexemes, not just borrowed ones. What does seem specific for loanwords is the competitive relationship they may enter into with any native equivalents, various factors determining the choice for one or the other. However, this too applies within

the native lexical stock as well, since there are many near-synonyms in any language, the choice of which is conditioned by all kinds of social, contextual, semantic and personal factors.

Just like loanwords cannot be the only source of evidence for a theory of lexical variation and change, an account of loanwords alone is not enough for a theory of contact-induced change either: the innovation and propagation of loanwords needs to be placed within a larger theory of contact-induced change that also takes into account loan translation, semantic extension, and all kinds of grammatical change (Croft 2000; Backus 2005). Perhaps the trickiest theoretical and methodological issue facing this field is how to handle the *Transition Problem*, identified by Weinreich, Labov & Herzog (1968) as one of five issues any theory of language change needs to tackle. If we conceptualize borrowing as a case of lexical language change, how on earth do we know whether a foreign word we see used in a particular language represents an established change in that language, an ongoing change, or only an incipient change that we managed to catch in its early stages? For example, when an individual Turkish-Dutch speaker in The Netherlands uses a particular Dutch word in his Turkish, we do not know to what degree that word is an established loanword in that person's Turkish, let alone in Immigrant Turkish in general. This occurrence will normally be analyzed as a case of codeswitching, but that says nothing about the degree of conventionalization. Borrowing is a diachronic process while codeswitching is a synchronic event. The Dutch word can thus be both: synchronically a codeswitch to Dutch, and diachronically a more or less established loanword in this particular variety of Turkish. To assess its status as a loanword, we would need information on its degree of entrenchment in the idiolect of the speaker, and its degree of conventionality in the speech community of which the speaker is a member. Its ability to be used (and perceived) as a switch to Dutch is relatively independent of this, as long as all Turkish speakers are bilingual and can potentially recognize any Dutch-origin element. What is meant by this is that codeswitching is taken in its literal sense, as a switch to Dutch. This can be done for any number of pragmatic reasons, such as attention grabbing, emphasizing, etc. However, it stands to reason that this potential decreases with increasing entrenchment, since the

effect of this entrenchment is to make the word in question a normal Turkish word. The more entrenched, the less its Dutch-origin nature stands out. As long as the population is bilingual, though, this potential can never be zero.

The extensive literature on bilingual speech makes it clear that there is considerable variation across speakers in codeswitching patterns. From a usage-based perspective, this means there must also be considerable variation in speakers' mental representations, including in the degree to which particular foreign-origin words are entrenched. In this perspective, linguistic competence depends on culture: the features of someone's social life determine what kinds of linguistic features she will use and be exposed to, and hence what will be entrenched to what degree (or, alternatively put: how proficient she will be in the various registers that play a role in her life).

From the perspective of cognitive sociolinguistics, then, it is not so much borrowability hierarchies that are interesting, but rather what borrowing can tell us about the nature of language change. The *cognitive* interest centers on issues of entrenchment and lexical semantics: how entrenched is a putative loanword (and therefore, to what degree can we say that the language has undergone change), and why was it borrowed in the first place (addressing an underlying semasiological dimension). In terms of Weinreich, Labov & Herzog (1968), the first question addresses the *Transition Problem*, and the second provides a piece of the *Actuation* puzzle. The *social* interest of the issue lies in the tension between the individual nature of entrenchment and the social nature of conventionalization. If the loanword is entrenched to different degrees by different speakers, then for whom is it entrenched more, and why? These questions are also part of the usage-based reformulation of the *Transition Problem*.

Perhaps a final word is in order here about the difference between borrowing and codeswitching. In the codeswitching literature, this has proved to be a very divisive issue, evaluations of the value of a theoretical proposal sometimes hinging on the question whether a particular counterexample should be classified as a codeswitch

or as a case of borrowing. To my mind, this debate is misguided, because a foreign-origin word can be both: borrowing and codeswitching are not directly comparable like that. The synchronic use of the word in question in a particular sentence recorded for the corpus cannot tell you much about the degree to which the word is integrated into the receiving system. To assess status as a loanword, we need to obtain information on its degree of entrenchment in the idiolects of speakers, and from that extrapolate its degree of conventionality in the speech community of which these speakers are members. The two categories are not mutually exclusive. A loanword is a foreign-origin word which is, to a certain extent, an accepted and established lexical item in the borrowing language. A codeswitch is a shift in mid-utterance or mid-discourse to material from the other language. In a bilingual context, these two categories do not exclude each other. What is needed for a word to be a loanword is that it is used often enough. For something to be used as a codeswitch, what is needed is some awareness of the foreign etymological origin. It is easy to see that in a bilingual situation, both conditions can apply to the same word at the same time.

This section has discussed what Cognitive Linguistics and sociolinguistics share, and what they have to offer each other. I have argued that the usage-based approach that underlies much of Cognitive Linguistics is compatible with the concerns of sociolinguistics, and that the study of contact phenomena, including the innovation and propagation of loanwords, is a suitable domain for exploring this link between two subfields. More in general, the lack of a rigorous distinction between lexicon and syntax in Cognitive Linguistics can help bring the studies of lexical contact phenomena (codeswitching, loanwords) and structural ones (contact-induced change) closer together, a prerequisite for a more general account of language change.

Methodological hurdles

Several methodological challenges plague the investigation of borrowability within the realms of traditional contact linguistics and sociolinguistics. One is the availability

of sufficient data. Take the basic question how pervasive loanwords are in current language use, in whatever modern language. Of course, one can search dictionaries to see how much of the vocabulary originated in another language. Loanword dictionaries, in fact, exist for many of the major languages, and they give a fine perspective on past contact situations and the degree to which the language has participated in the global flow of cultural influences. However, they are less useful for research questions that deal with synchronic language use. Which of those loanwords are, for instance, really in current use? And how frequently are they used? Is their frequency of use purely determined by the number of times the concept they encode is needed, or are they (still) in competition with a native equivalent?

And if the contact situation that gives rise to the borrowing is still ongoing, another set of questions remains hard to deal with for lack of relevant data. Who uses these loanwords and who doesn't? To what degree do they compete with native equivalents? To what degree is their usage dependent on communicative, contextual and stylistic factors? Such questions can perhaps all be subsumed under the general issue of the degree to which such putative loanwords are established in the borrowing language. Is there a direct link between loanword usage and the use of foreign-origin grammatical features? Is there a trade-off between using loanwords and employing loan translations?

Perhaps the above remark about the scarcity of data is a tad too pessimistic. After all, for many of the major languages, large spoken corpora are available and corpora of written data, such as newspaper archives, are relatively easy to come by. Those interested in the spread of loanwords may mine these monolingual corpora, but we should bear in mind that this will only provide information about one type of loanword. Loanwords enter languages through face-to-face contact between bilinguals or through the intermediary of elite bilinguals. The latter type is not unimportant, and is probably responsible for many of the Latin and Greek internationalisms in most of the world's modern languages. Its latest incarnation is the globalization-induced spread of English words worldwide through the media:

extensive knowledge of English or daily face-to-face contact between bilinguals is not necessary for English words to spread successfully around the globe. The tools of corpus linguistics can most certainly be used to investigate the spread of this type of loanword. The Corpus of Spoken Dutch (CGN), for example, a 10-million word sample representative of spoken registers in Holland and Flanders, will contain many English words, and identifying the frequency and contexts of their use, and of the speakers who use them, will go some way towards answering some of the abovementioned questions. Analyzing English words used in the CGN will certainly show the extent to which globalization affects the Dutch lexicon. On the other hand, even spoken corpora tend to be relatively limited in the amount of everyday informal interaction they can include, so their usefulness should not be overstated. They tend to make liberal use of data that are easier to process, such as public lectures. Note, also, that large corpora are normally not tagged for etymological origin of the words that are used, so that identifying these words will be a lot of work. More generally, though, these corpora will tell us little about how borrowing works in face-to-face contact between bilinguals.

However, it is not so easy to improve the availability of data. Borrowing tends to be from the dominant language in society into a dominated language, often the language of an immigrant or indigenous minority group. Immigrants are prone to shift to the majority language at some point; this makes it unlikely that any funding agency will spend large sums of money on building a large corpus of the minority language. The situation is better for indigenous minority languages, especially if they are threatened with shift and death, since conservation and documentation of the language may be perceived as a matter of national interest, of preserving an essential heritage. On the other hand, the corpus that might result is unlikely to accurately reflect loanword usage, since such languages will often be in the grips of purism. Language documentation will often be designed to maximize monolingual language use. Overall, funding agencies are not likely to stimulate the building of corpora of sufficient size for analyzing bilingual speech the way they do for the world's major languages. On the other hand, social developments in bilingual life (use of Internet-based modes of communication) and technological developments in

'E-Humanities' (e.g. new extraction techniques) may make more tools available than can currently be envisaged. On the whole, as I'm sure this paper also illustrates, there is relatively little corpus linguistic expertise among contact linguists.

The Dutch speakers recorded for the CGN who use some English words most likely know English fairly well, but they are not bilingual in the sense that they use both languages interchangeably in the same everyday settings, displaying codeswitching and intricate patterns of language choice. Much borrowing, however, takes place in spontaneous bilingual speech in everyday settings, and it's this situation that underlies much of the contact linguistic work on borrowing (e.g. Matras 2009). The tools of corpus linguistics cannot easily be applied to it, since no large-scale corpora are available, and presumably never will be.

The Turkish corpus collected by the author and associates is typical. It is as large as any bilingual corpus one is going to find, consisting of about half a million words of spoken Turkish conversation. About two thirds of it was collected from bilingual speakers in The Netherlands, from both first and second generation speakers. The rest of the corpus served as control data, and were collected in Turkey, in the same place as where most of the immigrants in the bilingual corpus had their roots (the central Anatolian town of Kırşehir). All data come from spoken everyday interaction: most were interviews with one to three individuals, conducted by an interviewer unknown to them before the recording. The recordings have yielded a stylistically fairly homogeneous set of data, which has been stored in a machine-readable form. This is a fairly typical corpus for contact linguistics; similar databases have been built elsewhere, including other ones for Immigrant Turkish. The gold standard is perhaps provided by the corpora built under the supervision of Shana Poplack in Ottawa (see www.sociolinguistics.uottawa.ca).

Tracing the diffusion of individual Dutch loanwords in Immigrant Turkish is impossible with these data. Basically, it's a problem of numbers: the corpus is simply too small. Loanwords tend to be content words, and even frequent content words do not occur that often in a corpus of half a million words. Given that loanwords

tend to have relatively specific meaning (Backus 2001), the typical loanword will have a low token frequency. In addition, lexical diffusion tends to be determined by social factors, such as social background of the speaker, and communicative goals, but the corpus was kept as homogeneous as possible in order to be able to compare bilingual and monolingual Turkish. And this typifies corpora of this kind, as they want to maximize the number of comparable utterances, rather than document the extent of variation. That is, there is little stylistic variation between recordings and little social variation between speakers.

To assess how widespread a particular loanword is, we essentially need a measure of the degree of conventionalization of that word. As always, this requires two different types of measure: the social measure of how many people use it, and the individual measure of how well entrenched it is in the linguistic competence of representative individual speakers. As we have seen, for the vast majority of bilingual settings, there are no large and balanced corpora, so there are no frequency data that provide a reliable picture of how widespread a loanword is. There are various problems if they are to be used to investigate the question of loanword diffusion.

First, the speakers captured on tape are few, and therefore may not be representative of the community. Informants for codeswitching studies will often have been selected precisely because they codeswitch a lot, which is all fine and good if the structure and pragmatics of codeswitching is the object of research, but it is clear that these speakers only cover part of the range of sociolinguistic variation present in the community. Loanwords used by them may not be used by everybody. Second, the conversation captured on tape may not be representative of community interaction either. Often, the corpus consists of only a few, or even just one, recording. It is, therefore, unlikely to capture the full communicative repertoire of the community.

The only methodological step that may possibly be defendable in using these data is some form of extrapolation. It sure stands to reason that if the use of a loanword is captured in such limited data, it probably is a word that is in general use in the

community. This can then be checked, in at least two ways. One solution would then be to search for more data concerning this particular word, e.g. by browsing Internet forums and blogs using the community language; the other one, especially advocated here, is to use these words as stimulus items in judgment tasks or as the basis of discussion in focus group interviews. Essentially, the question posed to informants then becomes something like 'I found you guys using this loanword in everyday conversation; how widespread is it really? Do you and/or people around you indeed use it freely?' There are, thus, reasons to invest in alternative methods for investigating the social diffusion of loanwords beyond the difficulty of building suitable corpora.

To summarize, while tracking loanword diffusion would tell us something about the rate of change, it is difficult to accomplish such tracking. The small corpora of bilingual speech collect at most a few hours speech of a limited number of speakers. Often, this won't turn up even a single instance of particular foreign-origin words that may well be established loanwords in the vernacular of the community. The situation is different for studying the spread of borrowed phonological or syntactic features: with due reservations, such corpora can be used to investigate their diffusion, as their token frequency will at least be quite high. Sociolinguistic variation analyses, of course, often rely on the quantitative analysis of just this kind of data. It is possible, for example, to track the use of a particular AAVE feature, such as copula *be*, in a large corpus of American English and check to what extent it has penetrated general usage. Similarly, even with a modest corpus of Immigrant Turkish, it is possible to track the occurrence of 'native' SOV and 'borrowed' SVO order (cf. Doğruöz & Backus 2007). But it is impossible to track the diffusion of a Dutch word this way, so for loanwords at least, we need alternative methods (see below).

To answer the types of questions a usage-based approach generates concerning loanwords, it is clear that data other than corpora are needed to get full answers. On the *cognitive* side, differential entrenchment levels of individual loanwords can be shown through psycholinguistic measurements, e.g. judgment tasks. Corpus frequencies, if a sizable corpus is available, can certainly be used as an additional

source, perhaps providing converging evidence, but for reasons outlined above (low or zero token frequency of individual content words), they are unlikely to provide us with very useful data by themselves. Corpus linguistics makes several tools available that have not been explored at all yet in connection to loanwords, as far as I know. A collostructional analysis (Gries and Stefanowitsch 2004) of loanwords could, for instance, show that speakers prefer to use loanwords in particular parts of a clause, such as the periphery (Treffers-Daller 1994) or special slots for loanwords (Poplack & Meechan 1995). One interesting question would be whether foreign words are implicated in the spread of foreign structure. In a collostructional analysis, the collostructional strength of foreign words and a particular foreign structure could be checked: if there is a significant attraction, then using foreign words appears to push the entrenchment of the foreign construction. That would suggest evidence that *lexical* codeswitching is a mechanism for contact-induced *grammatical* change, a hypothesis sometimes hinted at, but so far not empirically demonstrated.

On the *social* side, the social meaning of individual loanwords should equally be uncovered through some kind of attitude measurements, for example in acceptability tasks, or perhaps through focus groups. Again, the conversational analysis of occurrences in corpora can provide valuable additional, hopefully converging, data, but as the sole method it would rely too much on chance encounters ('found data').

Challenges

Following the usage-based approach, I defined change as the increase or decrease of the degree of *entrenchment* of a linguistic unit. Fine as this may be in the abstract, several potentially problematic questions are raised by this definition, and they bring further methodological challenges with them. Entrenchment of what exactly, for instance? And what kind of evidence is needed before we are able to say that a change is propagating at the community level, i.e. in how many individuals do we need to show fluctuations in entrenchment levels?

The discussion here is concerned with loanwords, but most usage-based approaches will hypothesize that the mechanisms are the same for schematic units (i.e. grammatical patterns) and partially schematic units (i.e. constructions in the sense of Construction Grammar), cf. Langacker (2008). On the basis of the discussion above, it would seem the methodology is fairly straightforward: you single out the unit to be investigated, you count how often it occurs in a corpus or, better, you measure subjects' responses to it in some suitable task testing cognitive accessibility to the form or evaluative judgment about it. However, so far we haven't problematized the term 'form-meaning unit', and maybe we should. On the form side, there is not much of a problem; at most we have to decide about whether to look at types or tokens, or at lemmas or word forms. However, forms tend to be polysemous, and hence we have to ask: what meaning do we look at? Should all meanings be taken together, so that each occurrence, no matter what the specific contextually determined meaning is, contributes to the entrenchment of one form-meaning unit? There doesn't seem to be an easy answer to this question.

What seems to make sense, though, is to assume that in case of true polysemy (rather than, say, homonymy), all uses count. If a Turkish speaker in Holland uses the Dutch word *feestje* 'party' several times, it may alternately refer to different kinds of parties, but by and large it all contributes to the entrenchment of the unit that comprises this form and a generalized meaning of 'party', glossing over the differences within a range of types of party. It may or may not overlap with the meaning of a small number of Turkish equivalents, such as *eğlence* and *parti* (most likely, the Dutch word will for most people take on the specific connotation of a party done the Dutch way, thus making the meaning that is being entrenched relatively specific).

Specificity helps the putative loanword in its competition with any native equivalent, as the specific meaning may make it more salient, or suitable, in many of the contexts where in principle both words would suffice. Encyclopedic characterization of meaning is key: a foreign word's attractiveness may lay solely in its pragmatic

impact or in the fact that the language it originates from has an association with cultural change (modernization, globalization, etc.). There are plenty of data in the codeswitching literature that suggests this, whenever examples are presented of insertions with highly specific cultural meanings or where switching is done to achieve a reference to a more powerful code. However, such data tell us little about the degree to which those loanwords are commonly seen as part of the lexical inventory in the receiving language.

The pre-contact situation is obvious: the entrenchment of the Dutch word is zero. But what is the pre-contact level of entrenchment of the Turkish words? Should we set them at 100%? That would only be justified if entrenchment can reach levels where further activation doesn't really do anything anymore. In reality, it is certainly imaginable that monolingual Turkish speakers, given the right methodology, will be shown to have different levels of entrenchment for *eğlence* and *parti*. In the contact situation, the entrenchment of each of the three words will be more than zero for most bilinguals; but how high they should be set seems to be an empirical question. Has the Dutch word reached the same levels as its Turkish counterparts? Have one or both of these decreased their entrenchment levels? Are the figures for the words related, so that for any individual speaker the entrenchment of one word predicts the entrenchment of the others? That would be a useful hypothesis, since the words may be expected to be in competition. Of course, without data from experiments and tasks that measure entrenchment in a suitable number of informants, this is just a theoretical game. What is urgently needed for a usage-based study of loanwords (and, by extension, of other contact-induced changes) is actual data on entrenchment levels (e.g. through judgment tasks), to provide an empirical basis for investigating the spread of loanwords, and their degree of integration into the repertoire of the speech community. If frequency data are available (but see above for the reason why I'm pessimistic about that), empirical testing of the usage-based assumption of a correlation between frequency of use and entrenchment also becomes possible. As far as I know, neither type of data is available at the moment for putative loanwords in bilingual situations.

One of the more urgent tasks for contact linguistics in the immediate future, I would think, is to develop the methodology for obtaining these kinds of data. We would then be able to come up with lists of successful Dutch loanwords, to be contrasted with less successful ones and words that never made it, and with Turkish words that remain well entrenched in the competence of Turkish speakers and ones that disappeared, or have weakened in entrenchment. That in turn would provide better empirical footing for that fundamental question asked repeatedly in contact linguistics and in historical linguistics: what explains borrowability hierarchies? We would be able to go beyond the usual explanations in terms of Parts of Speech, which are useful, but limited in scope.

Examples that can provide inspiration are readily at hand in other fields that are interested in exactly these same concepts: psycholinguistics, cognitive linguistics, and, especially, the nexus between these two fields. These fields have exploded in recent years with empirical investigations into entrenchment, mostly making use of variations on the conventional judgment task, such as Magnitude Estimation, lexical decision and speeded grammaticality judgment tasks (Schönefeld 2012). Such tasks can also be used to track the degree to which individual loanwords are deemed to be in common use in bilingual populations. Arguably, they provide better data on this issue than corpus data would, even if we did have a large corpus at our disposal. Ideally, both types of data provide converging evidence. Generally, studies in Cognitive Linguistics find good results when attempting to correlate corpus frequencies and behavioral or psycholinguistic measures. Elements that are frequent elicit shorter reaction times, for example, in lexical decision experiments. Obviously, for contact varieties we won't have as good a basis for frequency data as for the larger world languages, but experimental measurements are surely within reach.

Another challenge is to figure out what happens to lexical chunks larger than single words, which appear in great numbers as complex insertions in codeswitching data, but fail to make the cut in lists of loanwords. It is unlikely that this is simply an oversight on the part of the loanword list compilers. Once a contact setting sorts itself out, if that ever happens, the vocabulary of the borrowing language is enriched

with a bunch of loanwords, but borrowed phrases and collocations are few and far between. It is because of their rarity that the occasional French phrase (such as *je ne sais quoi*, or *le mot juste*) borrowed into English becomes a contact linguistic *cause célèbre*.

In the past, this question was not asked because codeswitching studies showed little interest in how the use of foreign-origin elements develops over a longer time period, i.e. in the diachronic development of borrowing patterns, while for the field of contact-induced change these cases are too lexical to be of more than passing interest. For the type of usage-based approach sketched here, the question is more interesting. If Turkish speakers in Holland routinely sprinkle their Turkish with Dutch phrases, as they have been observed to do, one would expect these phrases to become more and more entrenched in the competence of these speakers. The phrases in question are often adjective-noun (e.g. *short cycle*) or verb-object collocations (e.g. *run a program*), fixed and idiomatic prepositional phrases (e.g. *for what it's worth*) and assorted semi-idiomatic turns of phrase (e.g. *doesn't matter*). Would a Dutch Turkish develop in which these collocations and idioms become established loans? That is possible, but it is at odds with what we normally see in loanword layers. Various explanations are possible. It could be that most languages that incorporate this much foreign material eventually die, as their speakers simply shift completely to the other language. A usage-based hypothesis for this scenario would be that the foreign phrases tend to trigger more foreign material, such as subject and object pronouns, verb inflection, plural marking, etc, because of the strongly entrenched links to that other material. That means utterances will tend to become monolingual productions in the other language, and ultimately this leads to shift unless it is halted some way. Another explanation is that speakers at some point start to feel the need to halt this process, for example to protect the integrity of their 'native' language. Our data suggest that if speakers are forced somehow (e.g. by the choice of interlocutor) to speak monolingual Turkish, the incidence of loan translations and other forms of Dutch-influenced Turkish goes up. Phrases that could end up as multiword borrowings might instead end up as loan translations. This raises the interesting question whether the entrenchment of a foreign collocation is

transferable, as it were, to that of a literally translated native equivalent that didn't exist before contact.

Conclusions

This paper is an attempt to rethink the issue of loanwords from the perspective of an emerging Cognitive Sociolinguistics, and has worked out various theoretical and methodological implications. One is to rely less exclusively on corpus data and make better use of speaker's intuitions and metalinguistic knowledge. If we want to know about the degree to which a particular foreign-origin word has spread through the speech community, we can ask people. Loanwords are normally content words, and content words are normally low in frequency, so that corpus frequencies do not give a reliable picture about the overall use of these words in the speech community. This problem with corpora is exacerbated for bilingual speech because corpora will generally be relatively small in size. While the paper has focused on lexical cross-linguistic influence, the investigation of structural influence would also be better served by a combination of corpus and experimental data. While traditional approaches to linguistics, with their strict separation of lexicon and grammar, naturally focused on either one or the other, usage-based approaches call for a more integrated account.

References

Aikhenvald, Alexandra Y. (2002). *Language Contact in Amazonia.* Oxford: Oxford University Press.

Backus, Ad (2001). The role of semantic specificity in insertional codeswitching: evidence from Dutch-Turkish. In Rodolfo Jacobson (ed.). *Codeswitching Worldwide II,* 125–54. Berlin and New York: Mouton de Gruyter.

Backus, Ad (2003). Units in codeswitching: evidence for multimorphemic elements in the lexicon. *Linguistics*, 41(1), 83-132.

Backus, Ad (2005). Codeswitching and language change: One thing leads to another? *International Journal of Bilingualism* 9 (3/4), 307-40.

Barlow, Michael, and Suzanne Kemmer (2000). *Usage-based models of language.* Stanford: CSLI Publications.

Blommaert, Jan & Ad Backus (2011). Repertoires revisited: 'Knowing language' in superdiversity. *Working Papers in Urban Language and Literacies*, paper 67

Croft, William (2000). *Explaining Language Change: An Evolutionary Approach.* Longman, Harlow.

Doğruöz, Ayşe Seza and Ad Backus (2007). Postverbal elements in Immigrant Turkish: Evidence of change? *International Journal of Bilingualism*, 11(2), 185-220.

Field, Fredric (2002). *Linguistic Borrowing in Bilingual Contexts.* Amsterdam: John Benjamins.

Gries, Stefan Th. & Anatol Stefanowitsch (2004). Extending collostructional analysis; A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9, 97–129.

Heine, Bernd, and Tania Kuteva (2005). *Language Contact and Grammatical Change.* Cambridge: Cambridge University Press.

Hill, Jane & Kenneth Hill (1986). *Speaking Mexicano. Dynamics of Syncretic Language in Central Mexico.* Tucson: University of Arizona Press.

Johanson, Lars (2002). *Structural Factors in Turkic Language Contacts.* London: Curzon.

Kristiansen, Gitta & Dirven, Rene (2008). *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems.* Berlin: Mouton de Gruyter

Langacker, Ronald W. (2008). *Cognitive Grammar. A basic introduction.* Oxford: Oxford University Press.

Le Page, Robert & Andree Tabouret-Keller (1985). *Acts of identity. Creole-Based Approaches to Language and Ethnicity.* Cambridge, Cambridge University Press.

Matras, Yaron (2009). *Language Contact.* Cambridge: Cambridge University Press.

Mufwene, Salikoko (2008). *Language evolution. Contact, competition and change.* London: Continuum.

Muysken, Pieter (2000). *Bilingual speech: A typology of codemixing.* Cambridge: Cambridge University Press.

Myers-Scotton, Carol (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes.* New York: Oxford University Press.

Poplack, Shana & Marjorie Meechan (1995). Patterns of language mixture: nominal structure in Wolof-French and Fongbe-French bilingual discourse. In: P. Muysken & L. Milroy (eds.), *One speaker, two languages*, 199-232. Cambridge: Cambridge University Press.

Schönefeld, Doris (2012). *Converging evidence. Methodological and theoretical issues for linguistic research.* Amsterdam/Philadelphia: John Benjamins.

Thomason, Sarah Grey & Terrence Kaufman (1988). *Language Contact, Creolization, and Genetic Linguistics.* University of California Press, Berkeley, Los Angeles, London.

Treffers-Daller, Jeanine (1994). *Mixing two languages: French-Dutch contact in a comparative perspective.* Berlin: Mouton de Gruyter.

Van Hout, Roeland & Pieter Muysken (1994). Modeling lexical borrowing. *Language Variation and Change* 6, 39–62.

Verschik, Anna (2008). *Emerging Blingual Speech: from Monolingualism to Code-copying.* London: Continuum.

Weinreich, Uriel, William Labov & Marvin Herzog (1968). Empirical foundations for a theory of language change. In Lehmann, W.P., Malkiel, Y. (Eds.), *Directions for Historical Linguistics: A Symposium*, 95–195. University of Texas, Austin.