# Loanwords in the world's languages: A comparative handbook (review)

Yaron Matras

Language, Volume 88, Number 3, September 2012, pp. 647-652 (Review)

Published by Linguistic Society of America
*DOI: 10.1353/lan.2012.0052*

## REFERENCES

ARCHANGELI, DIANA. 1988. Aspects of underspecification theory. *Phonology* 5.183–208.
BLAHO, SYLVIA. 2008. *The syntax of phonology: A radically substance-free approach*. Tromsø: University of Tromsø dissertation.
CHOI, JOHN. 1992. *Phonetic underspecification and target-interpolation: An acoustic study of Marshallese vowel allophony*. Los Angeles: University of California, Los Angeles dissertation.
COHN, ABIGAIL. 1993. Nasalization in English: Phonology or phonetics. *Phonology* 10.43–81.
MORÉN, BRUCE. 2003. The parallel structures model of feature geometry. *Working Papers of the Cornell Phonetics Laboratory* 15.194–270.
ODDEN, DAVID. 1992. Simplicity of underlying representation as motivation for underspecification. *OSU Working Papers in Linguistics* 41.85–100.
VANCE, TIMOTHY. 1987. *An introduction to Japanese phonology*. Albany: State University of New York Press.

Department of Linguistics
Ohio State University
Columbus, OH 43210
[odden@ling.osu.edu]

**Loanwords in the world's languages:** A comparative handbook. Ed. by MARTIN HASPELMATH and URI TADMOR. Berlin: Mouton de Gruyter, 2009. Pp. 1,081. ISBN 9783110218435. $280 (Hb).

Reviewed by YARON MATRAS, *University of Manchester*

Lay people's first and most instinctive association with language and linguistic diversity relates to vocabulary. Words are considered the quintessential feature of a language more so than 'accent' or grammar. A group of experts in linguistic typology setting out to produce a volume that might trigger some interest among a lay audience would therefore be best advised to focus on the lexicon. The question 'which of the words we use are more likely to have been taken over from another language?' is a winner. In the age of globalization, everyone is aware of loanwords, shared words, and universal words. Even purists have no choice but to acknowledge the presence of lexical borrowings.

At the same time, studying lexical borrowings is by no means a mere popularization of linguistic typology. The very foundations of the comparative method in linguistics rely on recognizing cognates between languages and so on the notion of the diachronic stability of at least some parts of the lexicon. The realization that some meanings prove more reliable for this exercise than others has brought forward the idea of a 'basic lexicon'. Some linguists rely heavily on this idea in making the case for historical-genetic relationships among languages. Yet the precise definition of 'basic lexicon' has always been somewhat impressionistic. The Swadesh (1950) list remains to this day the principal tool and standard measure of lexical and so of language-genetic relatedness among languages. As work with the list became more ambitious and scholars ventured into the field of 'lexico-statistics', claiming to be able to reconstruct the time depth of separation between related languages on the basis of their shared lexicon, the issue of lexical borrowing took on a center-stage position as a potentially disrupting factor (e.g. Swadesh 1952, Embleton 1986, Renfrew et al. 2000). Needless to say, doubts about the reliability of the Swadesh list as an effective tool remain, and they continue to prompt those specializing in this field to re-examine the tool (see e.g. Starostin 2010). In this respect, the volume under review is a major empirical and theoretical breakthrough. For the very first time it approaches the issue of lexical borrowing in basic vocabulary systematically across a sample of languages and comes up with a new and empirically tested list of stable lexical meanings. This has potential implications for future research into language classification. It also has far-reaching implications for a theory of language contact and bilingualism.

The volume presents the outcome of the Loanword Typology project, led by the two editors, Martin Haspelmath and Uri Tadmor, at the Max Planck Institute (MPI) for Evolutionary Anthropology in Leipzig (and its affiliated field station in Jakarta). The contributors responded to an open call to join the project and were supported by MPI's technical team, who constructed a database to store the data, carry out calculations across the sample, and ultimately make the data available online (see below). The book follows in the footsteps of a collection that focuses on grammatical borrowing in a crosslinguistic sample of a similar size and similar diversity, also based on a questionnaire database (Matras & Sakel 2007). The two projects emerged at roughly the same time and partly in a shared discussion context, and there is even some overlap in the presence of languages and contributors (Michael Rießler on Kildin Saami, Viktor Elšík on Selice Romani, Zarina Estrada Fernández on Yaqui, Ewald Hekking and Dik Bakker on Otomi, Patience Epps on Hup, Jorge Gómez Rendón on Quichua, Uri Tadmor on Indonesian, Kristine Hildebrandt on Manange).

The collection has two parts. In Part 1, the two editors introduce the project methodology and database (1–34). MARTIN HASPELMATH then deals with concepts and issues in lexical borrowing (35–54), followed by a chapter by URI TADMOR on findings and results (55–75). Part 2 contains forty-one chapters on the borrowing behavior of individual languages. These chapters tend to follow a uniform format. They begin with an overview of the language community's history of cultural and linguistic contacts, often accompanied by a map. A brief description of integration patterns of loanwords follows. An analysis is then presented, relying on tables, of the number and distribution of loanwords by donor language and semantic domain. A short section is usually devoted to grammatical borrowing. Each chapter concludes with an appendix containing all loanwords recorded for the meanings contained in the project questionnaire of 1,460 lexical items, listed by donor language. Only some authors arrange the tables in order of hierarchical prominence (e.g. semantic domains in order of the percentage of loans, donor languages in order of their relevance), making a quick visual comparison based on a glance at the tables somewhat less convenient. The chapter appendices are entirely language-oriented in that they follow individual donor languages rather than semantic domains or individual word meaning (i.e. a preset order of words). But the reader who is motivated to learn more and rearrange lists over and over again following different criteria can do so by accessing the database directly on the project's website (see below).

Sampling languages for the purpose of contact linguistic studies poses some special challenges, not least because borrowing is sensitive to the power relations and the social functions of the languages involved (see Matras & Sakel 2007, and also Matras 2009). The loanword typology (LWT) language sample represents geographical, genetic, typological, and sociolinguistic diversity. The full list of languages included in the sample is: Swahili, Iraqw, Gawwada, Hausa, Kanuri, Tariffiyt, Seychelles Creole, Romanian, Romani, Lower Sorbian, Old High German, Dutch, English, Kildin Saami, Bezhta, Archi, Manange, Ket, Yakut, Oroqen, Japanese, Mandarin Chinese, Thai, Vietnamese, White Hming, Ceq Wong, Indonesian, Malagasy, Takia, Hawaiian, Gurindji, Yaqui, Tzotzil, Qeqchi, Oromi, Saramaccan, Quechua, Kalina, Hup, Wichi, and Mapudungun. Their diverse individual histories of contact can be read both from the tables of borrowings and from the surveys. We thus find languages like Swahili (THILO C. SCHADEBERG), with a broad range of historical contacts and loanwords derived from Arabic as well as items shared by Arabic and/or Hindi and Persian, and loans from English, Hindi, Portuguese, and Malagasy; or Romanian (KIM SCHULTE), which shows loanwords from Albanian, South Slavic, Bulgarian, Serbian, Ukrainian, Greek, Hungarian, Turkish, French, Latin, and Italian; alongside languages like Iraqw (MAARTEN MOUS and MARTHA QORRO), whose borrowings are derived almost entirely from Swahili; Bezhta (a Nakh Daghestanian language of the North Caucasus; BERNARD COMRIE and MADZHID KHALILOV), which has loans from Avar and Georgian; Old High German (ROLAND SCHUMANN), whose loans are mainly from Greek and Latin; and Otomi (EWALD HEKKING and DIK BAKKER), which borrows only from Spanish (though two borrowed items are from Nahuatl).

The material for the volume and for each language description is based on the project's questionnaire, the loanword typology meaning list, with 1,460 entries. The contributors took on the

task of identifying meaning equivalents in each of their languages, resulting in language-specific lists of between 1,000 and 2,000 items (not every language has exact meaning equivalents for every entry in the base list, and some have several equivalents for one base list meaning). The meaning list includes 1,310 items taken over from the list used by the Intercontinental Dictionary Series, to which 150 meanings were added, supplementing concepts of importance to geographical regions, or of significance to modern life (e.g. 'radio', 'hospital'), as well as items that appear on Swadesh's 207-item list. Contributors entered language-specific translations or counterparts for each meaning in the list into a database. Counterpart forms were recorded in transcription. A range of metadata information accompanies each entry: quantitative scales are used to indicate the degree to which a word is analyzable, the degree of certainty that the word is a loanword, the degree of its structural integration into the language, and its relative frequency (answering to the assumption that lexical stability increases with frequency). Information is included on etymology and the word's presumed or attested 'age' (the period in which it entered the language). The list is divided into twenty-four semantic fields, which include kinship, animals, motion, time, speech and language, law, religion and belief, modern world, the house, food and drink, the body, and more. This allows the user accessing the data to explore the rate of borrowing and retention in particular semantic domains (see below) and to compare the behavior of individual languages with respect to their lexical retention or replacement in different fields.

One of the most unique features of the volume is the fact that the data described in its individual chapters have been made freely accessible online via the World Loanword Database (WOLD; http://wold.livingsources.org/). Drawing on the generous resources of the Max Planck Institute for Evolutionary Anthropology, the project's sponsor, the loanword typology project follows in the footsteps of the World Atlas of Language Structures (WALS) in matching an online resource to the printed outcome of a collaborative project. Entries can be searched by language, semantic field, or meaning, and the user can filter results drawing on the full set of quantitative and qualitative metadata criteria that accompany the entries. This is a wonderful opportunity for researchers, students, and lay people to explore correlations and regularities in the corpus, and a new landmark in engaging wider audiences in the scientific investigation of linguistic universals.

The bulk of this (literally) heavyweight publication consists of the individual language descriptions; only some generalizations about the corpus are made in the general section that precedes them. A table summarizes the rate of lexical borrowing by language, determined on the basis of the percentage of loanwords in the 1,460 item list. Two languages are identified as showing 'very high borrowing', amounting to over 50% of the list: Selice Romani with 62.7% (a Romani dialect spoken in southern Slovakia; description by Viktor Elšík), and Tarifiyt Berber with 51.7% (Maarten Kossman). Around half of the sample languages are 'high borrowers', with over 25% loanwords in the list. They include English, Romanian, Indonesian, Japanese, Swahili, Thai, and Gurindji. The others are 'average borrowers', with anywhere from 10–25% loans; examples include Hausa, Dutch, Malagasy, Hawaiian, and Hup.

Only four languages are 'low borrowers' with less than 10% loans: Old High German, Manange, Ket, and Mandarin Chinese, which has the lowest borrowing rate in the sample: 1.2%. Ket (contribution by Edward Vajda), with 9.7% loans, is a Yeniseian language of Siberia with only around 200 speakers. They descend from hunter-gatherer-fishers who had little contact with other linguistic groups until the twentieth century. Apart from a peripheral handful of loans (altogether seven words from Mongolian, Evenki, Chinese, and Selkup), the bulk of the loanwords are Russian and pertain mainly to the semantic domains of the modern world, law, the house, and possession (see below). Manange (Kristine Hildebrandt), with 8.3% loans, borrows mainly from Nepali as well as a few words from English. As expected, a high proportion of loans appear in the domain of the modern world (48.3% of the items in the category). Other loans belong to the fields of clothing and grooming, agriculture and vegetation, and religion and belief. By contrast, Old High German loans, amounting to 5.8%, belong to the domains of food and drink, the house, and clothing and grooming.

Although little is said in the comparative summaries about the sociolinguistic circumstances of the sample languages, selecting the two extreme cases, namely Selice Romani and Mandarin

Chinese (see p. 58), one might conclude that high borrowing correlates with universal multilingualism, a minority language status, the absence of a written standard, and sociopolitical marginalization, while low borrowers show little or no bilingualism, a status as a majority language, a powerful standard, and a sociopolitically dominant population. The individual profiles of languages like Ket and Manange, however, show that the situation is much more complex and there is a whole set, not just one linear scale, of social factors that can encourage or impede borrowing.

One of the key questions of the project is whether some words are universally more prone to borrowing than others. Approaching the issue first in structural terms, the compilation shows that on average c. 25% of the vocabulary in the sample as a whole is borrowed, but that languages differ considerably in the ratio of borrowed content words to borrowed function words. The comparison does not seem very meaningful if one takes the position that distinct types of language-processing tasks are triggered by content and function words and that we are therefore dealing with very different motivations for borrowing (cf. Matras 2009). The same can be said for parts of speech. Among nouns, the percentage of loans is 31.2%, while among adjectives and adverbs it is 15.2%, and for verbs it is 14.0%. But the comparative evaluation of the corpus (55–75) is too brief to dwell on the question of whether this is due to frequency in the lexicon of entries belonging to different parts of speech, or to their frequency of occurrence in actual speech, or to their processing function, or indeed due to typological constraints that favor the borrowing of some word classes over others, at least in some languages (thus verbs often require morphological adaptation to assert their 'verbness'; see Matras 2009:175–87, Wohlgemuth 2009).

The semantic domain that is most prone to borrowing (see p. 64) is religion and belief (on average 41.2%), followed by clothing and grooming (38.6%), the house (37.2%), and law (34.3%). The domains with the lowest rates of borrowing are sense and perception (11.0%), spatial relations (14.0%), the body (14.2%), and kinship (15%) (64). Curiously, the domain 'modern world' is not included in the comparative list (64), and in fact is omitted from some of the individual contributions, too. A glance at the individual descriptions reveals the highest borrowing rates for this domain, which includes words like 'radio', 'television', 'driving license', 'government', 'nurse', and 'battery': Swahili shows 73.6% loans in the category 'modern world', Gawwada 67.0%, Kanuri 63.9%, Lower Sorbian 65.7%, Dutch 58.6%, Sakha (Yakut) 84.3%, and Yaqui 83.8%. We can thus assume that the overall borrowing rate is highest in this semantic field (modestly experienced with databases, I have been unable to figure out how to use the online resource to calculate the overall borrowing rate in the sample for a particular semantic field).

A less obvious semantic category is 'possession'. It groups together verbs expressing possession, such as 'to own', 'to have', and 'to keep', with a variety of expressions associated with money (e.g. 'coin', 'rich', 'tax', 'to pay', 'beggar', 'debt') as well as more general expressions related to trade (e.g. 'market', 'price', 'shop'). This field seems to be split between basic activities and commerce. The latter is very often perceived as external to the community and dependent on negotiations with others. With a loanword rate of 27.1%, the semantic field of possession scores slightly higher than average for the borrowing rates of semantic fields. Languages with average borrowing rates differ considerably in their borrowing rate in the field of possession, and we find a spread from Iraqw (12%), Dutch (12.4%), and Gawwada (13%), to Vietnamese (32.6%), Indonesian (34.4%), Hausa (35.8%), Romanian (44.7%), and Swahili (48.1%). Interestingly, Ket, with an otherwise very low borrowing rate, shows loans in 24.1% of word meanings in the field of possession. This may provide an insight into just how complex the interrelationships are between borrowing, cultural contact, and social activity domains. Societies whose commerce depended strongly on contact with dominant outsiders show a tendency to adopt terminology from those foreign traders, more so than those whose outside trade was in their own hands. Borrowing universals are therefore difficult to establish unless we limit ourselves to common sociocultural settings.

Nevertheless, some remarkable insights are gained from the sample as a whole. For entire sets of meanings we find low borrowing rates across the corpus. These include body parts, universally present nature phenomena, generic actions, basic properties, personal pronouns, and basic interrogatives—confirming predictions and observations made in Matras 2009:166–92. The project

leaders extract data on resistance to borrowing, adding to the score that reflects a word's loanword status, its scores for analyzability and age, and its representation score (the likelihood that its meaning is represented in the vocabulary of a language). These combine into a composite score on the basis of which a list of one hundred least borrowable or most stable words is compiled—the Leipzig-Jakarta 100-item list of basic vocabulary.

We now have a new tool to measure genetic affiliation and a new concept of what is stable or basic vocabulary: it includes body parts and especially external organs like 'mouth', 'ear', 'eye', 'arm', and 'nose'; universally present natural phenomena like 'water', 'fire', 'rain', 'night', 'star', 'wind', 'rock/stone', and others; generic animals terms like 'fish' and 'bird' as well as terms for creatures found wherever there are humans, such as 'louse', 'ant', 'fly', and 'dog'. Generic actions on the list include movement verbs like 'to go' and 'to come' as well as basic activities such as 'to eat', 'to drink', and 'to laugh', and sense perception verbs like 'to see' and 'to hear'; basic properties like 'big', 'small', 'old', and 'new' as well as the color terms 'black' and 'red'; the singular pronouns 'I', 'you', and 'he/she/it'; the interrogatives 'what', 'where', 'which', and more. The new list ends up bearing a close resemblance to the one that it replaces, having sixty-two items of overlap with the Swadesh list (73).

What can we make of the project's findings? We now have clear crosslinguistic evidence that aspects of social organization (modern world, possession or commerce, religion, law, household) are more prone to borrowing than concepts that mirror natural or physical surroundings (physical world, body, emotions, perception, space, kinship). But what is the explanation for this finding? The volume leaves little room for a theoretical discussion. Loanwords are explained (46) as motivated by cultural innovations, by taboos (against the use of established words), and otherwise by the prestige of the donor (source) language. But for those borrowings that are replacements of existing concepts rather than labels for new concepts, no explanation is offered as to why prestige should motivate loans in one domain but not in another, or why certain semantic constraints should work to resist borrowing.

'Prestige' is of course a sociolinguistic notion and can only be integrated effectively if some consideration is given to theoretical aspects of sociolinguistics in the evaluation of the data. For instance, one would wish to see discussions address the connection between the presence of language institutions such as literacy, a national standard, media and education, use of language in administration and religious functions, and a language's overall susceptibility to borrowing. Furthermore, contact-induced language change, just like internal change, can be 'prestigious', from the 'top', but also 'vernacular' or 'bottom-up'. Thus English has words like *estate* and *constitution*, both from French, once a high language that enjoyed social prestige, alongside words like *chavvy* (a stereotype word for youth of a working-class background) and *minge* (female genitals), both from Romani, the language of a highly stigmatized and marginalized minority. Information on register was included as an optional metadata slot in the database entries, but not assessed either for the sample as a whole, nor, as far as I could tell, in the individual contributions.

Ultimately, a theory of lexical borrowing requires a theory of bilingualism or multilingualism. 'Prestige' may be a way to capture the social relations that exist among two neighboring speech communities, but it is not enough in order to describe the motivation of bilingual speakers to selectively abandon certain words that were once reserved for communication in one set of interaction settings—words belonging to 'language A'—and to generalize instead counterpart words that were hitherto limited to a different set of interaction settings—or 'language B'. Loanwords are bilingual speakers' way of adjusting their overall repertoire of lexical words and the constraints on the selective use of words in certain settings, or with certain interlocutors. The insights into the hierarchical nature of lexical borrowing provide us with an excellent opportunity to explore how this process of renegotiating the bilingual lexical repertoire is related to the conceptualization of reality. As a first and very crude interpretation we might propose that shared lexical repertoire (i.e. lexical items that are generalized throughout the bilingual's repertoire and used irrespective of interaction setting or interlocutor, i.e. 'borrowed') is symbolic of activities that are shared with another, neighboring linguistic community: commerce, religion, administration, technology (whether the concepts are new, or established but replaced by loanwords). By contrast, personal

and family experiences (body, emotions, space) remain conceptually protected and individualized, and this is reflected in the enduring compartmentalization of the relevant linguistic expressions in the bilingual repertoire.

The fact that the volume offers no room for this (or other) kinds of interpretation shows the limitations of purely descriptive typological sampling. The volume delivers valuable data, but not a theory of lexical borrowing. Yet the data offer a factual basis for new insights into language functions and language processing in situations of multilingualism. In addition to an explanatory account of borrowing that could be inspired by the data, the volume under review also invites new discussions of the phenomenon of mixed languages (e.g. Matras & Bakker 2003). As the reader might recall, these are languages that are usually defined by having different sources of grammar and basic vocabulary. So far, no objective measure has been available for 'basic vocabulary', and it would be interesting to review data on mixed languages in order to ascertain whether their definition might be refined in light of the Leipzig-Jakarta list. Many mixed languages, for example, so-called 'Para-Romani' vocabularies, arise through lexical retention from an ancestral language once spoken in the community but abandoned. Here too, the Leipzig-Jakarta list offers an opportunity to compare lexical retention in situations of language shift (so-called language 'afterlife'; see Matras 2010) with lexical borrowing in the conventional sense.

## REFERENCES


EMBLETON, SHEILA M. 1986. *Statistics in historical linguistics*. Bochum: Brockmeyer.

MATRAS, YARON. 2009. *Language contact*. Cambridge: Cambridge University Press.

MATRAS, YARON. 2010. *Romani in Britain: The afterlife of a language*. Edinburgh: Edinburgh University Press.

MATRAS, YARON, and PETER BAKKER (eds.) 2003. *The mixed language debate: Theoretical and empirical advances*. Berlin: Mouton de Gruyter.

MATRAS, YARON, and JEANETTE SAKEL (eds.) 2007. *Grammatical borrowing in cross-linguistic perspective*. Berlin: Mouton de Gruyter.

RENFREW, COLIN; APRIL MCMAHON; and LARRY TRASK (eds.) 2000. *Time depth in historical linguistics*. Cambridge: McDonald Institute for Archaeological Research.

STAROSTIN, GEORGE. 2010. Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship* 3.79–116.

SWADESH, MORRIS. 1950. Salish internal relationships. *International Journal of American Linguistics* 16.157–67.

SWADESH, MORRIS. 1952. Lexicostatistical dating of prehistoric ethnic contacts. *Proceedings of American Philosophical Society* 96.452–63.

WOHLGEMUTH, JAN. 2009. *A typology of verbal borrowings*. Berlin: Mouton de Gruyter.



School of Arts, Languages and Cultures
University of Manchester
Oxford Road
Manchester M13 9PL, United Kingdom
[yaron.matras@manchester.ac.uk]


**Bare grammar:** Lectures on linguistic invariants. By EDWARD L. KEENAN and EDWARD P. STABLER. (Stanford monographs in linguistics.) Stanford, CA: CSLI Publications, 2003. Pp. vi, 192. ISBN 9781575861883. $22.50.

Reviewed by RICHARD T. OEHRLE

Linguistic structure occurs at different scales. At one level, where we model the relation of language to consciously observable physical events (e.g. speech, writing, signing), the postulation of linguistic structure plays an essential role in explaining why we can judge one and the same physical entity in more than one way and why we can judge distinct physical entities to be linguisti-