Borrowability and the notion of basic vocabulary

Uri Tadmor, Martin Haspelmath and Bradley Taylor
Max Planck Institute for Evolutionary Anthropology

This paper reports on a collaborative quantitative study of loanwords in 41 languages, aimed at identifying meanings and groups of meanings that are borrowing-resistant. We find that nouns are more borrowable than adjectives or verbs, that content words are more borrowable than function words, and that different semantic fields also show different proportions of loanwords. Several issues arise when one tries to establish a list of the most borrowing-resistant meanings: Our data include degrees of likelihood of borrowing, not all meanings have counterparts in all languages, many words are compounds or derivatives and hence almost by definition non-loanwords. We also have data on the age of words. There are thus multiple factors that play a role, and we propose a way of combining the factors to yield a new 100-item list of basic vocabulary, called the Leipzig-Jakarta list.

Keywords: loanword, borrowing, basic vocabulary, stability, language contact

1. Assessing degrees of lexical borrowability

Predicting borrowing behavior is important in historical and comparative linguistics for a variety of reasons. Most importantly, borrowing is often a confounding factor in assessing genealogical relatedness of languages. If we were able to determine in general which words are more or less likely to be borrowed, based on their meanings, we would be in a better position to distinguish diffusional similarities from similarities that are due to common ancestry.

Until now, not much has been known about borrowability in general. For some grammatical domains, tentative borrowability scales have been set up (Matras 1998, 2007, Field 2002), and recent years have seen a lot of new research on grammatical borrowing (Aikhenvald & Dixon 2007, Matras & Sakel 2007). However, no systematic cross-linguistic research on lexical borrowing has been carried out before our project (the Loanword Typology project, see §3). A tradition of research examines the stability

Diachronica 27:2 (2010), **226–246**. DOI 10.1075/dia.27.2.04tad
ISSN 0176–4225 / E-ISSN 1569–9714

of lexical meanings (e.g., Dolgopolsky 1986, Lohr 1998, Holman et al. 2008), but it does not specifically address borrowing; in fact, Holman et al. emphasize that stability in general and resistance to borrowing pattern differently.

Linguists have long been in agreement that "basic vocabulary" or "core vocabulary" is more resistant to borrowing than less basic vocabulary, but what exactly is meant by this is often left unclear. However, we can approach this question empirically. If the kinds of words that are borrowed or are resistant to borrowing across languages are not a random selection, but systematically tend to come from certain meaning domains but not others, then we can come up with a list of hard-to-borrow vocabulary by examining a representative set of languages with known loanwords. This is what we have done in the Loanword Typology (LWT) project. One of our results is a list of basic vocabulary based on data from 41 languages from all continents, the Leipzig-Jakarta list of basic vocabulary (see §9).[1] However, as will be explained later, it turned out that low borrowability by itself was not a sufficient criterion and needed to be supplemented by a few other criteria.

2. The notion of basic vocabulary and the Swadesh 100 list

The term "basic vocabulary" can mean different things in different domains of linguistics. For example, in second language acquisition "basic vocabulary" is that part of the lexicon that "would have as wide a communicative range as possible using a minimum number of words of general meaning" (McCarthy 1999: 233). In corpus linguistics, it may be equated with the most frequent words. In historical and comparative linguistics, basic vocabulary has typically been associated with stability, universality, simplicity, and resistance to borrowing.

As Hymes (2006 [1971]: 254) pointed out, "[t]he notion of basic vocabulary is at least as old as comparative linguistics". Swadesh's great contribution was in formulating a standard list of basic vocabulary, utilizing it for various lexicostatistical studies, and perhaps most importantly — inspiring scholars around the world to use it for thousands of languages.

In his writings, Swadesh was quite explicit in describing how he created and refined his 100-item list. He describes the purpose of the list as follows (Swadesh 2006 [1971]: 19):

> In counting and statistics, it is convenient to operate with *representative samples*, that is, a portion of the entire mass of facts so selected as to reflect the essential facts. For our lexical measure of linguistic divergence we need some kind of selected word list, a list of words for which equivalents are found in each language or language variant …

1. The list is named after the locations where it was produced.

Although the major purpose of the list was to help determine relationships among languages already known to be related, Swadesh also pointed out that "[t]he technique of the diagnostic vocabulary, which was developed as an instrument of glottochronological measurement, has come to have other uses, above all some related to problems of remote relationship" (Swadesh 2006 [1971]: 279). In other words, the list was seen as a means of establishing previously unknown genealogical connections, not only determining the degree of relatedness within known groups.

Swadesh defined basic vocabulary as "concepts and experiences common to all human groups" (Swadesh 1950: 157) and "the fundamental everyday vocabulary of any language — as against the specialized or 'cultural' vocabulary" (Swadesh 1952: 452). Later he spelled out more explicitly the elements that constituted his basic vocabulary (Swadesh 2006 [1971]: 275): "… universal and simple things, qualities, and activities, which depend to the least degree possible on the particular environment and cultural state of the group", including "pronouns, some quantitative concepts, parts and simple activities of the body, movements, and some general qualities of size, color, and so on". Explicitly excluded were "words of a cultural nature, words that in many languages are sound-imitative (onomatopoeic) … [and] terms with very specific meanings".

How did Swadesh create his 100-item list? Based on his intuitions, around 1948 he composed a list of meanings that suited his criteria, which he first tested on English (Swadesh 1950: 161). As he later explained, "[t]he first research making use of the diagnostic list led to changes, the elimination of some elements and the substitution of others, and finally the selection of the hundred words" (Swadesh 2006 [1971]: 275).[2] Quantitative studies were apparently not part of his methodology.

3. The Loanword Typology project

In the Loanword Typology (LWT) project, which we coordinated between 2004 and 2009, each language was the responsibility of an author who is a specialist of the language and its history. Each of the authors (or author teams) provided counterparts

2. The number of items on the list fluctuated. The original list tested on English contained 225 items. The list in Swadesh 1950 contained 165 items, many of which, however, were culture-specific (e.g., 'canoe', 'moccasin'). In Swadesh 1950, he only used 121 of these, plus 44 extra forms relevant to Salishan languages which were not used on the general list because they were culture-specific (e.g. 'canoe', 'moccasin'). The final list of 100 items can be found in Swadesh 2006 [1971], although it was formulated years before the publication of the first edition of that book. Many linguists are not aware that the popularly used 200-item list was one of the intermediate lists, not what Swadesh considered to be his final product. The 207-item list, which is also commonly used, was in fact never used by Swadesh himself. Rather, it contains all the meanings on the 200-item list plus seven items that appear on the final 100-item list but not on the 200-item list. Below, we will consider only the final 100-item list.

(=translational equivalents) for lexical meanings on a fixed list of 1460 meanings (the LWT meaning list). When the counterpart word was judged to be a loanword, supplementary information was provided on what is known about the historical circumstances under which it was borrowed. The resulting combined database (which we call the *World Loanword Database*, Haspelmath & Tadmor 2009b) is a lexical database comprising 41 individual language subdatabases (available online at http://wold.livingsources.org).[3] We hope that these languages are reasonably representative of the world's languages; while they do not come close to the ideal of a fully random sample, we do not think that the biases of the sample lead to major skewings in the results. (In contrast to grammatical typology, where genealogical and geographical bias is a well-known confounding factor, language contact typology seems to be much less affected by such factors.)

Each subdatabase contains about 1,000–2,000 words or counterparts of the meanings on the Loanword Typology meaning list. The number of words in each subdatabase varies, because sometimes a language had no counterpart for a particular meaning, while in other cases it had several counterparts. Considerable work by the language experts went into these subdatabases, and each of them constitutes a separate online publication. In addition, each author team wrote a prose chapter describing the borrowing situation in their language. These chapters and a general description of the project and its results are found in Haspelmath & Tadmor (2009a).

The LWT meaning list consists of 1,460 lexical meanings, most of which were adopted from the meaning list of the Intercontinental Dictionary Series, which in turn was based on Buck (1949). We opted for a meaning list that is significantly longer than Swadesh's list, because we wanted to get a more comprehensive picture of the lexicon

---

3. The languages are:

Africa: Swahili, Iraqw, Gawwada, Hausa, Kanuri, Tarifiyt Berber, Seychelles Creole

Europe: English, Old High German, Lower Sorbian, Dutch, Romanian, Selice Romani, Kildin Saami, Bezhta, Archi

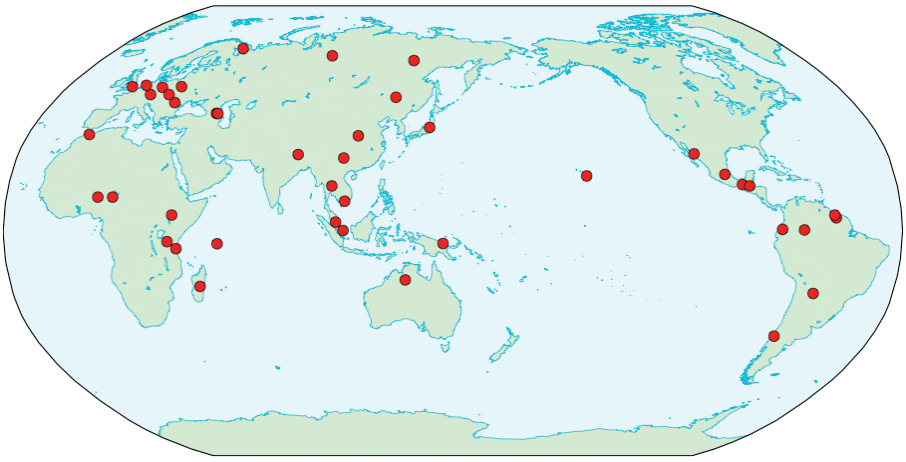
Asia: Manange, Sakha, Mandarin Chinese, Thai, Vietnamese, White Hmong, Japanese, Ket, Oroqen, Ceq Wong, Indonesian

Pacific: Malagasy, Takia, Hawaiian, Gurindji

Americas: Yaqui, Zinacantán Tzotzil, Q'eqchi', Otomi, Saramaccan, Imbabura Quechua, Kari'na, Hup, Wichí, Mapudungun

In selecting languages for inclusion in the project, an effort was made to represent the world's genealogical, geographical, typological, and sociolinguistic diversity. However, the overriding factors were practical. Languages could only be included if a specialist in the language volunteered to invest the considerable amount of time and effort needed to complete the database and to write a book chapter based on the findings. This has no doubt led to a bias in favour of languages with many loanwords, and in favour of well-studied languages.

of each project language. Unlike Swadesh, who determined a priori what constituted basic vocabulary based on his intuitions, and then proceeded to refine his list by trial and error, we wanted the composition of any list derived from the LWT project to be empirically based. We also added about 160 meanings — many of them having to do with the modern world ('hospital', 'newspaper', 'radio') — because we were interested in the extent to which loanwords or native neologisms are used for such new concepts. Many of the meanings on the list were not culture-free, so often a language did not have a counterpart of a given meaning. The geographical distribution of the languages is shown in Map 1.



Map 1. The geographical distribution of the 41 languages in the Loanword Typology project

For each counterpart word, the database gives the (orthographic and/or transcribed) form of the word, information about the analyzability of this word, a morpheme-by-morpheme gloss (for analyzable words), information about loanword status, information about the age of the word, and optional further information of various kinds. For each loanword, the database gives the donor language and the source word (with its meaning), as well as some information about the borrowing circumstances. The loanword status is not a simple binary distinction, but a point on a scale between "No evidence for borrowing" and "Clearly borrowed". For calculating the loanword rates of the project languages (as well as for statistics relating to semantic word classes and semantic field as discussed in §4), we regarded as loanwords all words that were marked as "clearly borrowed" or as "probably borrowed".

One result of our project is a ranking of the languages with respect to the proportion of (clear or probable) loanwords in their vocabulary. The leading borrower is Selice Romani, with 63% loanwords, followed by Tarifiyt Berber (52%), Gurindji (46%), Romanian (42%), and English (41%). The language with the lowest percent-

age of loanwords is Mandarin Chinese (1%), followed by Old High German (6%) and Manange (8%). The average in our sample is 24% loanwords, but it should be noted that our sample is biased toward languages with many loanwords, because linguists working on such languages were more interested in contributing to our project. More details on the differences in the borrowing patterns of the various project languages can be found in Tadmor (2009) and in the online World Loanword Database.

4. Differences among semantic word classes and semantic fields

4.1 Nouns vs. verbs (and adjectives)

It has long been known that languages are more likely to borrow nouns than verbs. This is not only due to the fact that languages have more nouns than verbs. In the consolidated database from our 41 languages, the verb-to-noun ratio is 1:2.5, but the corresponding ratio among the loanwords is 1:5.5 (Table 1). While almost a third of all nouns are loanwords, less than a sixth of the verbs are loanwords. Possible reasons for this are discussed in great detail in Wohlgemuth (2009). Interestingly, adjectives (and adverbs) are almost as hard to borrow as verbs — this is a much less well-known fact which has hardly received any attention so far.

Table 1. Semantic nouns, verbs, and adjectives

Semantic word class	All words	Loanwords	Loanwords as % of total
Nouns	34,355	10,712	31.2%
Verbs	13,808	1,932	14.0%
Adjectives and adverbs	5,284	803	15.2%
All content words	**53,446**	**13,446**	**25.2%**

4.2 Content words vs. function words

Words with grammatical meanings ("function words") are even harder to borrow than verbs. As Table 2 shows, only about 12% of all function words are borrowed.

Table 2. Content words vs. function words

Category	All words	Loanwords	Loanwords as % of total
Content words	53,446	13,446	25.2%
Function words	4,071	492	12.1%
All words	**57,517**	**13,938**	**24.2%**

4.3 Differences among semantic fields

As already seen, not all word meanings are equally often borrowed. Cultural items, such as words relating to religion, clothing, the house, and law, tend to be borrowed often, as shown in Table 3, which ranks semantic fields by percentage of loanwords in the combined database. The semantic fields that are used in this table are the fields of Buck (1949), which are also retained in the Intercontinental Dictionary Series. At the bottom of Table 3, we find semantic fields with relatively culture-free meanings, such as words relating to sense perception, spatial relations, and body parts. Thus, our findings broadly confirm the old view that words with culture-free meanings are less likely to be borrowed.[4] However, since we have information about each individual meaning, we can be much more specific.

Table 3. Semantic fields, ranked by loanword percentage

Semantic field	No. of meanings	Loanwords as % of total
Religion and belief	26	41.2%
Clothing and grooming	59	38.6%
The house	47	37.2%
Law	26	34.3%
Social and political relations	36	31.0%
Agriculture and vegetation	75	30.0%
Food and drink	81	29.3%
Warfare and hunting	40	27.9%
Possession	46	27.1%
Animals	116	25.5%
Cognition	51	24.2%
Basic actions and technology	78	23.8%
Time	57	23.2%
Speech and language	41	22.3%
Quantity	39	20.5%
Emotions and values	48	19.9%
The physical world	75	19.8%
Motion	82	17.3%

4. A reviewer makes the interesting valid point that different semantic fields in Table 3 have different proportions of nouns vs. everything else, and that this could account for some of the differences. One of the reasons why the fields of *Sense perception* and *Spatial relations* show few loanwords is that they contain a much greater proportion of adjectives and verbs than *Clothing and grooming* or *The house*.

Table 3. (*continued*)

Semantic field	No. of meanings	Loanwords as % of total
Kinship	85	15.0%
The body	159	14.2%
Spatial relations	75	14.0%
Sense perception	49	11.0%
All words		**24.2%**

5. The most borrowing-resistant meanings

5.1 Meanings with the fewest (probable or clear) loanword counterparts

One of the main goals of the LWT project was to determine which word meanings are least likely to be borrowed. This task, however, turned out to be far from straightforward. The first problem was that our database provided two different ways of determining the least borrowed meanings. The first was to count the percentage of "clearly borrowed" and "probably borrowed" words among all counterparts for each meaning. A ranking obtained using this method is provided in Table 4. The second method was to assign gradually decreasing numerical values to each of the five loanword statuses[5] and then to calculate the average borrowability scores for each meaning. This was the basis of the ranking in Table 5.[6]

As can readily be seen, the rankings obtained using the two methods are rather different from each other. In Table 4, we see that only 17 meanings on the LWT list have no counterparts that are (clearly or probably) loanwords, among them 13 deictic meanings, three verbal meanings, and one adjectival meaning. There is not a single nominal meaning among them. That meanings of this kind should be at the top of the list of low-borrowability items is not surprising, in view of the fact that function words and verbs are generally much more borrowing-resistant than nouns, as we saw earlier. This also means that a list of the 100 most borrowing-resistant meanings would be very different from the Swadesh list and other high-stability lists, which contain few deictic and grammatical meanings. In Table 5, we see that there are only five meanings (*he/she/it*, *we [inclusive]*, *we [exclusive]*, *up*, *this*) that only have counterparts with no evidence for borrowing, all of them deictics. Since the ranking in Table 5 utilizes a graded score rather than an arbitrary cut-off point as the ranking in Table 4, it was used as the basis for the final list.

5. The assigned values were: "no evidence for borrowing", 1.00; "very little evidence for borrowing", .75; "perhaps borrowed", .50; "probably borrowed', .25; and "clearly borrowed", 0.

6. Because of space limitations, only the first 25 items are listed in each table.

Table 4. LWT meanings ranked by percentage of clearly and probably borrowed counterparts (top 25)

	LWT label	% of clearly and probably borrowed counterparts
1	*he/she/it*	0.00%
1	*we*	0.00%
1	*we (inclusive)*	0.00%
1	*we (exclusive)*	0.00%
1	*itch*	0.00%
1	*spin*	0.00%
1	*rise*	0.00%
1	*up*	0.00%
1	*day after tomorrow*	0.00%
1	*bitter*	0.00%
1	*how?*	0.00%
1	*where?*	0.00%
1	*which?*	0.00%
1	*why?*	0.00%
1	*this*	0.00%
1	*it*	0.00%
17	*say*	1.11%
17	*younger sister*	1.11%
19	*run*	1.21%
20	*married woman*	1.34%
21	*raw*	1.60%
22	*throw*	1.79%
23	*thatch*	1.83%
24	*there*	1.85%
25	*that*	1.88%

6. Representation

The second problem to be tackled was that the meanings in our list differed markedly in **representation**, i.e. the number of languages for which the database has counterparts. While all languages have counterparts for the deictics *where*, *why*, *which*, *there*, *here*, *how*, and for basic verbs like *rise*, *lie down*, *stand*, some highly specific meanings like *mother-in-law of a man*, *netbag*, *tumpline*, and *larch* are only represented in a few of the project languages. If a particular meaning has counterparts in all 41 project

Table 5. LWT meanings ranked by borrowability score (top 25)

	LWT label	Borrowed score
1	*he/she/it*	1.00
1	*we (inclusive)*	1.00
1	*we (exclusive)*	1.00
1	*up*	1.00
1	*this*	1.00
6	*where?*	0.997
7	*why?*	0.995
8	*which?*	0.994
9	*we*	0.991
10	*married woman*	0.990
11	*younger sister*	0.989
11	*rise*	0.989
13	*day after tomorrow*	0.987
13	*spin*	0.987
15	*stinking*	0.982
15	*bring*	0.982
17	*day before yesterday*	0.981
17	*there*	0.981
17	*lie down*	0.981
17	*stand*	0.981
17	*here*	0.981
22	*how?*	0.980
23	*run*	0.976
24	*behind*	0.975
24	*bitter*	0.975

languages, none of which are loanwords, this is excellent evidence that this meaning has very low borrowability. But if a meaning has a counterpart in only one project language, and that word is not a loanword, that hardly constitutes any evidence. It would be improper to make any generalizations based on evidence from just one language (or from just a few languages). Yet if we ranked meanings purely by percentage of loanword counterparts, the meaning with unborrowed counterparts in all project languages and the meaning with just one counterpart that happens to be unborrowed would receive the same score and would have an identical ranking. Obviously, such an approach would be deeply flawed.

In order to address this problem, we computed the representation rate of each meaning as the percentage of the project languages that had a counterpart for it. This score eventually constituted one of the four scores which together made up our composite score (see §9).

7. Analyzability

Some of the low-borrowability meanings in Tables 4 and 5, such as 'younger sister' and 'day after tomorrow', have many counterpart words that are analyzable. This means that these words were most probably created in the language, and not borrowed from another language. For such words the non-borrowed status is not surprising. Words with such meanings are rarely borrowed, but not because of some inherent **resistance to borrowing** — they simply tend not to be loanwords because they are often made up from the resources of the language. In order to take into account the effect of this factor, we computed an average analyzability (or simplicity) score for each meaning.[7]

Analyzable words included complex and compound words as well as phrasal expressions, and semi-analyzable words were those whose complexity is transparent only to linguists, or words containing so-called 'cranberry morphs'. Some of the meanings in Table 4 and Table 5 such as 'to stand', 'bitter', and 'we' show a strong tendency to have unanalyzable counterparts (analyzability score of over .85), while meanings such as 'day after tomorrow' and 'married woman' are much more often analyzable (analyzability score of under .65).

8. Age

The age of words has implications for their usefulness as evidence for borrowability. The longer a word has existed in a language without being replaced by a loanword, the better evidence it constitutes for the low borrowability of its meaning. Therefore information about word ages was collected in a systematic way. For each word, the contributors were asked to give an age, i.e. an approximate year or time period when it was first attested or the oldest time period for which it can be reconstructed. This is relatively easy only for languages with a long attested history such as English and Dutch (where for many loanwords we know the decade in which they were first used, at least in writing). Even for these languages, the age of very old non-loanwords (those going back to Proto-Germanic or even Proto-Indo-European) can be estimated only very roughly. However, such reconstructions are available for many languages even

7. The following values were used: unanalyzable, 1.00; semi-analyzable, 0.75; analyzable, 0.50. The minimum score is 0.50 rather than 0 to avoid overweighting this variable.

Table 6. LWT meanings ranked by age score (top 25)

	LWT label	Age score
1	*fire*	0.939
2	*water*	0.926
3	*tongue*	0.908
4	*nose*	0.906
5	*wing*	0.904
5	*mouth*	0.904
5	*bone*	0.904
8	*arm/hand*	0.903
9	*wind*	0.900
10	*horn*	0.898
10	*the rain*	0.898
10	*to take*	0.898
13	*leg/foot*	0.897
13	*two*	0.897
15	*three*	0.894
16	*one*	0.893
16	*he/she/it*	0.893
16	*you (singular)*	0.893
19	*ash*	0.891
20	*blood*	0.890
21	*flesh/meat*	0.889
22	*ear*	0.888
23	*go*	0.887
24	*name*	0.886
25	*fish*	0.885

where there is no long written history (e.g., for Mayan languages, or for Austronesian languages). We have age information on 88% of our words.

Table 6 gives a ranking of lexical meanings by the average age of the counterparts in the combined database according to their age scores.[8] These scores are deliberately similar to the borrowability scores (cf. §3), in that the oldest words (which are least likely to be loanwords) are assigned a value of 1, and those which are attested only very

8. For this ranking, we assigned the following age values: words first attested or reconstructed earlier than 1000, 1.00; earlier than 1500, 0.90; earlier than 1800, 0.80; earlier than 1900, 0.70; earlier than 1950, 0.60; earlier than 2007, 0.50. The age scale is nonlinear to compensate for the uncertainty in estimates of older ages.

recently and might therefore be (undetected) loanwords are assigned a value of only 0.5. Of course, even the oldest words might be loanwords in the sense that they were borrowed at an even earlier, prehistoric stage, but the very fact that they are old shows that they have not been replaced by loanwords for a very long time.

The 25 meanings in Table 10 contain 12 body-part meanings (or 13 if 'meat' is included), five meanings relating to nature, two basic verbs, three low numerals, and two personal pronouns. With the possible exception of the numerals, these are culture-free meanings in the sense that they represent concepts known to every human society.

9. The Leipzig-Jakarta list of basic vocabulary

If we take into account all four factors discussed above (***borrowed score, representation score, analyzability score, and age score***) and multiply them by each other, we arrive at a composite score. The list of the 100 top-ranking items based on this score are provided in Table 7.[9] This is no longer a simple borrowability ranking, because the borrowability score was just one of four scores used to derive the composite score. In fact, it is a basic vocabulary list that takes into consideration the features normally associated with basic vocabulary in historical and comparative linguistics: resistance to borrowing (the borrowed score), universality (the representation score), simplicity (the analyzability score), and stability (the age score).

The meanings on the Leipzig-Jakarta list can be broken down into the following categories (items also on the Swadesh list are shown in **boldface**):

natural phenomena	***water***, ***fire***, ***night***, *wind*, ***rain***, ***smoke***, ***stone/rock***, *salt*, ***sand***, ***soil***, ***ash***, *shade/shadow*, ***star***
human body parts	***nose***, ***mouth***, ***tongue***, ***eye***, ***tooth***, ***hair***, ***ear***, ***arm/hand***, ***neck***, ***breast***, *navel*, ***liver***, *back*, ***leg/foot***, *thigh*, ***knee***, ***skin/hide***, ***flesh/meat***, ***bone***, ***blood***
animal and plant parts	*wing*, ***horn***, ***tail***, ***egg***, ***root***, ***leaf***, *wood*
humans and animals	*child (descendant)*, ***fish***, ***bird***, ***dog***, *ant*, *fly*, ***head louse***
cultural items	*house*, ***name***, *rope*
properties	*old*, ***new***, ***big***, ***small***, ***long***, *wide*, *far*, *thick*, ***good***, ***red***, ***black***, *heavy*, *sweet*, *bitter*, *hard*

9. The limitation to 100 items is somewhat arbitrary, and is in part in homage to Morris Swadesh. However, for items significantly lower on the list, their ranking on the combined list is less significant as it is often due to a single factor. (Incidentally, it was necessary to make several minor adjustments to the computer-generated ranking, and several meaning labels were edited. This is described in detail in Tadmor 2009.)

actions	*go*, ***come***, *run*, *fall*, *carry*, *take*, ***eat***, ***drink***, *cry/weep*, *tie*, *laugh*, *suck*, *hide*, ***stand***, ***bite***, *hit/beat*, *do/make*, ***burn (intr.)***, *blow*, ***know***, ***see***, ***hear***, ***give***, ***say***, ***crush/grind***
deictic/grammatical	***1SG pronoun***, ***2SG pronoun***, ***3SG*** *pronoun*, ***who?***, ***what?***, ***this***, ***one***, ***not***, *yesterday*, *in*

Table 7. The Leipzig-Jakarta List of Basic Vocabulary

Rank	Word meaning	Borrowed score	Age score	Analyzability score	Representation score	Composite score
1	*fire*	0.965	0.939	0.995	1.000	0.901
2	*nose*	0.973	0.906	0.980	1.000	0.864
3	*to go*	0.963	0.887	0.974	1.000	0.832
4	*water*	0.909	0.926	0.987	1.000	0.831
5	*mouth*	0.920	0.904	0.982	1.000	0.817
6	*tongue*	0.934	0.908	0.954	1.000	0.808
7	*blood*	0.904	0.890	1.000	1.000	0.805
7	*bone*	0.918	0.904	0.971	1.000	0.805
9	*2SG pronoun*	0.958	0.893	0.933	1.000	0.798
9	*root*	0.944	0.869	0.973	1.000	0.798
11	*to come*	0.968	0.876	0.940	1.000	0.796
12	*breast*	0.947	0.856	0.967	1.000	0.783
13	*rain*	0.916	0.898	0.950	1.000	0.782
14	1SG *pronoun*	0.970	0.875	0.936	0.976	0.776
15	*name*	0.915	0.886	0.955	1.000	0.774
15	*louse*	0.950	0.861	0.946	1.000	0.774
17	*wing*	0.884	0.904	0.968	1.000	0.773
18	*flesh/meat*	0.877	0.892	0.986	1.000	0.771
19	*arm/hand*	0.881	0.903	0.966	1.000	0.768
20	*fly*	0.948	0.858	0.942	1.000	0.766
20	*night*	0.931	0.880	0.934	1.000	0.766
22	*ear*	0.896	0.888	0.961	1.000	0.764
23	*neck*	0.895	0.881	0.964	1.000	0.760
23	*far*	0.944	0.850	0.948	1.000	0.760
25	*to do/make*	0.947	0.877	0.914	1.000	0.759
26	*house*	0.893	0.876	0.969	1.000	0.758
27	*stone/rock*	0.895	0.882	0.958	1.000	0.756
28	*bitter*	0.975	0.872	0.889	1.000	0.755
28	*to say*	0.972	0.837	0.928	1.000	0.755
28	*tooth*	0.882	0.877	0.975	1.000	0.755

Table 7. (*continued*)

Rank	Word meaning	Borrowed score	Age score	Analyzability score	Representation score	Composite score
31	*hair*	0.944	0.871	0.917	1.000	0.754
32	*big*	0.889	0.864	0.980	1.000	0.753
32	*one*	0.870	0.893	0.969	1.000	0.753
34	*who?*	0.968	0.838	0.924	1.000	0.749
34	*3SG pronoun*	1.000	0.893	0.955	0.878	0.749
36	*to hit/beat*	0.955	0.827	0.947	1.000	0.748
37	*leg/foot*	0.856	0.897	0.972	1.000	0.747
38	*horn*	0.840	0.898	0.987	1.000	0.745
38	*this*	1.000	0.851	0.897	0.976	0.745
38	*fish*	0.855	0.885	0.984	1.000	0.745
41	*yesterday*	0.958	0.843	0.922	1.000	0.744
42	*to drink*	0.904	0.877	0.934	1.000	0.741
42	*black*	0.951	0.866	0.899	1.000	0.741
42	*navel*	0.878	0.860	0.982	1.000	0.741
45	*to stand*	0.981	0.847	0.889	1.000	0.738
46	*to bite*	0.964	0.861	0.887	1.000	0.736
46	*back*	0.918	0.868	0.924	1.000	0.736
48	*wind*	0.828	0.900	0.987	1.000	0.736
49	*smoke*	0.916	0.863	0.929	1.000	0.734
50	*what?*	0.971	0.804	0.939	1.000	0.732
51	*child (kin term)*	0.929	0.866	0.930	0.976	0.730
52	*egg*	0.910	0.846	0.945	1.000	0.728
53	*to give*	0.913	0.878	0.907	1.000	0.727
53	*new*	0.920	0.860	0.920	1.000	0.727
53	*to burn (intr.)*	0.951	0.860	0.889	1.000	0.727
56	*not*	0.965	0.880	0.974	0.878	0.726
56	*good*	0.893	0.860	0.945	1.000	0.726
58	*to know*	0.933	0.856	0.908	1.000	0.725
59	*knee*	0.911	0.862	0.922	1.000	0.724
59	*sand*	0.901	0.866	0.928	1.000	0.724
61	*to laugh*	0.942	0.844	0.910	1.000	0.723
61	*to hear*	0.953	0.848	0.895	1.000	0.723
63	*soil*	0.900	0.883	0.954	0.951	0.722
64	*leaf*	0.897	0.823	0.977	1.000	0.721
64	*red*	0.926	0.864	0.900	1.000	0.721
66	*liver*	0.869	0.857	0.967	1.000	0.720

Table 7. (*continued*)

Rank	Word meaning	Borrowed score	Age score	Analyzability score	Representation score	Composite score
67	*to hide*	0.928	0.847	0.913	1.000	0.718
67	*skin/hide*	0.889	0.875	0.924	1.000	0.718
67	*to suck*	0.940	0.860	0.888	1.000	0.718
70	*to carry*	0.919	0.838	0.953	0.976	0.717
71	*ant*	0.865	0.850	0.975	1.000	0.716
71	*heavy*	0.911	0.874	0.901	1.000	0.716
71	*to take*	0.900	0.898	0.887	1.000	0.716
74	*old*	0.896	0.867	0.920	1.000	0.715
75	*to eat*	0.920	0.840	0.925	1.000	0.714
76	*thigh*	0.906	0.856	0.918	1.000	0.712
76	*thick*	0.950	0.827	0.906	1.000	0.712
78	*long*	0.956	0.824	0.898	1.000	0.707
79	*to blow*	0.962	0.857	0.878	0.976	0.706
80	*wood*	0.860	0.871	0.940	1.000	0.705
81	*to run*	0.976	0.833	0.867	1.000	0.704
81	*to fall*	0.946	0.825	0.903	1.000	0.704
83	*eye*	0.904	0.847	0.918	1.000	0.703
84	*ash*	0.853	0.891	0.921	1.000	0.699
84	*tail*	0.883	0.813	0.973	1.000	0.699
84	*dog*	0.838	0.869	0.960	1.000	0.699
87	*to cry/weep*	0.871	0.871	0.921	1.000	0.698
88	*to tie*	0.879	0.836	0.948	1.000	0.697
89	*to see*	0.918	0.842	0.900	1.000	0.695
89	*sweet*	0.914	0.857	0.887	1.000	0.695
91	*rope*	0.848	0.824	0.993	1.000	0.694
91	*shade/shadow*	0.887	0.840	0.931	1.000	0.694
91	*bird*	0.842	0.857	0.962	1.000	0.694
91	*salt*	0.848	0.838	0.976	1.000	0.694
91	*small*	0.909	0.790	0.966	1.000	0.694
96	*wide*	0.955	0.819	0.885	1.000	0.692
97	*star*	0.830	0.859	0.970	1.000	0.691
97	*in*	0.948	0.856	0.943	0.902	0.691
99	*hard*	0.918	0.833	0.903	1.000	0.690
100	*to crush/grind*	0.919	0.845	0.886	1.000	0.688

10. The Leipzig-Jakarta list vs. the Swadesh 100 list and three other stability lists

There is a fair degree of correlation between the Swadesh 100 list and the Leipzig-Jakarta list: 62 items on the lists overlap (those in boldface in the list). This means that a total of 38 items on the Leipzig-Jakarta list do not appear on the Swadesh list and vice versa. Swadesh's intuitions thus turn out to have been good, although a 38% difference is substantial and can lead to rather different lexicostatistical results. Moreover, our findings indicate that quite a few items on the Swadesh list are not basic (see the rankings in Table 8). At any rate, the major advantage of the Leipzig-Jakarta list is that it has a strong empirical foundation and is thus a more reliable tool for scientific purposes.

Table 9 compares the 100 meanings of the Leipzig-Jakarta list (in the middle column) with the meanings on three stability lists: Dolgopolsky's (1986) list of 23 stable meanings, Lohr's (1998: 54) list of 61 meanings, and the ASJP's list of 40 meanings (Holman et al. 2008: 336–339, 351–352). Dolgopolsky's and Lohr's lists were established by looking at the kinds of meanings expressed by words that are reconstructed for protolanguages of various families. The ASJP list consists of the 40 most stable meanings of the Swadesh 100 list, where stable meanings are identified due to their greater tendency to yield cognates within groups of closely related languages.

What is striking in particular is that many adjectival and verbal meanings appear on the Leipzig-Jakarta list, but are not part of the stability lists. It thus seems that some meanings may be subject to change (e.g., semantic change, or replacement by novel formations), but not so much subject to borrowing.

Table 8. Items on the Swadesh list but not on the Leipzig-Jakarta list

Item	Our ranking	Item	Our ranking	Item	Our ranking
sit	106	*sleep*	155	*bark*	301
fingernail	107	*white*	157	*walk*	321
man	115	*kill*	159	*swim*	322
belly	118	*many*	166	*seed*	327
two	119	*that*	174	*all*	338
lie	121	*sun*	178	*tree*	345
cloud	123	*woman*	183	*we*	347
fly	134	*dry*	192	*moon*	358
head	137	*grease*	219	*round*	376
hot	143	*heart*	220	*green*	412
cold	146	*yellow*	232	*person*	531
feather	147	*path*	271		
full	153	*die*	291		

Table 9. Items on the Dolgopolsky list (**boldface**), the Lohr 61 list (italics), and the ASJP 40 list (underlined)

semantic domain	meanings on Leipzig-Jakarta list	other meanings
natural phenomena	***water***, *fire*, **night**, wind, rain, smoke, stone/rock, ***salt***, sand, soil, ash, shade/shadow, star	mountain, ***sun***, *day*, *darkness*, *light*, *moon*, *sky*
human body parts	nose, mouth, **tongue**, **eye**, **tooth**, hair, **ear**, **arm/hand**, neck, breast, navel, *liver*, back, *leg/foot*, thigh, knee, skin/hide, *flesh/meat*, *bone*, ***blood***	***fingernail***, **heart**, *head*, *shoulder*, *stomach*
animal and plant parts	wing, **horn**, tail, egg, root, leaf, wood	*tree*
humans and animals	child (descendant), fish, bird, dog, ant, fly, **(head) louse**	*person*, *snake*
cultural items	*house*, ***name***, rope	*path*
properties	*old*, new, *big*, small, *long*, wide, far, thick, good, red, black, heavy, sweet, bitter, hard	**full**, *other*, *thin*
actions	*go*, *come*, run, fall, *carry*, *take*, *eat*, drink, *cry/weep*, tie, laugh, *suck*, *hide*, stand, bite, *hit/beat*, do/make, burn (intr.), *blow*, know, *see*, hear, *give*, say, crush/grind	***die***, *breathe*, *grind*, *cut*, *defecate*, *grow*, *lie (down)*, *press*, *sleep*, *smell (intr.)*, *split*, *spread*, *turn*, *wrap*
deictic/ grammatical	**1SG pronoun**, **2SG pronoun**, 3SG pronoun, **who?**, what?, *this*, one, ***not***, yesterday, in	***two***, *three*, *four*, we, *over*

11. Conclusions

The large-scale comparative study of loanwords that we carried out with more than 40 colleagues has yielded many results, only some of which we have reported on here. We found that nouns are borrowed much more often than verbs and adjectives, and that content words are borrowed much more often than function words. We also found that different semantic fields of words show different degrees of borrowability, as was suspected before.

We then focused on establishing a list of the least borrowable word meanings, to be compared with Swadesh's 100-item list that was intended to be a list of basic, non-cultural items, not necessarily a list of borrowing-resistant meanings. However, deriving a single ranking from our data was not straightforward, because a number of factors other than the raw borrowability rate were found to be important: the representation rate (the degree to which meanings have counterparts in our data), analyzability (the degree to which words have complex counterparts), and age. These correspond to the notions of universality, simplicity, and stability, respectively, which have long been

associated with the notion of basic vocabulary in historical and comparative linguistics. Thus, we ended up with a list that takes into account all of these factors, which we call the Leipzig-Jakarta list of 100 basic word meanings. This list is similar to Swadesh's 100-item list and various other lists of stable word meanings, but resistance to borrowing and stability are not exactly the same concepts.

The language sample and the resulting database are not free from bias, and the way the results were calculated was necessarily arbitrary to a certain extent. Nevertheless, we consider the Leipzig-Jakarta list to have an advantage over Swadesh's 100-item list, in that it was empirically derived, rather than being based on intuition.

References

Aikhenvald, Alexandra Y. & R.M.W. Dixon, eds. 2007. *Grammars in Contact: A cross-linguistic typology*. Oxford: Oxford University Press.

Buck, Carl Darling. 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*. Chicago: University of Chicago Press.

Dolgopolsky, Aharon B. 1986. "A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia". *Typology, Relationship and Time: A collection of papers on language change and relationship by Soviet linguists* ed. by Vitalij V. Shevoroshkin & Thomas L. Markey, 27–50. Ann Arbor: Karoma.

Field, Fredric W. 2002. *Linguistic Borrowing in Bilingual Contexts.* Amsterdam: Benjamins.

Haspelmath, Martin & Uri Tadmor, eds. 2009a. *Loanwords in the World's Languages: A comparative handbook*. Berlin: De Gruyter Mouton.

Haspelmath, Martin & Uri Tadmor, eds. 2009b. *World Loanword Database.* Munich: Max Planck Digital Library. http://wold.livingsources.org/

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller & Dik Bakker. 2008. "Explorations in automated language classification". *Folia Linguistica* 42:2.331–354.

Hymes, Dell. 2006 [1971]. "From the First Yale School to World Prehistory". *The Origin and Diversification of Language* ed. by Sherzer Joel, 228–270.

Lohr, Marisa. 1998. *Methods for the Genetic Classification of Languages.* PhD dissertation, University of Cambridge.

Matras, Yaron, 1998. "Utterance modifiers and universals of grammatical borrowing". *Linguistics* 36:2.281–331.

Matras, Yaron, 2007. "The borrowability of structural categories". *Grammatical Borrowing in Cross-Linguistic Perspective* ed. by Yaron Matras & Jeanette Sakel, 31–73. Berlin: Mouton de Gruyter.

Matras, Yaron & Jeanette Sakel, eds. 2007. *Grammatical Borrowing in Cross-Linguistic Perspective*. Berlin: Mouton de Gruyter.

McCarthy, Michael. 1999. "What constitutes a basic vocabulary for spoken communication?" *English Language and Literature* 1.233–249.

Swadesh, Morris, 1950. "Salish internal relationships". *International Journal of American Linguistics* 16:4.155–167.

Swadesh, Morris, 1952. "Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos". *Proceedings of the American Philosophical Society* 96:4.452–463.
Swadesh, Morris, 1955. "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics* 21:2.121–137.
Swadesh, Morris, 2006 [1971]. *The Origin and Diversification of Language*. New Brunswick, New Jersey: Transaction Publishers.
Tadmor, Uri, 2009. "Loanwords in the world's Languages: Findings and results". *Loanwords in the World's Languages: A comparative handbook* ed. by Martin Haspelmath & Uri Tadmor, 55–75. Berlin: De Gruyter Mouton.
Wohlgemuth, Jan. 2009. *A Typology of Verbal Borrowings*. (= *Trends in Linguistics* 211.) Berlin: Mouton de Gruyter.

Résumé

Cet article présente une étude en collaboration concernant des emprunts lexicaux dans 41 langues dans le but d'identifier des signifiés ou des groupes de signifiés qui résistent à l'emprunt. Nous constatons que les noms s'empruntent plus facilement que les verbes ou les adjectifs, que des mots lexicaux s'empruntent plus facilement que des mots à fonction grammaticale et que différents champs sémantiques contiennent différentes proportions d'emprunts lexicaux. Si on essaie de dresser une liste des signifiés le moins souvent empruntés différents problèmes surgissent : nos données fournissent des informations sur le degré de probabilité d'un emprunt, certains signifiés n'ont pas d'équivalents dans toutes les langues, beaucoup de mots sont des composés ou des dérivés et, de ce fait, sont presque par définition des termes non-empruntés. Nous disposons également d'informations concernant l'âge des mots. De multiples facteurs rentrent donc en jeu et nous proposons une combinaison des facteurs dans une liste de cent signifiés du vocabulaire de base : la liste Leipzig-Jakarta.

Zusammenfassung

In diesem Aufsatz berichten wir über ein Gemeinschaftsprojekt zur Erforschung von Lehnwörtern in 41 Sprachen, mit dem Ziel, entlehnungsresistente Bedeutungen und Bedeutungsgruppen zu identifizieren. Es stellt sich heraus, dass Substantive entlehnbarer als Adjektive oder Verben sind, dass Inhaltswörter entlehnbarer als Funktionswörter sind, und dass verschiedene semantische Felder verschiedene Anteile an Lehnwörtern zeigen. Bei dem Versuch, eine Liste der entlehnungsresistentesten Bedeutungen zu erstellen, ergeben sich einige Probleme: Wir haben Informationen über mehr oder wenig wahrscheinliche Entlehnung, nicht alle Bedeutungen haben Entsprechungen in allen sprachen, viele Wörter sind Komposita oder Ableitungen und daher fast per definitionem Nichtlehnwörter, und wir haben auch Daten über das Alter der Wörter. Es spielen also viele Faktoren eine Rolle, und wir schlagen eine bestimmte Kombination der Faktoren vor, die eine neue Liste von 100 Bedeutungen für Grundvokabular ergibt, die Leipzig-Jakarta-Liste.

Authors' addresses

Uri Tadmor
Max Planck Institute for Evolutionary Anthropology
Department of Linguistics
Deutscher Platz 6
04103 Leipzig, Germany

uritadmor@yahoo.com

Martin Haspelmath
Max Planck Institute for Evolutionary Anthropology
Department of Linguistics
Deutscher Platz 6
04103 Leipzig, Germany

haspelmt@eva.mpg.de

Bradley Taylor
Max Planck Institute for Evolutionary Anthropology
Department of Linguistics
Deutscher Platz 6
04103 Leipzig, Germany

taylor@eva.mpg.de