Creating a Database on Verb Borrowing Patterns

JAN WOHLGEMUTH
Max Planck Institute for Evolutionary Anthropology, Leipzig

[This paper describes the methodological and practical issues involved in collecting and analyzing data on loan verb accommodation patterns in an associative database. The main difficulties in this process are the availability of information on language contact situations and the abstract classification of borrowing situations as well as other sociolinguistic and metalinguistic information regarding lexical borrowings.]

1. Introduction

The question as to why most languages have more trouble borrowing verbs than nouns, and as to the possible mechanisms and paths by which verbs are being borrowed, was addressed in a preliminary way by Edith Moravcsik in her 1975 paper "Verb borrowing". On the basis of a very small sample of languages, primarily Modern Greek, Hungarian, German, and English, she argues that borrowed verbs are in most cases integrated by means of a denominal verbalizer or by a light verb construction *do* + *loan-verb*.

More recently, George Huttar gave a brief summary of the state of the art on *Linguist List* in March 2002; and in mid-2004 Søren Wichmann collected further data on the various mechanisms of verbal borrowing through the *Linguist List* (Wichmann 2004a). As far as I know, however, no truly substantial typological research has been undertaken on this field thus far.

My dissertation topic is "Towards a typology of verbal borrowing", and the first step towards such a typology will be the collection and classification of data and examples of verb borrowings from a large number of languages. In this paper, I would like to present the database that I have set up for this purpose and to discuss its structure and the methodological problems involved with it.

2. Methodological Issues

As Wichmann (2004b, 2004c) has shown, Moravcsik's generalizations in fact do not hold, and there exist more patterns of verbal borrowings, even within one recipient language. Accordingly, one cannot just make generalizations like "languages of the type p always use borrowing mechanism x". The task for typology is now to detect and explain the variety of these patterns and their distribution across languages.

To do this, few methodological and terminological problems must first be addressed. Such issues are e.g.:

1. Metadata selection
2. Language contact information
3. Borrowing patterns
4. Terminology and definitions
5. Sampling
6. Data availability

I will address each of these issues in the following sections.

2.1 Metadata selection

Since we do not yet know exactly what language-internal and language-external factors might determine the choice of any particular borrowing strategy, it is advisable to collect not only examples of borrowed verbs, but also metadata on the corresponding language contact situation and the languages involved. In most cases, *languages involved* means just two languages, namely the DONOR LANGUAGE (from which the item is borrowed) and the RECIPIENT LANGUAGE (into which the item is borrowed). Sometimes, however, there are more languages involved, e.g. in borrowings that basically are from the ULTIMATE DONOR LANGUAGE Latin, but have been taken over into German or English through the IMMEDIATE DONOR LANGUAGE French. In the database, it is the immediate donor language will be the one linked to typological features, while the ultimate donor language that will only be noted in a remarks field; in some cases, the particular borrowing from the ultimate to the immediate donor will itself be treated as a separate example. The immediate donor language is more important to this study, since the borrowing per se involves the taking over of an actual lexical item from that language, regardless of the word's ultimate origin.

2.1.1 Typological / Structural information

First and foremost, the language-internal features of both the (immediate) donor language and the recipient language are relevant. When one thinks of potential obstacles to verb borrowings, morphosyntactic and phonological differences between the two languages immediately come to mind. Such differences include – but are not limited to – phonotactic constraints with regard to the number and structure of syllables, the orientation of affixation (values: *prefixing and suffixing*, *infixing*, *predominantly prefixing*, *predominantly suffixing*, *no affixation*), the overall morphosyntactic nature of the language (values: *inflectional*, *agglutinative*, *isolating*, *incorporating*), etc.

However, one cannot determine in advance which typological features will actually turn out to be relevant, either for any given example involving a given pair of languages or for the study as a whole. Therefore, I need to collect as much typological information as possible on the languages in my sample. As will be explained below (section 3.2), I have incorporated the database of the WORLD ATLAS OF LANGUAGE STRUCTURES (WALS) into the study so as to have a wide range of typological information at hand.

2.1.2 Sociolinguistic information

Language-external factors also play an important role in lexical and grammatical borrowing. Apart from information on the actual situation when, where and why a verb borrowing occurred, it would be useful to have background knowledge about the size of the speaker communities involved, and their attitudes towards language change and borrowing of lexical items in general. The significance of any example will differ greatly depending on whether it is the only (verbal) borrowing in the language or whether the speakers readily adopt words from other languages.

Other information that might be relevant to understanding the context of the borrowing is the geographical location of the languages involved. This immediately enables one to identify neighboring languages and to find possible areal distributions of borrowing patterns.

While the aforementioned two sets of variables – typological and sociolinguistic – could in principle apply to all examples of verb borrowings in a given pair of languages, one should also bear in mind that they may be specific to an individual borrowing and can

change over time. This can be the case both due to the phonological or morphological structure of the item in question, and to shifts in the social settings over the course of time (cf. section 2.2).

2.1.3 Lexical Information

The lexical information accompanying the examples of verbal borrowing can of course vary from lexeme to lexeme, even within a given language pair. For the purposes of my study I will collect data on the valency (values: *intransitive*, *transitive*, *ditransitive*, *not applicable*) of the word in both the recipient and the donor languages. I will also include data on the lexical status of the borrowed verb in the recipient language: is it an insertion into the lexicon (basically filling a lexical gap), an added synonym to a pre-existing (native) word, or does it replace a word that thereby becomes obsolete? In some cases where I include nonce or ad hoc forms (cf. paragraph 2.4.2); their status will also be marked in this field.

2.1.4 Other metadata

Apart from the above features, some additional information will be included in my database. This primarily involves details about the source (bibliographical reference, page and example numbers) and the degree of reliability of the data, ranking from very high to very low. Reliability here does not chiefly refer to the author(s) cited but rather to the degree of certainty with which one can state the form, status and origin (donor language and word-form therein) of the borrowed verb.

2.2 Language contact information

Borrowings can only take place when two or more languages are somehow in contact. The intensity of this contact has an impact on what kinds of lexical items get borrowed and how. The question is how to insightfully generalize language contact situations, since most contact situations are individual by their very nature.

A very broad taxonomy is provided by the five-point scale of intensity of contact given by Thomason & Kaufman (1988):

(1): casual contact
(2): slightly more intense contact
(3): more intense contact
(4): strong cultural pressure
(5): very strong cultural pressure

These rather abstract degrees of contact intensity, however, are of course not types of contact situations, and this information alone is unlikely to be sufficient for answering the question as to what contact situation may lead to what kind of borrowing behavior. Thus I will provisionally collect information on contact situations in the form of an open list of abstract types. And though I will try to assign every new example I get to one of the already existing situation types, the list may grow over the course of the study.

So far, I have been tentatively assuming the following general scenarios where borrowing may occur:

(01): substrate to colonial language
(02): superstrate colonial language
(03): geographical neighbor
(04): occasional contact (trade etc.)
(05): bilingual individual

(06): substrate to areal lingua franca
(07): superstrate areal lingua franca
(08): substrate migrant language
(09): science and technology, "geek talk"
(10): unknown
(11): secret language, word games, ludling
(12): substrate to areal official language
(13): superstrate areal language
(14): forced bilingualism
(15): multilingual society
(16): diglossia
(17): language attrition
(18): religion, missions, cult
(19): cultural prestige
(20): domain-specific (other)
(21): media etc.

These scenarios are neither exhaustive nor mutually exclusive. In the database, multiple factors from the above list can be combined so as to characterize the particular language contact situation or the circumstances of the particular borrowing as accurately as possible. It is important to stress that this information is related to an individual example, since the contact situation is not fixed for any pair of languages and may well change over time or between different domains.

Historical information on the contact situation of donor and recipient language, such as date of first contact or the duration of the contact (or that particular situation), will not be incorporated into the database. The approximate date of the borrowing, however, will be included if information is available.

2.3 Borrowing patterns

As indicated above, Wichmann (2004b, 2004c) offers a set of structural borrowing patterns with subtypes. I have adapted this set to accommodate the different patterns I have encountered so far. This list is not meant to be exhaustive and may grow if other patterns come to my attention.

Table 1: Types of loan verb embedding (insertion) patterns

MACRO TYPE	SUBTYPE
M 1 - Direct Insertion	S 11 - Direct insertion of root or infinitive-like stem
	S 12 - Direct insertion of inflected form
	S 13 - Direct insertion across word class
M 2 - Indirect Insertion	S 21 - Affixation with a verbalizer
	S 22 - Affixation with a causative/factitive
	S 23 - Affixation with a special borrowing affix
M 3 - Light Verb Strategy	S 31 - Light verb "*do*", "*make*"
	S 32 - Light verb "*go*"
	S 3x - Other light verb
M 4 - Paradigm Insertion	S 41 - Borrowing of verb plus inflectional paradigm
M 5 - other	S 51 - Loan translation

For the sake of space, and since this paper is mostly concerned with the structure of the database, I will not elaborate on these patterns here. A more detailed account can be found in Wichmann (2004b, 2004c) and Wichmann/Wohlgemuth (forthc.).

2.4 Terminology and Definitions

A number of terminological issues arise in conjunction with this collection of loan verbs. Actually both elements of the term *verb borrowing* need to be defined properly in order to yield useful results.

2.4.1 "*Verb*"

While I believe that a universal, cross-linguistic definition of this term is impossible, there are certain parameters which can be used to establish a word class with that label in most languages and to assign particular words to that class. (cf. Baker 2003:23f.). Yet several questions remain even after defining that category:

What exactly should be taken into account in the collection? Only examples where the word in question is a verb in both the donor and the recipient languages? What about verbalized borrowings where the donor-language root is not a verb but a member of another word class? What if the recipient language has a fuzzy noun-verb distinction so that almost any root can be treated like a verb (as e.g. in Māori or Nootka)?

For the time being, I will include examples into the database, if the particular word form functions as a verb (or behaves in a predominantly "verby" way) in the RECIPIENT language, according to Baker's (2003) criteria. This may, from time to time, have the disadvantage of excluding some stative verbs, but also has the advantage of excluding most problematic forms where e.g. a nominal serves as head of the borrowed predicate.

2.4.2 "*Borrowing*"

Apart from the issue of word classes, one still has to address the question as to whether the form actually is a borrowing or rather an ad hoc instance of code-switching on the word level. While this probably cannot be determined for every given example, there is at least one rule of thumb which can be applied: If the word appears in a dictionary of the recipient language or is frequently used in non-metalinguistic contexts, it is more likely to be an established borrowing than an ad hoc form. On the other hand, some authors clearly mark particular "borrowings" or code-switches as not widely used. Such examples may often provide insight into the processes of adapting foreign verbs into the recipient language; hence they will be included in the database, but be marked as questionable forms.

The same holds true for loan translations or calques, which will only be added to the collection if they contribute, or supplement, relevant information on borrowing patterns or overall attitudes towards borrowings in the given language.

2.5 Sampling

Since it is not yet clear which factors the choice of borrowing strategies depends upon, it is advisable not to limit the study to any "representative" sample of recipient (or donor) languages or language pairs. (Furthermore, it is hard to determine what "representative" might mean here.) This should ensure that as many different combinations of languages and contact situations as possible make it into the database. Furthermore, it might prove difficult to find examples for verb borrowings in all the languages or language pairs of any predetermined sample. Leaving out languages for which one could not find appropriate data would then inevitably skew the sample and thus compromise its representativity.

Accordingly, I decided not to limit the number or the genealogical and geographical distribution of the languages taken into consideration. I will, however, try to collect data from the broadest possible range of languages and avoid multiple examples which are too similar to each other in both the languages involved and the patterns used (which for example rules out many Romance-to-Germanic borrowings found in Europe).
A tentative goal is to have examples from at least 200 pairs of languages (more, if possible), and no less than 80 different recipient languages. Nonetheless, my attempts to obtain data from a sound sample of languages distributed across all areas and genera will undoubtedly be constrained by the limited availability of data and by the uneven distribution of grammatical and sociolinguistic information on (verb) borrowings.

2.6 Data availability
As mentioned above, it is not a simple undertaking to collect information on verbal borrowings together with the desired metadata for all of the languages involved. While most modern grammars no longer simply ignore loanwords as "improper language", accounts of the language-specific contact situations and the background of borrowings are normally very brief or not found at all in grammatical descriptions.
For some languages and contact situations, especially those in Europe, data is abundant and all relevant information is readily available. To a lesser extent this is also true for more recent borrowings in colonial and modern contexts worldwide. Information on pre-colonial language contact outside Europe, however, is scarce, to say the least.

3. The database
With all these issues in mind, let us now turn to the database itself, which is specially designed for the purpose of collecting the examples and meta-information.

3.1 Software
Not only because it is used as the tool of the Leipzig Loanword Typology Project, but also because it is one of the few applications capable of handling formatted text and Unicode characters, the loan verb database is managed using FileMaker™ Pro 7.
The structure of the tables and associations of this database are shown below in figure 1. Table 2 on the following page gives the definitions of the field names and abbreviations used in that diagram.

3.2 Structure
The main key to all the information stored is the ID of the particular example, since the examples are the basic units of the database. From any given example, one can then access all information regarding that example (pattern, translation, source, bibliography) as well as general information on the two languages involved.
Typological information on the donor and recipient languages is linked through the language IDs to the lists of languages, families, countries, and typological features in the database of the *World Atlas of Language Structures* (=WALS) (Haspelmath et al. (eds.), 2005). This database is incorporated into my own database and is thus fully searchable (cf. 2.1.1 above, and 3.3 below).

Figure 1: FileMaker™ Pro 7 overview of database field definitions and table relationships

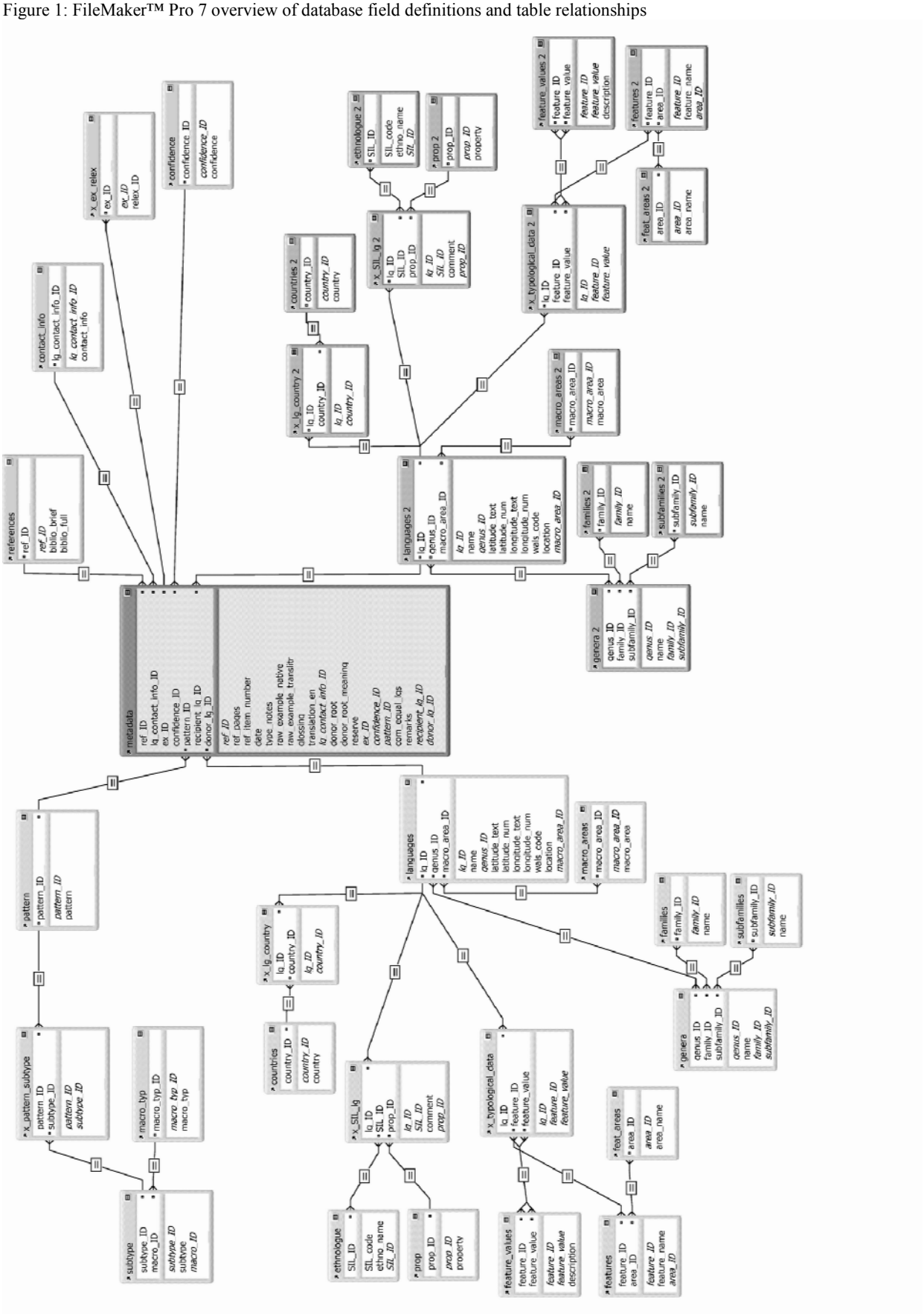

Table 2: Field names and abbreviations used in the database overview:

Table/Field	Abbreviation	Explanation
confidence		degree of reliability of the information
x_ex_relex		cross-reference to related examples
contact_info		information on the type of language contact situation
reference		bibliographical reference for the sources of the examples
	biblio_brief	short bibliographical reference; as used in texts
	biblio_full	full bibliographical reference; as used in reference lists
pattern		the particular pattern of accommodating the borrowed verb / form
x_pattern_subtype		cross-link of patterns to their subtypes
subtypes		subtypes of accommodation of borrowed forms
macro_typ		higher-level type accommodation of borrowed forms
metadata		the main table to which all data is linked
	ref pages	bibliographical information: on what page is the example
	ref item number	example or paragraph number of the example have in the source
	date	estimated date of borrowing
	type notes	any information directly concerning the borrowing pattern
	raw example native	the example in orthographic or phone(ma)tic representation
	raw example translitr	same, transliterated (Latinized or IPA) and hyphenated for glossing
	glossing	interlinear glossing of the example
	translation en	English translation of the example
	donor root	borrowed lexeme as found in the donor language
	donor root meaning	meaning of that form in the donor language
	reserve	additional field for information on the donor language form
	com equal lgs	consistency check disallowing borrowing from a language into itself
	remarks	any further (meta)information on the particular example
languages*		table to access information on donor / recipient language
	name	name of the language as used in this database
	latitude text / num	geographical position (point); segmented, for map generation
	longitude text / num	
	wals code	abbreviation used for this language in WALS
	location	geographical position (point) as one string
x_lg_country*		cross-linking languages to the countries they are spoken in
countries*		list of countries and territories
x_SIL_lg*		cross-linking (wals) language info and ethnologue language info
	comment	indicates whether there are more varieties of that lg. in the database
prop*	property	indicates "is a variety of x" or "has *n* varieties"
ethnologue*	SIL code	SIL (ethnologue) code of the language
	ethno name	SIL (ethnologue) name of the language
x_typological_data*		cross-linking language and typological information
features*		the 142 WALS features (e.g. *feature 38: article*)
	feature name	names of the features, as in WALS
feat_area*	area name	typological area (lexicon, syntax, morphology etc)
feature_value*		indicating which property a feature has in the given language
	feature value	numerical value, e.g. "*5*" linked to the description
	description	description of the value, e.g. "*no definite or indefinite article*"
genera*		linking every language to its genealogical relatives (e.g. *Meso-Philippine*) or to the non-genealogical classes (isolates, creoles, artificial languages, sign languages)
families*		highest-level grouping e.g. *Austronesian*
subfamilies*		intermediate-level grouping e.g. *Western Malayo-Polynesian*
macro area*		linking countries to the six macro-regions of the world (Africa, Australia-New Guinea, Eurasia, North America, Southeast Asia and Oceania, South America)
transitivity		transitivity (itr., trans. ditrans., n/a) of borrowed and original form
LW status		is the word a replacement, insertion, or synonym in the recipient lexicon?

* indicates structures and data originally from the WALS database.

3.3 Using the database

The database has a unique structure due to the fact that I am not working with any particular language or language pair or with fixed donor-recipient relationships. This means that all languages are both possible donor languages and possible recipient languages, and one can sometimes find loan relationships in both directions, e.g. German borrowings into English and vice versa.

Therefore, the language metadata and WALS data are shown twice on the above diagram 1, once for both the donor and once for the recipient language. This allows for database queries like "show me an example of a verb borrowed from an exclusively prefixing language into an exclusively suffixing language" or "show me borrowings from or into languages spoken in Indonesia", and so forth. Furthermore, some peculiarities of loanword adaptations may readily be explained by features of the recipient language. Thus, e.g. the loss of consonants in clusters should not be a surprise, if the recipient language does not allow complex clusters (WALS Map 12, "Complex Syllables" by Ian Maddieson).

4. Concluding remark

While establishing a database to collect examples is but a first step in a project investigating the typology of verb borrowings, such a database – if implemented correctly – will facilitate the work enormously, enabling far quicker access to information and easy detection of correlations governing the behavior of loan verbs.

References:

Baker, M. C. (2003): *Lexical Categories. Verbs, Nouns, and Adjectives*. Cambridge: Cambridge University Press- (Cambridge Studies in Linguistics; 102).

Haspelmath, M. / Dryer, M. S. / Gil, D. / Comrie, B. (eds.) 2005: *The World Atlas of Language Structures*. (Book with interactive CD-ROM) Oxford: Oxford University Press.

Huttar, G. 2003: http://www.linguistlist.org/issues/13/13-588.html

Moravcsik, E. 1975: "Verb borrowing". In: *Wiener Linguistische Gazette* 8, 3-30.

Moravcsik, E. 1978: "Language contact." In: Greenberg H. / Ferguson Ch. (eds.): *Universals of human language*. Vol. 1, 93-122.

Moravcsik, E. 2003: *Borrowed Verbs*. Manuscript, 2003.

Thomason, S. G. / Kaufman, T. 1988: *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Wichmann S. 2004a: http://www.linguistlist.org/issues/15/15-1674.html

Wichmann, S. 2004b: "Structural patterns of verb borrowing". Paper presented at the Workshop on Loan Word Typology; Leipzig, 1-2 May 2004.

Wichmann, S. 2004c: "Loan verbs in a typological perspective". Paper presented at the seminar on contact linguistics, The Linguistic Circle of Copenhagen; 14 December 2004.

Wichmann, S. / Wohlgemuth, J. (forthc.): "Loan verb patterns in a typological perspective". In: Stolz Th. / Palomo, R. / Bakker, D. (eds.): Proceedings of the "Romanicisation worldwide" conference, Bremen 4-8 May 2005.