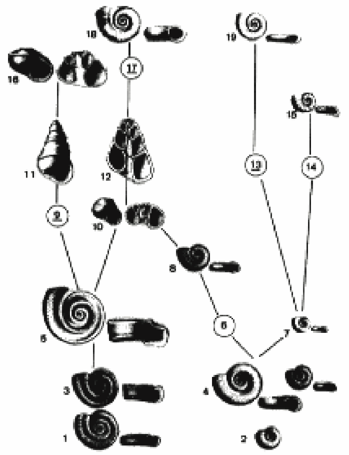
THE GENEALOGICAL WORLD OF PHYLOGENETIC NETWORKS



Automatic detection of borrowing (Open problems in computational diversity linguistics 2)

List, Johann-Mattis

March 2019

Cite as: List, Johann-Mattis (2019): Automatic detection of borrowing (Open problems in computational diversity linguistics 2). The Genealogical World of Phylogenetic Networks 6.1.

http://phylonetworks.blogspot.com/2019/03/automatic-detection-of-borrowing-open.html

Automatic detection of borrowing (Open problems in computational diversity linguistics 2)

The second task on my list of 10 open problems in computational diversity linguistics deals with detecting *borrowings* or *language contact*. The prototypical case of language contact would be *lexical borrowing*, where a word is borrowed from one language into another, such as English *job*, which was adopted by Germans in the rather specific meaning of *temporary occupation*. More complex cases involve *semantic* borrowing, where a way of denoting something is borrowed, not the form itself, such as, for example, the use of the word for *mouse* to denote a *computer mouse* in many languages of the world.

Even less well understood are cases where specific aspects of grammar have been transferred. German has, for example, a certain number of neuter nouns, all borrowed from Ancient Greek or Latin, in which the plural is built according to (or inspired by) the Greek model: *Lexikon* has *Lexika* as plural, *Komma* has *Kommata* as plural, and *Kompositum* has *Komposita* as plural. While these cases are spurious in German and thus rather harmless (as are the similar examples in English), there are other cases of language contact where scholars not only suspect that plural forms have been borrowed along with the words (as in German), but that entire paradigms and strategies of grammatical marking have been adopted by one language from a neighboring variety as a result of close language contact.

Why borrowing is hard to detect

Unless we witness them happening directly, most cases of borrowing are difficult to demonstrate consistently. By comparison with lexical borrowing, however, the borrowing of grammar is probably the hardest to show, especially when dealing with abstract categories that could have actually emerged independently. The reason why borrowing is *generally* hard to deal with, not only in computational approaches, is that detecting borrowing and demonstrating language contact presupposes that alternative explanations are all excluded, such as universal tendencies of language change (i.e., "convergent evolution" in the biological sense), common inheritance, or simple chance.

While we need to exclude alternative possibilities to prove any of the four major types of similarities (*coincidental*, *natural*, *genealogical*, or *contact-induced*, see List 2014: 55-57), we have a much harder time in doing so when dealing with borrowings, because linguistics does not know even one procedure for the identification of borrowings. Instead, we resort to a mix of different types of evidence, which are qualitatively weighted and discussed by the experts. While historical linguistics has developed sophisticated techniques to show that language similarities are genealogical, it has not succeeded to reach the same level of sophistication for the identification of borrowings.

In this regard, techniques for contact detection are not much different from other, more specific, types of linguistic reconstruction, such as the "philological reconstruction" of ancient pronunciations (Jarceva 1990, Sturtevant 1920), the reconstruction of detailed etymologies (Malkiel 1954), or the reconstruction of syntax (Willis 2011).

Traditional strategies for detecting borrowing

It is not easy to give an exhaustive and clear-cut overview of all of the qualitative methods that scholars make use of in order to detect borrowings among languages. This is at least partially due to the nature of "cumulative-evidence arguments" (Berg 1998) — or arguments based on *consilience* (Whewell 1840, Wilson 1998) — which are always more difficult to formalize than clear-cut procedures that yield simple, binary results. Despite the difficulty in determining exact workflows, we can identify a couple of proxies that scholars use to assess whether a given trait has been borrowed or not.

One important class of hints are *conflicts* with possible genealogical explanations. A first type of conflict is represented by similarities shared among unrelated or distantly related languages. Since English *mountain* is reflected only in English, with similar words only in Romance, we could take this as evidence that the English word was borrowed. Since these conflicts arise from the supposed phylogeny of the languages under consideration, we can speak of *phylogeny-related arguments for interference*.

A second conflict involves the traits themselves, most prominently observed in the case of irregular sound correspondence patterns. German *Damm*, for example, is related to English *dam*, but since the expected correspondence for cognates between English and German would yield a German reflex *Tamm* (as it is still reflected in Old High German, see Kluge 2002), we can take this as evidence that the modern German term was borrowed (Pfeifer 1993). We can call these cases *trait-related arguments for contact*.

In addition to observations of conflicts, two further types of evidence are of great importance for inferring contact. The first one is *areal proximity*, and the second one is the assumed *borrowability* of traits. Given that language contact requires the direct contact of speakers of different languages, it is self-evident that geographical proximity, including proximity by means of travel routes, is a necessary argument when proposing contact relations between different varieties.

Furthermore, since direct evidence confirms that linguistic interference does not act to the same degree on all levels of linguistic organisation, the notion of borrowability also plays an important role. Although scholars tend to have different opinions about the concept, most would probably agree with the borrowability scale proposed by Aikhenvald (2007, p. 5), which ranges from "inflectional morphology" and "core vocabulary", representing aspects resistant to borrowing, up to "discourse structure" and the "structure of idioms", representing aspects that are easy to borrow. How core vocabulary can be defined, and how the borrowability of individual concepts can be determined and ranked, however, has been subject to controversial discussions (Lee and Sagart 2008, Starostin 1995, Tadmor 2009, Zenner et al. 2014).

Computational strategies for contact inference

Despite the large number of quantitative applications proposed during the past two decades, computational approaches for the inference of contact situations are still in their infancy. As of now, none of the few approaches proposed in the past can compete with the classical methods. The reasons for this are twofold. First, given the multiple types of evidence employed by the classical approaches, the formalization of the problem of borrowing detection is difficult. Second, given the limited number and suitability of datasets annotated for different types of linguistic interference, scholars have a hard time in developing algorithms, since they lack data for testing and training.

In principle, all algorithms for contact inference proposed so far make use of the strategies used in the classical approaches. Thus, they infer or determine shared traits among two or more languages, and then determine conflicts in these traits, taking geographical closeness and borrowability into account. In contrast to classical approaches, which combine different types of evidence, computational approaches are usually restricted to one type.

The automatic methods proposed so far can be divided into three classes. The first class employs phylogeny-related conflicts to identify those traits whose evolution cannot be explained with a given phylogenetic tree, explaining the conflicts as resulting from contact. Examples include work where I was involved myself (Nelson-Sathi et al. 2011, List et al. 2014), some early and interesting approaches which did not receive too much attention (Minett and Wang 2003), or have been mostly forgotten by now (Nakhleh et al. 2005), along with a recent study on grammatical features (Cathcart et al. 2018).

The second class uses techniques for automatic sequence comparison to search for similar words, but not cognate words, across different languages. Here, the most prominent examples include the work by Ark et al. (2007), and later Mennecier et al. (2016), who searched for similar words among languages known to be *not* related. Further examples include the work by Boc et al. (2010) and Willems et al. (2016), who experimented with tree reconciliation approaches, based on word trees derived from sequence-alignment techniques. There is also an experimental study where I was again involved myself (Hantgan and List forthcoming), in which we tried to identify borrowings by comparing two automatically inferred similarities among words from related and unrelated languages: surface similarities, as reflected by naive alignment algorithms, and deep similarities, reflected by advanced methods that take sound correspondences into account (List 2014).

The third class searches for distribution-related conflicts by comparing the amount of shared words within sublists of differing degrees of borrowability. This class is best represented by Sergey Yakhontov's (1926-2018) work on stable and unstable concept lists (Starostin 1991), which assumed that deep historical relations should surface in those parts of the lexicon that are stable and resistant to borrowing, while recent contact-induced relations would surface rather in those parts of the lexicon that are more prone to borrowing. Yakhontov's work was independently re-invented by Chén (1996), and McMahon et al. (2005); but given how difficult it turned out to distinguish concepts prone to borrowing from those resistant to borrowing, it has been largely disregarded for some time now.

Problems with computational strategies for contact inference

All three classes of approaches discussed so far have certain shortcomings. Phylogeny-based inference of borrowing, for example, tends to drastically overestimate the number of borrowed traits, simply because conflicts in a phylogeny *can* result from undetected borrowings in the data but they never *need to* (see Appendix 1 of Morrison 2011 on *causes of reticulation* in biology, which has many parallels to linguistics). Saying that all instances in which a dataset conflicts with a given phylogeny are borrowings is therefore generally a bad idea. It can be used as a very rough heuristics to come up with potentially wrongly annotated homologies in a dataset, which could then be checked again by experts, but deriving stronger claims from it seems problematic.

While sequence comparison techniques applied to unrelated languages are basically safe in my opinion, and the results are very reliable, unless one compares words that occur in all languages, such as "mama" and "papa" (Jakobson 1960, see also "Mama and papa" on Wikipedia).

Using methods for tree reconciliation on individual word trees, calculated from word distances based on phonetic alignment techniques or similar, yields the same problems of over-counting conflicts as we get for phylogeny-based approaches to borrowing. The problem here is a general misunderstanding of the concept differences between gene trees in biology, where surface similarity of gene sequences is thought to reflect evolutionary history, and word trees in linguistics. While we can use qualitative methods to draw a word tree for a given set of homologous words, the surface similarity among the words says little, if anything, about their evolutionary history.

Attempts to distinguish borrowed from inherited traits with sublists have lost their popularity in most recent studies. When properly applied, they might, indeed, provide some evidence in the search for borrowings or deep homologies. So far, however, all stability rankings of concepts that have been proposed have been based on too small an amount of either concepts (we would need rankings for some 1,000 concepts at least), or languages from which the information was derived. If we could manage to get reliable counts on some 1,000 concepts for a larger sample of the world's languages, this might greatly help our field, as it would provide us with a starting point from which people could search (even qualitatively) for borrowings in their data.

Outlook

Assuming that currently we have no realistic way to operationalize arguments based on consilience, there is no direct hope to have a fully automatic method for detecting borrowings any time soon. By developing promising existing methods further, however, there is a hope that we can learn a lot more about borrowing processes in the world's languages. What is needed here are, of course, the data that we need in order to apply the methods.

In addition to the above-mentioned automatic approaches for borrowing detection, so far, nobody has tried to use trait-related conflicts to infer borrowings. Since these are usually considered to be quite reliable by experts in historical linguistics, it seems inevitable to work in this direction as well, if we want to tackle the problem of consistent automatic detection of borrowing. Here, my recently proposed framework for a consistent handling and identification of *patterns of sound correspondences across multiple languages* (List 2019), could definitely be useful, although it will again be challenging to find the right balance of parameters and interpretation, since not all conflicts in sound correspondences necessarily result from borrowings.

Whether it will be possible to identify even the *direction* of borrowings, when developing these methods further, is an open question. Borrowability accounts might help here, but again, since no clear-cut strategies are being used by scholars, it is difficult to formalize any of the existing qualitative approaches. The greatest challenge will perhaps consist in the creation of a database of known borrowings that could assist digital linguists in testing and training new approaches.

References

Aikhenvald, Alexandra Y. (2007) Grammars in contact. A cross-linguistic perspective. In: Aikhenvald, Alexandra Y. and Dixon, Robert M. W. (eds.) *Grammars in Contact*. Oxford:Oxford University Press. 1-66.

van der Ark, René and Mennecier, Philippe and Nerbonne, John and Manni, Franz (2007) Preliminary identification of language groups and loan words in Central Asia. In: *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*, pp. 13-20.

Berg, Thomas (1998) *Linguistic Structure and Change: an Explanation from Language Processing*. Gloucestershire:Clarendon Press.

Boc, Alix and Di Sciullo, Anna Maria and Makarenkov, Vladimir (2010) Classification of the Indo-European languages using a phylogenetic network approach. In: Locarek-Junge, H. and Weihs, C. (eds.) *Classification as a Tool for Research*. Berlin and Heidelberg:Springer. 647-655.

Cathcart, Chundra and Carling, Gerd and Larson, Filip and Johansson, Richard and Round, Erich (2018) Areal pressure in grammatical evolution. An Indo-European case study. *Diachronica* 35.1: 1-34.

Chén Bǎoyà 陈保亚 (1996) Lùn yǔyán jiēchù yǔ yǔyán liánméng 论语言接触与语言联盟 [Language Contact and Language Unions]. Běijīng 北京:Yǔwén 语文.

Hantgan, Abbie and List, Johann-Mattis (forthcoming) Bangime: Secret language, language isolate, or language island? *Journal of Language Contact*.

Jakobson, Roman (1960): Why 'Mama' and ‘Papa'?. In: *Perspectives in Psychological Theory: Essays in Honor of Heinz Werner*, pp. 124-134.

Jarceva, V. N. (1990) Lingvistil'eskij enciklopedil'eskij slovar'. Moscow: Sovetskaja Enciklopedija.

Kluge, Friedrich (2002) *Etymologisches Wörterbuch der deutschen Sprache*. Berlin:de Gruyter.

Lee, Yeon-Ju and Sagart, Laurent (2008) No limits to borrowing: The case of Bai and Chinese. *Diachronica* 25.3: 357-385.

List, Johann-Mattis and Nelson-Sathi, Shijulal and Geisler, Hans and Martin, William (2014) Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays* 36.2: 141-150.

List, Johann-Mattis (2014) *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis (2019) Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 1.45: 137-161.

Malkiel, Yakov (1954): Etymology and the structure of word families. *Word* 10.2-3: 265-274.

McMahon, April and Heggarty, Paul and McMahon, Robert and Slaska, Natalia (2005) Swadesh sublists and the benefits of borrowing: an Andean case study. *Transactions of the Philological Society* 103: 147-170.

Phillipe Mennecier and John Nerbonne and Evelyne Heyer and Franz Manni (2016) A Central Asian language survey. *Language Dynamics and Change* 6.1: 57–98.

Minett, James W. and Wang, William S.-Y. (2003) On detecting borrowing. *Diachronica* 20.2: 289–330.

Morrison, D. A. (2011) An Introduction to Phylogenetic Networks. Uppsala: RJR Productions.

Nakhleh, Luay and Ringe, Don and Warnow, Tandy (2005) Perfect Phylogenetic

Networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* 81.2: 382-420.

Nelson-Sathi, Shijulal and List, Johann-Mattis and Geisler, Hans and Fangerau, Heiner and Gray, Russell D. and Martin, William and Dagan, Tal (2011) Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society of London B: Biological Sciences* 278.1713: 1794-1803.

Pfeifer, Wolfgang (1993) *Etymologisches Wörterbuch des Deutschen*. Berlin: Akademie.

Starostin, Sergej Anatolévic (1991) Altajskaja problema i proischoždenije japonskogo jazyka [The Altaic Problem and the Origin of the Japanese Language]. Moscow: Nauka.

Starostin, Sergej Anatolévic (1995) Old Chinese vocabulary: A historical perspective. In: Wang, William S.-Y. (ed.) *The Ancestry of the Chinese Language*. Berkeley: University of California Press, pp. 225-251.

Sturtevant, Edgar H. (1920) *The Pronunciation of Greek and Latin*. Chicago: University of Chicago Press.

Tadmor, Uri (2009): Loanwords in the world’s languages. Findings and results. In: Haspelmath, Martin and Tadmor, Uri (eds.) *Loanwords in the World's Languages*. Berlin and New York: de Gruyter, pp. 55-75.

Whewell, William D. D. (1847) *The Philosophy of the Inductive Sciences, Founded Upon Their History*. London: John W. Parker.

Willems, Matthieu and Lord, Etienne and Laforest, Louise and Labelle, Gilbert and Lapointe, François-Joseph and Di Sciullo, Anna Maria and Makarenkov, Vladimir (2016) Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology* 16.1: 1-18.

David Willis (2011) Reconstructing last week's weather: Syntactic reconstruction and Brythonic free relatives. *Journal of Linguistics* 47.2: 407-446.

Wilson, Edward O. (1998) *Consilience: the Unity of Knowledge*. New York: Vintage Books.

Zenner, Eline and Dirk Speelman and Dirk Geeraerts (2014) Core vocabulary, borrowability and entrenchment. *Diachronica* 31.1: 74–105.