Oleg Belyaev

Temporal stability of features in Jazyki Mira

September 28, 2008

Abstract

Stability of a language feature reflects its susceptibility to change, i. e., if defined in more rigorous terms, it is the probability that the feature remains unchanged during an arbitrary period of time. The idea that some features are more prone to mutate due to areal contact than others is not a new one and has been elaborated on, among others, by Edward Sapir (Sapir 1921) and, later, by Joseph Greenberg (Greenberg 1978). (Nichols 1992) was, however, the first study to propose precise mathematical metrics for calculating the parameter of stability and applying them to a relatively wide range of the world's languages.

New large typological databases like WALS in its digital form (Haspelmath et al. 2008) provide new opportunities for conducting similar quantitative research. A recent study notable in this regard is (Wichmann & Holman n. d.), where a metric similar in spirit to Nichols' metrics is applied to the data of the WALS database, making it the first such effort to use the data of thousands of different languages.

Jazyki Mira ('Languages of the World', JM) is another large database analogous to WALS which has been developed in the Institute of Linguistics of the Russian Academy of Sciences starting from the mid-1980s (cf. (Polyakov & Solovyev 2006) for a full description). It is mostly limited in scope to Eurasia and consequently contains only 318 languages. JM's advantage is, however, in its detail, both in terms of the representation of different local genera and in terms of the feature set: JM contains 3821 binary features on most aspects of grammar, and each language is supposed to be exhaustively described by them. A calculation of stability conducted on JM's data would not only be useful to double-check the results received on WALS and to verify if stability is tied to linguistic areas, but would also serve as a plausible means of comparison of the two databases' performance.

Since (Wichmann & Holman n. d.) is, to the author's knowledge, the only quantitative approach to this problem to be based on WALS or any other typological database, it seemed natural to apply the same algorithm to JM, using the same classification of languages and comparing the results for features with more-or-less reliable correspondences.

The results that have been received to this moment are mixed. The first problem was establishing corresponding features for WALS and JM, which is a complicated undertaking in its own right. Even for the relatively small set of 42 most reliable pairs absolute values for stabilities do not seem to correlate in any plausible way. On the other hand, approximate evaluation of these values based on a four-way division of the whole range of percentages (as used in

(Wichmann & Holman n. d.)) shows a much better correspondence: 23 (∼55%) fall into the same categories, 13 (∼31%) – into adjacent ones, and only 6 (∼14%) have no correlation at all. This means that the correspondences between stabilities are acceptable for 86% of all feature pairs. When comparing with statements in the literature (as it is done in the original paper), the results are even better, with only 2 pairs with uncertain correlation out of 13. This seems to bring us to the conclusion that while the databases are quite different in data structure and scope, they do often reflect the same typological realities and are suitable for similar objectives with similar final results. Among other things, this means that one of them can be used for double-checking the results gained on the other one.

Another important observation can be drawn from comparing the frequency distribution of JM stabilites with the distribution of WALS stabilities (for features decomposed into binary form). Both of them resemble Gaussian distributions and both have the mean at about 0.2. Generally speaking, the most stable features seem to constitute a relatively small 'core' of language.

A possible way of testing the plausibility of JM's results would be to conduct a simulation similar to the one conducted on the WALS results, but another possibility is using stability values as weights while calculating typological (dis)similarity of languages. In principle, more stable features should better reflect genetic relationships, while the unstable ones should show more evidence of areal contact. Such a result would demonstrate the metric's plausibility in practice.

The author would like to thank Søren Wichmann, Eric W. Holman, Vladimir Polyakov, Valery Solovyev and Dmitry Egorov for their helpful comments, suggestions and collaboration on this project.

References

[Greenberg 1978] Greenberg, Joseph H. *Diachrony, synchrony and language universals.* Universals of Human Language, Vol. III: Word Structure, ed. by Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik, 47–82. Stanford: Stanford University Press, 1978.

[Haspelmath et al. 2008] Haspelmath, Martin, Matthew S. Dryer, David Gil and Bernard Comrie (eds.) *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library, 2008. Available online at http://wals.info .

[Nichols 1992] Nichols, Johanna. *Linguistic Diversity in Space and Time.* Chicago: The University of Chicago Press, 1992.

[Polyakov & Solovyev 2006] Polyakov, Vladimir N. Valery D. Solovyev. *Компьютерные модели и методы в типологии и компаративистике* (Computational Models and Methods in Typology and Comparative Linguistics). Kazan: Kazanskij Gosudarstvennyj Universitet, 2006.

[Sapir 1921] Sapir, Edward. *Language: An introduction to the study of speech.* New York: Harcourt, Brace and company, 1921.

[Wichmann & Holman n. d.] Wichmann, Søren and Eric W. Holman: *Assessing temporal stability for linguistic typological features.*, pending publication, available online: http://email.eva.mpg.de/w̃ichmann/WichmannHolmanIniSubmit.pdf .

Margaret J. Blake & Harald Hammarström

Swadesh, meet Levenshtein: A Quantitative Method from Dialectology Applied to Phonetic Glottochronological Data

Glottochronology and lexicostatistics, despite their flaws, do provide useful information about language relationships and lexical change. However, it is crucial that lexical change not be viewed as equivalent to language change, and that other complementary tools and techniques be developed to give a better picture of overall language change, which also includes morphosyntactic change, phonetic/phonological change, etc. Additionally, the assumption that language change occurs at a constant rate for all languages is one of the major flaws in Swadesh's original approach to glottochronology; simple inspection of, for example, modern Icelandic and modern Danish against Old Norse suggests quite the opposite (see Trudgill 2007 for exploration of the causes of this difference). To that end, the authors propose a novel technique for quantifying phonetic change, using the Levenshtein string distance algorithm to calculate degree of phonetic change through time upon a cognate subset of the Swadesh-200 list for the insular and peninsular Scandinavian languages, as well as their parent language, Old Norse.

The Levenshtein distance algorithm calculates the difference between two strings by calculating the minimum number of additions, deletions, and substitutions necessary to transform one string into another, normalized by dividing by the length of the longer string. In linguistic applications, the algorithm is typically applied to phonetic transcriptions of words rather than the strings themselves. In addition, modifications have been made to the Levenshtein algorithm to make it more reflective of language change, such as calculating only vowel-to-vowel and consonant-to-consonant changes (Heeringa & Gooskens 2003), as well as the more conservative Almeida-Braun variation, which considers a greater number of phonetic features and incorporates each set of relevant features as an axis in a multidimensional distance calculation (Heeringa & Braun 2003). Levenshtein distance has been shown to correspond significantly with perceptual evaluations of dialect distance (Gooskens & Heeringa 2004), and thereby may yield useful results for other areas of language study where difference and/or change are involved.

Although there are numerous criticisms of the validity of the Swadesh list, it does have many strengths to offer (see e.g. Renfrew et al., eds. 2000 for discussion of the relevant issues): it is pre-existing, readily available for many languages, and contains enough cognates for closely-related languages to provide a statistically significant data set (Kessler 2001). Criticisms have also been leveled against the Levenshtein algorithm (Heggarty 2006), but both methods can yield meaningful results when applied in a conservative, linguistically-informed manner (ibid., Heggarty 2000). To that end, the Scandinavian languages provide an ideal data set, as they are long-documented, well-studied, and have a long history of the written word from which to draw data (as well as extensive reconstruction of former incarnations of their phonetic systems). The accuracy of our analysis will be judged against non-glottochronological measures, such as socio-historical data (König and van der Auwera 1994), data concerning intermediate forms between the parent and modern varieties, and mutual intelligibility judgments (Delsing & Åkesson 2005).

References

Delsing, L.-O. & Åkesson, K.L. (2005) Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska. Nordic Ministerråd, Copenhagen.

Gooskens, C. & Heeringa, W. (2004) Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. Language Variation and Change, **16**, 189-207.

Heeringa, W. & Braun, A. (2003) The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances. Computers and the Humanities, **37**, 257-271.

Heeringa, W. & Gooskens, C. (2003) Norwegian Dialects Examined Perceptually and Acoustically. Computers and the Humanities, **37**, 293-315.

Heggarty, P. (2000) Quantifying change over time in phonetics. In Time Depth in Historical Linguistics, (Renfrew, C., McMahon, A. & Trask, L., eds.) The McDonald Institute for Archaeological Research, Cambridge, pp. 531-562.

Heggarty, P. (2006) Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully Be Applied to Language Data – and to Dating Language? In Phylogenetic Methods and the Prehistory of Languages, (Forster, P. & Renfrew, C., eds.) McDonald Institute for Archaeological Research, Cambridge, pp. 183-194.

Kessler, B. (2001) The Significance of Word Lists. Center for the Study of Language and Information, Stanford, CA.

König, E. & van der Auwera, J. (Eds.) (1994) *The Germanic Languages*. Routledge, London.

Renfrew, C., McMahon, A. & Trask, L. (Eds.) (2000) Time Depth in Historical Linguistics The McDonald Institute for Archaeological Research, Cambridge.

Trudgill, P. (2007) Sociolinguistic dialect typology: contact and isolation in Nordic dialects. In Nordisk dialektologi og sociolingvistik, (Arboe, T., ed.) Peter Skautrup Centret for Jysk Dialekforskning, Århus, 33-53.

Lexical Stability Across Deep Divides:
Lessons from Austronesian-Ongan Comparisons

Juliette Blevins
Max Planck Institute for Evolutionary Anthropology

Application of the comparative method suggests that Onge and Jarawa, two languages of the Andaman Islands, might be distantly related to Proto-Austronesian (Blevins 2007). A range of regular sound correspondences between Proto-Ongan and Proto-Austronesian are proposed, and a number of basic vocabulary items are reconstructed for Proto-Austronesian-Ongan. Reconstructable items are compared with the stability indices of Swadesh items proposed by Holman et al. (2008). If these languages are indeed related, significant meaning shifts have occurred among the most stable items. Semantic shifs include: Proto-Austronesian LIVER = Proto-Ongan BLOOD; Proto-Austronesian ARTERY/VEIN/MUSCLE/SINEW/TENDON = Proto-Ongan LIVER. Overall, findings suggest that when time depths are potentially deeper than those used to assess stability, some semantic leeway should be allowed when regular sound correspondences are in evidence.

Blevins, Juliette. 2007. A long lost sister of Proto-Austronesian? Proto-Ongan, mother of Jarawa and Onge of the Andaman Islands. *Oceanic Linguistics* 46:154-198.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42: 331-354

Swadesh and Prehistoric Linguistics

Daniel Cazés-Menache

Once Swadesh arrived at the main principle, method and techniques of lexicostatistic glottochronology, he and his collegues and students continued the search on the origins and evolution of language. He proposed a system of protophonemes and the rules of "horizontal" and "vertical" variations. With this view, he completed the Mexican Archives of World Languages, in which vocabularies having between 600 and 2000 words are included with their stems, and for each language the same number of phonetic and semantic reconstructions and approximations. This permitted punctual and massive comparisons between languages supposed to be of the same origin, and brought out similarities that could approximate the relationships between very different languages. This procedure followed the principles of known reconstruction of prtolanguages (documented for more than 10 centuries) and led to new proposed genetic classifications (mainly of languages with less than 10 centuries).

Thus emerged the World Linguistic Network in which all languages presently spoken are related at different (little or very big) time profundities.

Later, I made complete comparisons and reconstructions of the Oto-Pamean (hña-maklasinka-meko mychrophylum) and proveed that careful complete reconstructions permit more accurate comparisons and lead to smaller time depths than the obtained in general and provisional comparison-reconstructios.

Following the methods of cultural reconstruction from shared terms at different depths, according to Swadesh and Bounak, I propose a scheme of paleoanthopological and social organization reconstruction, together with a linguistic one, parallelized with the development of thinking and image (sculptures, paintings, writing and mythical representations).

It is my intention to present this complex cultural evolutive theory as a souvenir of my more tan a decade of work as student, assistant and collegue of Morris (Mauricio) Swadesh.

Dr. Daniel Cazés-Menache.
Researcher (since 1984) and Director (2000-08) of the Centro de Investigaciones Interdisciplinarias en Ciencias y Humanifdades,–Universidad Nacional Autónoma de México.
Chargé de recherches at the Centre National de la Recherche Scientifique, Laboratoire de Linguistique Amérindienne (1970-79).
Founder and director of the Colegio de Antropología of the Universidad Autónoma de Puebla (Mexico, 1979-81), and its Secretary General (1981-84).

Lexical and geographical distances as a tool to address the demographic history underlying the Bantu migrations

de Filippo C.[1], Mundry R.[1], Bostoen K.[2], Pakendorf B.[1]

1. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.
2. Royal Museum for Central Africa, Tervuren, Belgium.

During the last few decades, the evolution of Bantu languages and the spread of Bantu-speaking peoples across sub-Saharan Africa from their presumed homeland in the Benue Valley have been under debate in various disciplines. Bantu language trees constructed using the Swadesh word lists of Bastin et al.[1] have produced conflicting results, not only in resolving the phylogeny of the Bantu family but also in suggesting putative models – given a certain tree – for the migrations of Bantu-speaking groups.
In order to test different models of migrations, in this study we generated a matrix of lexical distances among 95 Bantu languages based on the 92-word lists of Bastin et al.[1], as well as matrices of hypothetical geographic distances according to different models of migrations. A correlation approach (Mantel's test) among these matrices of distances was then applied. The results indicate that the best correlation of lexical distances is with the actual geographic distances among groups, rather than the hypothetical distances predicted by any of the models. Therefore, the models of migration tested here are rejected as the major causes of the linguistic patterns observed.
This highlights the complex history of sub-Saharan Africa and the intricate demographic scenario of the Bantu migrations. This would also point out the importance that further multidisciplinary approaches (e.g. by means of genetic data) would have to elucidate the migrations of Bantu-speaking populations.

Reference

1 Bastin, Y., Coupez, A. & Mann, M. (1999). *Continuity and divergence in the Bantu languages : perspectives from a lexicostatistic study*. Tervuren, Belgium: Muse\0301e royal de l'Afrique centrale

Anthony Grant

On making more of qualitative lexicostatistics

In many fields in the social science world, approaches to arriving at the solution of a problem may be *quantitative* (involving a numerically or statistically-based approach) or *qualitative*, in which the nature of the material under examination is itself examined closely. Most work on lexicostatistics has naturally been quantitative in nature because it involves the use of statistical techniques, as was most of Swadesh's work in the 1950s and 1960s (e.g. Swadesh 1950, 1951, 1952, 1955), but this has not always been so. (The Indo-European work done by the team under Don Ringe, for instance Ringe, Taylor and Warnow 2002, is an exception). Lexicostatistics, when used in attempts at the classification and filiation of groups of lects, lends itself admirably to qualitative approaches which use character-based methods in order to examine the degree to which a set of referents for the same gloss can be said to be cognate or non-cognate.

I present case studies of the use of qualitative lexicostatistics in examining subgrouping in a wide range of the world's families, and suggest that by using the techniques inherent in Swadesh's writings one can arrive – without extra effort - at a much more sharply nuanced picture of the historical and other interrelationships between groups of languages which derive from a single common ancestor.

References:

Ringe, Don, Ann Taylor and Tandy Warnow. 2002. 'Computational cladistics and Indo-European.' *Transactions of the Philological Society* 101: 51-117.

Swadesh, Morris. 1950. 'Salish internal relationships.' *International Journal of American Linguistics* 16: 157-167.

---- 1951. 'Diffusional cumulation and archaic residue as historical explanations.' *Southwestern Journal of Anthropology* 7: 1-21.

---- 1952. 'Lexico-statistical dating of prehistoric ethnic contacts.' *Proceedings of the American Philosophical Society* 96: 452-463.

---- 1955. 'Toward greater accuracy in lexico-statistical dating.' *International Journal of American Linguistics* 21: 121-137.

Harald Hammerström

A Full-Scale Test of The Language Farming Dispersal Hypothesis

One attempt at explaining why some language families are big (while others are small) is the hypothesis that the families that are now large became large because their ancestors had a technological advantage, most often farming[1] (Renfrew 1997).

While it has been pointed by Wichmann (2005), it is not clear that we need an explanation of this kind at all, since simple language-split models may also produce language family sizes observed. However, as we shall show, the large families are not a random selection, as one would expect from a simple language split model.

There have been many case studies of the language/farming-dispersal hypothesis for specific families, e.g., Blench (2006), Blench (2005), Holden (2002), Diamond and Bellwood (2003), and a large number of papers in Bellwood and Renfrew (2002). What is lacking is a cross-linguistic test, accounting for *all* factual data.

We have a compiled a database of *every* attested language families in the world and (bluntly but sensitively) assessed their category as either a hunter-gatherer or agricultural family. (For the data to be complete, it is hard to use a more fine-grained categorization.) We also have rough data on location and geospatial size of all families.

The following two tests will be discussed:

- Does the farming have any explanatory power in predicting which families are large (and which are not)?
- Does the geospatial distribution of the observed farming language families show an east-west spread (rather than a north-south) as predicted if the cause of their spread is farming, cf. (Diamond 1997)?

References

Bellwood, P. and C. Renfrew (Eds.) (2002). *Examining the farming/language disperal hypothesis*. McDonald Institute Monographs. McDonald Institute for Archaeological Research, Oxford.

Blench, R. (2005). From the mountains to the valleys: Understanding ethnolinguistic geography in southeast asia. In L. Sagart, R. M. Blench, and A. Sanchez-Mazas (Eds.), *The peopling of East Asia*, pp. 31–50. Routledge, London & New York.

Blench, R. (2006). *Language, Archaeology and the African Past*, Volume 10 of *African Archaeology Series*. Altamira Press, Lanham, MD.

Diamond, J. (1997). *Guns, germs and steel: the fates of human societies*. Cape, London.

Diamond, J. and P. Bellwood (2003). Farmer and their languages: The first expansions. *Science 300*, 596–603.

Evans, N. and R. Jones (1997). The cradle of the pama-nyungans: Archaeological and linguistic speculations. In P. McConvell and N. Evans (Eds.), *Archaeology and Linguistics: Aboriginal Australia in Global Perspective*, pp. 385–417. Oxford University Press, Melbourne.

[1] For a case invoked where the technological advantage was not farming, see Evans and Jones (1997).

Holden, C. J. (2002). Bantu language trees reflect the spread of farming across sub-saharan africa: a maximum-parsimony analysis. *Proceedings of the Royal Society of London, Series B 269*, 793–799.

Renfrew, C. (1997). World linguistic diversity and farming dispersals. In R. M. Blench and M. Spriggs (Eds.), *Archaeology and Language, I*, Volume 27 of *One World Archaeology*, pp. 82–90. Routledge, London & New York.

Wichmann, S. (2005). On the power-law distribution of language family sizes. *Journal of Linguistics 41*, 117–131.

PAUL HEGGARTY

BEYOND LEXICOSTATISTICS: HOW TO GET MORE OUT OF 'WORD LIST' COMPARISONS

If lexicostatistics could speak, it might justifiably assert: "reports of my death have been greatly exaggerated". For, glottochronology aside, various facets of Swadesh's basic lexicostatistical 'idea' seem alive and well, in a new breed of modern derivatives. This paper first reviews the dominant trends today, then presents alternative approaches to take quantitative lexical comparison in other new directions. Illustrative case-studies range from subfamilies of Indo-European to two major language families of the New World.

A persistent ambiguity has attended 'lexicostatistics', in that methods that go by this name have variously sought to answer two very different types of inquiry:

- An information-type question of *degree*: *how closely* related are languages A and B (within their *known* family)?
- A yes/no-type question: are languages A and B related or not?

These represent two opposing directions in which lexicostatistical methodology might be refined to extract more mileage out of it, extending its range at either the 'shallow' or the 'deep end' of its applicability. The thrust of recent work has been in the latter direction, seeking to isolate the most reliable signal diagnostic of deep-time relatedness, by excluding 'less stable' meanings (which also simplifies data collection). A raft of recent studies hone their lists down far beyond Swadesh's 200 and then 100 items, to a minimal 'most stable' core of just 55, 40, 35, or even 23 meanings.

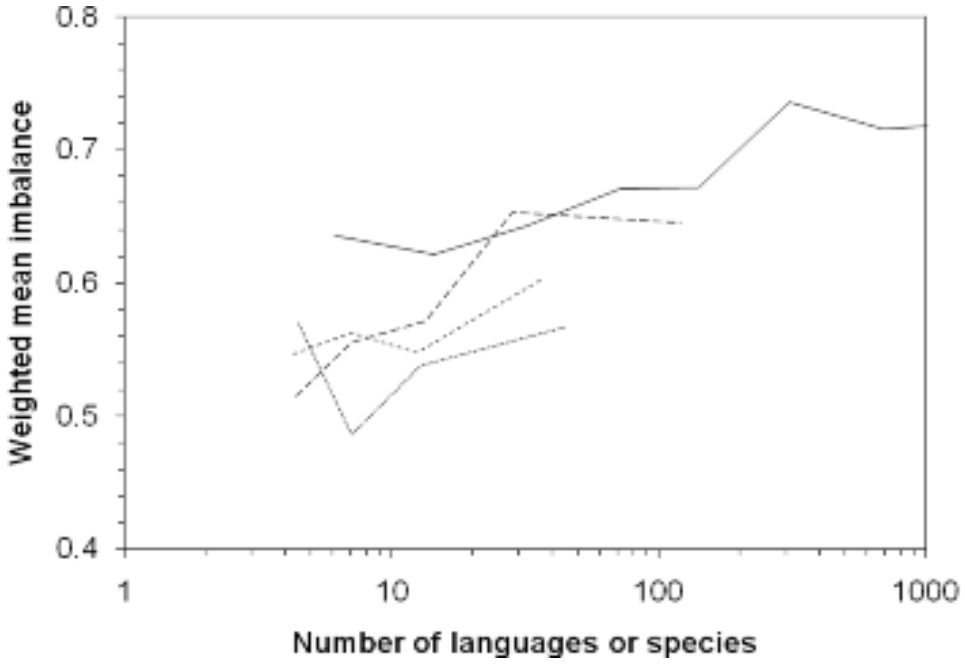
In this paper I argue that we would do well *not* to discard the less stable meanings. Firstly, for 'shallow end' purposes the 'binary straightjacket' of lexicostatistics is already all too blunt a characterisation of *degree* of overlap in lexical semantics; *a fortiori* if we limit the list to the most stable, i.e. least variable, data. In phonetics, a new methodology offers a 'resolution per word' beyond the wildest dreams of lexicostatistics, discriminating even to the accent level. To extend quantification in lexical semantics likewise into dialectology, I propose a radically new method to extract, from each individual meaning, measurements considerably finer-grained than just a binary 'cognate, yes or no?' datum. Again, *less* stable meanings offer richer data. A lesson duly emerges for enthusiasts of phylogenetic analyses too: unrefined lexicostatistical 'encoding' inherently biases results towards more tree-like outputs than the real language data warrant, misrepresenting also the prehistory of speaker populations.

More unexpected is how useful the less stable meanings can prove even at the 'deep end'. Here it is the *contrast* with more stable meanings, and the detailed *gradient* between them, that provide a stark and highly informative perspective on whether given languages are distantly related. By again abandoning certain tenets of traditional lexicostatistics, the fraught case-by-case judgement of cognacy (automated or otherwise) can be sidestepped entirely — especially useful when wordforms are clearly correlate, but data and scholarship are inconclusive or insufficient to confirm whether contact or common origin is the explanation. The Andes provide an ideal test case: the method proposed adduces powerful evidence for the debates on Quechua-Aymara relatedness, and on how far 'borrowability' influences the stability of particular meaning slots.

Eric W. Holman

Do languages originate and become extinct at constant rates?

Morris Swadesh explored the possibility that the words for a given meaning in basic vocabulary are replaced at a stochastically constant rate over time. Other possibilities for constancy are the rate at which languages split to form additional daughter languages, and the rate at which languages become extinct without descendants. The constancy of these rates defines a simple birth and death process, which has already been applied to the origination and extinction of species and used to predict the shape of phylogenetic trees in biology. One standard measure of tree shape is weighted mean imbalance as defined by Purvis et al., 2002, *Journal of Theoretical Biology* 214:99-103. For any binary node in a tree, imbalance takes the value 0 if the two branches are as balanced as possible (with equal numbers of languages or with numbers differing by only one), and it takes the value 1 if the branches are as unbalanced as possible (with one language on one branch and all the rest on the other), and it takes intermediate values for intermediate degrees of balance. According to the birth and death model, weighted mean imbalance has expected value .5, independent of the birth and death rates and the total number of languages on the two branches.



The chart shows weighted mean imbalance as a function of the total number of languages (or species) on the branches, for four sets of phylogenetic trees. The dotted line refers to language trees in the Ethnologue classification, which were constructed manually. The long-dashed line refers to trees constructed by the neighbor-joining algorithm from language similarities produced by the Automatic Similarity Judgment Program (ASJP, see http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm) based on Levenshtein distances in a 40-item Swadesh lexicostatistical list. The short-dashed line also refers to the Ethnologue trees but includes only the languages that are also attested in ASJP. For comparison, the solid line refers to a set of published biological phylogenetic trees, which were constructed automatically from morphological or genetic data. Each curve is significantly above .5 on average, contradicting the birth and death model. The ASJP and biological curves, which have the most binary nodes, also have significantly positive slopes.

The usual explanation for unbalanced biological trees is that species on some branches are better adapted to their environment and therefore have higher origination rates or lower extinction rates. Alternative explanations are explored for languages, which are not thought to differ in adaptation. Average number of speakers per language is no greater on branches with many languages than on branches with few. There is, however, indirect evidence for temporary reductions in rates of origination and extinction associated with the adoption of writing.

Gary Holton, Marian Klamer, František Kratochvil

Lexicostatistical and Comparative method applied to the Papuan languages of Alor-Pantar (Eastern Indonesia): A (re)assessment.

In the study of under-described languages, the lexicostatistical method has proven to be a useful tool for initial genetic classification. However, these preliminary groupings tend to persist long after new data have become available. Ideally, the outcomes of the lexicostatistical method should be reassessed once sufficient data are available to apply the bottom-up approach of the comparative method to refine the outcomes obtained by the preliminary tool.

The present paper attempts to do this using recently available data from seventeen eastern Indonesian languages spoken in the islands of Alor and Pantar. Our comparative data consists of an expanded Swadesh list (200+ items) for each language and of dictionaries for a number of languages.

Based on an examination of possessive prefixes, Capell (1944) originally proposed that the Alor-Pantar languages were related to the West Papuan Phylum languages of North Maluku and the Bird's Head of New Guinea. This hypothesis was later countered by Wurm et al (1975), who classified these languages as members of the putative Trans-New Guinea Phylum. The first attempts to examine internal subgrouping were made by Stokhof (1975), based on lexicostatistical analysis of 117 item Swadesh lists. Stokhof also identified a number of grammatical features with potential for further subgrouping (such as number systems). Based on Stokhof's and Capell's data and their conclusions, Pawley (2001) and Ross (2005) included the Alor and Pantar languages (along with the non-Austronesian languages of Timor) in the large Trans-New Guinea family. Recently, this classification has been questioned by Donohue (2007), who proposes yet another type of affiliation for Timor-Alor-Pantar languages. All of this classification work suffers from a paucity of available data.

By applying bottom-up reconstruction techniques to larger data sets we are able to directly evaluate preliminary classifications based on lexicostatistics, as well as to make direct lexical comparisons with Trans-New Guinea languages of the New Guinea mainland. This work in turn informs our knowledge of prehistoric settlement of Alor-Pantar, complementing emerging genetic and archaeological evidence (cf. Capelli et. al. 1999; Mona et. al. 2007). Klamer (to appear) states that it is unclear "whether the Papuan languages presently spoken in the Alor-Pantar are the result of east-west migrations from the New Guinea highlands between 6,000 and 4,000 BP, or whether they are remnants of an earlier population that had migrated west-east some 20,000 years ago through the Lesser Sunda islands, with a subsequent trek into the highlands of New Guinea." The general consensus is that although the individual languages might be results of later migrations, Papuan populations in Alor and Pantar predate the arrival of the Austronesians. There is archaeological evidence that Austronesians reached neighbouring Timor island by 4,500 BP (cf. Higham 1996:298). The genetic studies suggest a gene flow from Austronesian speaking populations predominantly via maternal line (cf. Handoko 2001), while the paternal line is characterized by Papuan haplogroup (Keyser et.al 2001).

In our paper, we will re-assess the preliminary classifications by Stokhof, Pawley, and Ross and the subgrouping by Donohue in the light of the new data available to us. We will also attempt to resolve the migration route question. Finally, we will elaborate on the benefits of the lexicostatistical and comparative method in language description.

References:

Capell, A. 1944. Peoples and languages of Timor. *Oceania* 14.191-219.

Capelli, Cristian, James F. Wilson, Martin Richards, Michael P. H. Stumpf, Fiona Gratrix, Stephen Oppenheimer, Peter Underhill, Vincenzo L. Pascali, Tsang-Ming Ko and David B. Goldstein. 2001. A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *American Journal of Human Genetics* 68:432-443.

Donohue, M. 2007. The phonological history of the non-Austronesian languages of southern Indonesia. Paper presented at ENUS 5, August 2007. Kupang, Indonesia.

Higham, Charles. 1996. *The Bronze Age of Southeast Asia.* Cambridge: Cambridge University Press.

Klamer, M. To appear. *A grammar of Teiwa*: Under review.

Kayser Manfred, Silke Brauer, Gunter Weiss, Wulf Schiefenhövel, Peter A. Underhill and Mark Stoneking. 2001. Independent histories of human Y chromosomes from Melanesia and Australia. *American Journal of Human Genetetics* 68:173–190.

Mona, Stefano, Mila Tommaseo-Ponzetta, Silke Brauer, Herawati Sudoyo, Sangkot Marzuki and Manfred Kayser. 2007. Patterns of Y-Chromosome Diversity Intersect with the Trans-New Guinea Hypothesis. *Molecular Biology and Evolution* 24(11):2546–2555.

Pawley, A. 2001. The Proto Trans New Guinea obstruents: arguments from top-down reconstruction. In *The boy from Bundaberg : studies in Melanesian linguistics in honour of Tom Dutton*. T. E. Dutton, A. Pawley, M. Ross and D. T. Tryon (ed). 261-300. Canberra: Pacific Linguistics Research School of Pacific and Asian Studies Australian National University.

Ross, M. 2005. Pronouns as a preliminary diagnostics for grouping Papuan languages. In *Papuan pasts: Investigations into the cultural, linguistic and biological history of the Papuan speaking peoples*. A. Pawley and R. Attenborough (ed). 30. Canberra: Pacific Linguistics.

Stokhof, W. A. L. 1975. *Preliminary notes on the Alor and Pantar languages (East Indonesia)*. Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.

Bart Jacobs

A diachronic and comparative analysis of Papiamentu's Swadesh list

This paper is concerned with the origins of Papiamentu and offers a diachronic and comparative analysis of Papiamentu's 100-word Swadesh list (cf. Hancock 1975) to support the claim made by Quint (2000b) that Spanish relexification of an early Upper Guinea Creole variety resulted in what we now know as Papiamentu.

With this purpose, I have roughly divided the Swadesh list up into content words and function words. About function words Muysken & Smith (1990:883) note that "they are normally less susceptible to replacement due to processes of historical change than content words". Consequently, if Papiamentu results from (partial) relexification of an early Upper Guinea Creole variety, we should find evidence for this in the functional catgories. Indeed, according to this prediction, the content words on the Swadesh list are principally of Spanish origin, while the functional elements (e.g. 'that', 'this', 'when', 'where', 'who', 'why', 'you (pl.)') demonstrate remarkable correspondences in form and use with Upper Guinea Creole.

Furthermore, for the benefit of this paper, several recently published early Papiamentu texts (e.g. Conradi 1844, Van Dissel 1865) have been closely studied. These texts allow describing various salient soundchanges that have lead to some of the modern Papiamentu forms found on the Swadesh list, this way providing valuable insight into the relexifcation process responsible for the pronounced Spanish character of modern Papiamentu's content words. I will make a selection of diachronically interesting items that star on the list and, where possible, contrast the modern Papiamentu form with the early form as found in the early texts in order to demonstrate that, if we look closely, we find strong indications of the historical ties between Papiamentu and Upper Guinea Creole not only in the functional categories.

In addition, some Spanish derived Swadesh items in Papiamentu will be compared with their equivalents in Chabacano to argue against a significant role of Old Spanish in Papiamentu's formation.

In its totality, then, the paper aims to present a rich collection of 'Swadesh-list-related' observations that are of interest to the origins of Papiamentu in general and its relationships with Upper Guinea Creole in particular.

Filippo Petroni and Maurizio Serva

Indo-European and Austronesian trees reconstruction

Languages evolve in time according to a process in which reproduction, mutation and extinction are all possible. This is very similar to haploid evolution for asexual organisms or for mtDNA of complex ones. Exploiting this similarity, it is possible, in principle, to verify hypothesis concerning the relationship among languages and to reconstruct their family tree. The key point is the definition of the distance among pairs of languages in analogy with the genetic distance among pairs of organisms. Distances can be evaluated comparing grammar and/or vocabulary but while it is difficult, if not impossible, to quantify grammar distance, it is possible to measure a distance from vocabulary differences. The method used by glottochronology, computes distances from the percentage of shared ``cognates'' which are words with a common historical origin. The weak point of this method is that subjective judgment plays a relevant role. Here we define the distance of two languages by considering a renormalized edit distance among words with same meaning and averaging on the two hundred words contained in a Swadesh list. In our approach the vocabulary of a language is the analogous of DNA for organisms. The advantage is that we avoid subjectivity and, furthermore, reproducibility of results is granted. We apply our method to the Indo-European and the Austronesian group considering, in both cases, fifty different languages. The two trees obtained are, for many aspects, similar to those found by glottochronologists with some important differences concerning the position of few languages. In order to support these different results we separately analyze the structure of distances of these languages with respect to all the others.

Vladimir Polyakov

NEW APPROACHES TO LANGUAGE SIMILARITY MEASURES

October 13, 2008

Similarity measures are used in phylogenetic calculations with the aim of establishing genetic relationships between languages. Recently it has been established (Polyakov & Solovyev, Wichmann et al.) that genetic relationships can be reflected when using measures based on typological data. The noise one gets when using WALS, however, and the influence of areal contact on the performance of Jazyki Mira/Languages of the World (JM), appear to have significant impact on the outcome of the calculations. Therefore, an important goal is to create a measure where the influence of areal contact is minimal.

A test set of 48 languages (hereafter "A. A. Kibrik's set") has been proposed by the "Languages of the World" research group at the Institute of Linguistics of the Russian Academy of Sciences in 2005. A method of estimating the quality of a similarity measure has also been proposed by author, based on ranking the languages of each of the main eight families of the set in accordance with the "prototype language" of each family. After the measure is calculated, the languages are sorted eight times based on their closeness to each of the prototype languages. The quality of the measure is estimated as follows:

$$K = (K_1+K_2+K_3+K_4+K_5+K_6+K_7+K_8)/8,$$

where $Ki = Np_i/Ng_i$,

Np_i - the number of related languages which have been placed after the prototype.

Ng_i - the number of related languages in each group.

The tables with measures can be found at www.dblang2008.narod.ru.

While testing the database, a lot of work has been put into choosing the best possible similarity measure and researching the influence of different factors on its quality. Among these factors are: the types of features, their frequencies, their hierarchical positions in the feature set, and also the contribution of various aspects of how the language has been described. It has been found that the best quality values are obtained when using a measure which is simply the number of all conterminous features without any restrictions on their frequency, hierarchical position or the section to which they belong. In this case, full correspondence with the traditional genetic classification is obtained for two grups (Uralic and Turkic) and the quality estimation K is 0.67. For any other combinations of measures, the results were worse. Also, calculating similarities based on only one of the sections of the feature set always gives worse results than using the entire set.

A new set of 39 languages based on overlap of WALS, the Automatic Similarity Judgement Program (ASJP) project, and JM has been offered by Valery Solovyev and specified by Søren Wichmann in 2007. It can be used not

only for estimating the performance of different metrics on JM's data, but also for comparing the genetic trees received from three different sources. Sample trees obtained from the three databases can be ranked as follows: the ASJP tree is the most reliable, the JM tree is second and the WALS tree is the least reliable (the trees can be found at www.dblang2008.narod.ru).

More subtle techinques of estimating the quality of a measure have also been proposed. After calculating the metric the languages are sorted 39 times based on their similarity to each of them. The quality is estimated as follows:

$$K = \sum(K_i)/39,\ i = i \ldots 39$$

$$Ki = Np_i/Ng_i,$$

where:

where Np_i is the number of related languages placed after each language;

and Ng_i is the number of related languages in each group.

Different methods of comparing the resulting trees also exist. In this case the quality of similarity measure is the edit distance.

New heuristics which would allow for improving the quality of similarity measures have been investigated recently. For example, restricting the features based on their frequencies increases the measure's quality up to 0.697 and restricting them based on the sections of the feature set

yields an increase to 0.76. Both results have been obtained on A. A. Kibrik's set.

In the future it is planned to include factors such as stability and genetic markers in the calculation. It appears more preferable to use the second set because it allows for comparison with other resources such as WALS and ASJP.

Valery Solovyev

Is Grammochronology Possible?

Up to now, only lexical, but not grammatical information has been used for determining the age of language groups. This can generally be explained by two reasons. The first one is the lack of grammatical descriptions of various languages in a systematic and standartized fashion which would allow for statistical methods to be applied. The second reason is the widespread opinion that grammatical features change very irregularly across different languages.

The first of these problems can be considered solved with the appearance of large typological databases like WALS and *Jazyki Mira* [1]. As for the second, the opinion that grammars change irregularly is mostly a subjective statement and requires verification.

In this report the results of several preliminary studies based on the *Jazyki Mira* database are presented. The most important goal at the current stage is to develop a methodology for research in this direction.

We use the standard Hamming distance as a measure of differences between languages, i. e. the number of features whose values are different for the compared languages. If the speed with which grammatical features change is significantly different for various languages, then the distances between them will vary substantially even for languges with the same divergence date.

The average distances between languages among 9 different genetic groups of approximately the same age of 2-3 thousand years are as follows: Indo-Iranian — 242, Italic — 195, Celtic — 215, Germanic — 226, Balto-Slavic — 234, Finno-Ugric — 234, Turkic — 193, Mongolian — 158, Tunguso-Manchurian — 177. The data show that while there is a difference between the distances, it is not as large as to exceed 50%.

The speed with which the lexicon of different languages changes is also not entirely regular. Calculations based on Starostin's adjusted formula [2] show that 4 to 6 words from the 100-item Swadesh list change in different languages during a period of 1000 years. This means that the variation here is also limited to no more than 50%.

Therefore, it would seem that grammatical changes occur at a similar rate in different languages. One could probably receive a better result by counting not all of the grammatical features contained in a database, but only a subset of the most stable ones. Preliminary data on the stability of features for WALS and *Jazyki Mira* can be found in [3,4].

An important part of Swadesh's approach was the notion that the rate of temporal change is static. Sergey Starostin has later adjusted this proposition [2] by introducing a formula with a nonlinear dependence between the number of lexical changes and time. The dependence of the speed of grammatical change is probably more complex and determining it is a task to be accomplished in the future.

The main difference between lexicon and grammar is that while the number of words is virtually limitless, the number of grammatical features is relatively low: the *Jazyki Mira* database contains 3821 such features. Therfore, a language evolving over time in the limit space of grammatical features would inevitably return to its earlier state — these are so-called *back mutations*, which are almost impossible in lexical evolution [5]. A possible mathematical model for this process would perhaps be the movement of points in non-Euclidean space (a hyperboloid in Lobachevsky's geometry). Grammochronology, if established, could be useful in determining the age of language families and macrofamilies on greater time depths than glottochronology.

Preliminary data shows that grammatical change can be considered suitable for determining the times of divergence for languages, although a lot of work still needs to be done for an adequate mathematical model to be created.

References

1. Polyakov V., Solovyev V. Computer Models and Methods in Typology and Comparative Linguistics. Kazan: Kazanskiy Gosudarstvennyy Universitet. 2006.

2. Starostin S. Comparative-historical linguistics and Lexicostatistics. In: Historical linguistics and lexicostatistics. Melbourne, 1999. pp. 3-50.

3. Wichmann S., Holman E.M. Assessing temporal stability for linguistic typological features. http://email.eva.mpg.de/˜wichmann/WichmannHolmanIniSubmit.pdf

4. Polyakov V., Solovyev V., Wichmann S., Belyaev O. Using WALS and Jazyki Mira. (submitted).

5. Nakhley L., Ringe D., Warnow T. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. Language, v.81, pp. 382-420.

Jennifer Sullivan and April McMahon

Phonetic comparison, varieties, and networks: Swadesh's influence lives on here too.

While his eponymous basic vocabulary lists and the study of *language* divergence may be Swadesh's most appreciated legacies, he also took great interest in phonetics (e.g. 1934, 1935, 1937, 1941, 1947), and his techniques of lexicostatistics and glottochronology were as equally applicable to *varieties* as to separate languages (e.g. 1950, 1972: 19-20, 276-7). We demonstrate that phonetic quantification of varieties follows very much in the tradition of Swadesh's own work (see also Embleton (1986, 2000) and Heggarty (2000)). First, we draw these strands of Swadesh's work together, from his work on the vowels of English varieties (1947), to his "Mesh Principle" which captures more complex patterns of variety and language relationships than a simple bifurcating tree (1972: 285-92).

Swadesh's views on phonetic similarity in varieties are couched within the older isogloss system (e.g. 1972: 16). In the second part of our paper, we use a more recent phonetic feature methods with a subset of Germanic/English varieties from data in McMahon et al (2005-07). Our results have identified the great need for Swadesh's "Mesh Principle" to display the complexity of the relationships between varieties adequately. For example, though Standard American English and RP always achieve the highest percentage similarity scores across the different methods, the subset of words in which the rhoticity contrast (see Swadesh 1947) between these varieties is exposed behaves differently. Also, a subset of words with particularly retentive pronunciations pulls the Buckie variety away from Standard Scottish English and more towards a different language, High German. Such complexities are lost within methods which assign a single aggregate score of similarity between a pair of varieties. Through the separate use of very simple artificial data, we have demonstrated this and other problems with existing feature methods.

This work leads us to the final part of the paper, in which we attempt to extend to phonetics one proposal of great foresight from Swadesh. Within the context of inferring ancestral relationships, Swadesh outlined ways of assessing whether two languages were more similar than would be expected by chance (1972: 120). Yet outside of ancestry, what does it mean for two *varieties* to be more *phonetically* similar than chance? The percentage similarity scores between two varieties may not be fully interpreted until we can assess them against a baseline chance level. Just as contemporary methods from evolutionary biology can display the type of network models similar to what Swadesh envisioned for varieties in his "Mesh Principle", so too can we adapt techniques from this field (specifically those for testing for a phylogenetic signal) to the problem of inferring a chance level of phonetic similarity. This problem also requires us to remain linguistically grounded, through the incorporation of frequencies, phonetic typology and the lack of independence of phonetic features. We are working on these challenges currently. What we emphasise overall is that Swadesh's influence is palpable, even in domains outside those for which he is best remembered.

References

Embleton, Sheila. 1986. *Statistics in historical linguistics*. Bochum: Brockmeyer.

Embleton, Sheila. 2000. Lexicostatistics/Glottochronology: From Swadesh to Sankoff to Starostin to future horizons. In Colin Renfrew, April McMahon & Larry Trask (eds.). 2000. *Time depth in historical linguistics. Vol. 1*, 143-65. Cambridge: The McDonald Institute for Archaeological Research.

Heggarty, Paul. 2000. Quantifying change over time in phonetics. In Colin Renfrew, April McMahon & Larry Trask (eds.). 2000. *Time depth in historical linguistics. Vol. 2*, 531-562. Cambridge: The McDonald Institute for Archaeological Research.

McMahon, April, Warren Maguire & Paul Heggarty. 2005-07. Sound comparions: Dialect and language comparison and classification by phonetic similarity. http://www.soundcomparisons.com/ (Sep 2008)

Swadesh, Morris. 1935. The phonetics of Chitimacha. *Language* 10. 345-62

Swadesh, Morris. 1935. The vowels of Chicago English. *Language* 11. 148-51.

Swadesh, Morris. 1937. A method for phonetic accuracy and speed. *American Anthropologist* 39. 728-32.

Swadesh, Morris. 1941. Observation of pattern impact on the phonetics of bilinguals. In Leslie Spier, A. Irving Hallowell & Stanley S. Newman (eds.). *Language, Culture and Personality: Essays in Memory of Edward Sapir*, 37-45. Menasha, WI: Banta.

Swadesh, Morris. 1947. On the analysis of English syllabics. *Language* 23. 137-50.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 21. 121-37.

Swadesh, Morris (ed. Joel Sherzer). 1972. *The origin and diversification of language*. London: Routledge & Kegan Paul.

Measuring the Borrowability of Word Meanings

Uri Tadmor & Martin Haspelmath

This paper presents some of the initial results of the Loanword Typology Project, a large-scale international research project on lexical borrowing. This collaborative project involves several dozen scholars who work on languages representing the geographical, typological, and genealogical diversity of the world's languages. Each contributor was asked to compile an extensive lexical database based on a fixed list of over 1,400 meanings, and the individual databases were then integrated into one consolidated database.

Unlike word lists traditionally used for lexical comparison and analysis, our database allows for an unlimited number of words to be linked to a single meaning and conversely for an unlimited number of meanings to be linked to one word. Moreover, in addition to the word form itself, a wealth of other information is provided for each lexical item, such as morphological structure, age, and loanword status (ranging from 'no evidence for borrowing' to 'clearly borrowed').

One of the major results of the project is a list of all the meanings in the database ranked by how often the counterparts of each meaning are represented by loanwords. For ease of presentation this short conference paper, we will focus mainly on the first 20 least borrowable items on the list.

The list includes seven verbal meanings compared to only four nominal meanings, confirming a commonly made yet hitherto unproven claim that nouns are more borrowable than verbs. The four least borrowable verbal meanings represent semantically broad, typically polysemous verbs: 'stand', 'make', 'go', 'carry'. The next three are basic bodily functions: 'eat', 'hear', 'suck'. The three least borrowable nominal meanings are body parts ('mouth', 'nose', udder'), followed by a plant part ('root'). There are no man-made objects on the short list, the least borrowable noun in this category being 'house' (number 58). All the other items in the top 100 are culture-free. The short list of least borrowable meanings also includes two adjectives ('sharp', 'thick') and seven grammatical or deictic meanings. The fact that more than a third of top 20 least borrowable meanings are grammatical/deictic (typically represented by function words in the individual languages), despite the very low proportion of such meanings in languages' vocabularies overall, confirms a long-held yet hitherto unproven claim that function words are more resistant to borrowing than content words. These seven items include the pronominal meanings 'I', 'you (singular)', and 'he/she/it', the demonstratives 'this' and 'that', plus 'in' and 'yesterday'.

The results of our study are thus not particularly surprising, but they provide a solid empirical basis for what has so far only been suspected. They also allow us to go significantly beyond intuitive definitions of hard-to-borrow meanings such as those underlying the Swadesh list. While the words corresponding to the 207 meanings on this list have a 15% chance of being loanwords in the languages of our sample, the 200 least borrowable meanings on our list have loanword counterparts in only 5% of the cases. Thus, historical linguists who are interested in the most stable meanings now have a serious alternative to the Swadesh list.

Mihail Vasilyev

Glottochronology and Lexicostatistics
Starostin's method: Past and Present Perspectives

Glottochronology, developed by Morris Swadesh in the 1950s, many times served as a subject to a sharp criticism. The series of critical works ([Fodor 1961], [Bergsland and Vogt 1962], [Chrétien 1962] etc.) had brought into question both its basic assumptions (Swadesh's postulates) and mathematical apparatus till it seemed to be completely discredited. Since then a lot of attempts to modify glottochronology have been made. One of them, proposed by the Russian linguist Sergei Starostin, gradually developed into a holistic approach towards not only glottochronology but also lexicostatistics in whole.

In his works S. Starostin detects two chief reasons, explaining the inadequacy of the glottochronological method in most cases:

1. Regarding of borrowings, contained in basic lists (BL), as lexical replacements (or as cognates) in glottochronological calculations (thus exclusion of eleven Danish, three Swedish and two German loans from Riksmal's basic list convincingly resolves Bergsland and Vogt's controversy).

2. Incorrectness of the analogy between radioactive decay and lexical replacement. Starostin reconsiders the 3rd and the 4th Swadesh's postulates by bringing forward the following improvements:

a) A word in contrast to a neutron can become 'older' and the probability of its retention in BL diminishes with the course of time. Thereby the rate of replacement is not a constant, but increases in direct proportion to time.

b) Different items of the BL are not homogeneous by their stability and have different retention rates. Therefore words should be replaced in turn, beginning with the least stable and going on to the more stable. This causes the average rate of divergence to slow-down, as the most stable items progressively dominate in the wordlist.

As shown in Starostin's works, these improvements made it possible to obtain much more reliable datings than those of classical glottochronology. On the other hand, S. Starostin himself pointed out several shortcomings of the method such as:

1. A contradictory nature of the introduced improvements, which is hardly can be interpreted according to the present conception of language development.

2. Objective problems with synonyms and loanwords in compiling the wordlist.

3. Impossibility of any statistical procedure both in glottochronology and lexicostatistics.

In this paper a thorough examination of Starostin's method will be given. On its basis, we will try to reveal the main reasons of the imperfections, mentioned above and several others. After this a new approach (alternative to 'root glottochronology') towards the modeling of lexical processes in comparative-historical linguistics will be proposed and considered.

Computational methods for inferring evolutionary histories of languages

Tandy Warnow

Languages evolve through what is called "genetic descent", but also through lateral transfer, and distinguishing between the two can be difficult.
In this talk I will describe the work that our group is doing modelling language evolution so as to be able to estimate exchange between languages in contact, and the methods we have developed for inferring evolutionary histories including borrowing.
I will also describe our analysis of Indo-European using our new methods.

State of the art of the ASJP project

Søren Wichmann (MPI-EVA & Leiden University) and the ASJP Consortium

Since early 2007 a group of scholars, including myself, have pursued the idea of classifying the world's languages by a computer-automated lexicostatistical method (the Automated Similarity Judgment Program or ASJP). We have built up a database which currently contains data for around 2500 languages, 10% of which are standard 100-item Swadesh list and 90% of which are 40-item subsets of this list, representing the most stable items. The data have allowed for various statistical tests, including testing the basic premise of lexicostatistics and glottochronology that words change at a regular rate. Evidence from many languages have been brought forward against this assumption, but has tended to be anecdotal in nature. A systematic test across languages shows that within a tolerable margin of error the assumption is actually correct. This justifies pursuing both lexicostatistics and glottochronology. We will show some results for the world's languages in these regards, and in addition demonstrate a method for identifying areas of high genealogical diversity within families, a method which may be used as input to the inference of homelands.