Authors' final version (post-review, pre-editing). Please refer to the published version, as changes were made in this one: Lingua Posnaniensis 58:2 (2016).

Transitivizing-detransitivizing typology and language family history

Riho Grünthal (University of Helsinki)
Johanna Nichols (University of California, Berkeley)

Abstract

The transitivizing/detransitivizing typology of Nichols, Peterson, Barnes 2004 also proves useful to historical linguistics. We focus on language families of northern Eurasia, chiefly the three oldest families (Indo-European, Uralic, Nakh-Daghestanian), some of their daughter branches aged about 2000-3000 years, and one younger family for which we have data on enough daughters to support a family phylogeny (Tungusic). We use the 18-pair wordlist of Nichols et al. 2004, which typologizes each pair of verbs depending on which of the two is derived. We make some improvements in the coding of grammatical properties and the typologization of pairs. NeighborNet trees based on this information reveal family-wide linguistic geography and areal trends. Adding minimal information about the cognacy or non-cognacy of the roots of the wordlist items produces NeighborNet trees which approximate well the known phylogeny of the family. Thus very small closed datasets, collected originally for typology, yield rich information about language family history – strikingly, a mere 18 verbs (9 pairs), coded for morphological type and cognacy, yield a very good genealogical tree – while historical methods have also improved the typology.

Keywords: phylogeny, Uralic, Slavic, causative, transitive

Transitivizing-detransitivizing typology and language family history[1]

1. Introduction

The basic valence orientation typology of Nichols, Peterson, and Barnes 2004 uses a fixed wordlist of 18 verb pairs to typologize languages by their preferred realization of what is often known as the causative alternation.[2] The present paper shows that the typology is also useful for historical linguistics in providing grammatical characters to use in phylogeny, complementing existing work based on wordlists and shared phonological innovations.

The causative alternation is the relationship between verbs in pairs such as 'fear' and 'frighten, scare' or '(come to a) boil' and '(bring to a) boil', where the second member is semantically the causative of the first. Henceforth we will use the abbreviated terms *non-causative* and *causative* for the two verb types: e.g. 'fear' is the non-causative and 'scare, frighten' the causative in their pair. We call each such pair of verbs a *causative pair*. The two verbs differ in semantics ('scare, frighten' adds causation by an agent or force), argument structure (the agent or force is the A of the causative verb, and the S or A of the non-causative becomes the causee of the causative verb), valence or argument coding (the causative verb is usually transitive: the new A is usually its subject and the causee is usually an object), and often the morphology of the verb (the causative verb may have a causative affix or other marker of derived transitivity, and/or the non-causative one may have a marker of derived intransitivity). Note that, in using the convenient shorthand terms *causative* and *non-causative* for the individual verbs and *causative pair* for the set we do not imply that the causative verb always has a morpheme called *causative* in descriptions of the language: there is no necessary correlation between the morphology and the causative vs. non-causative status of the verb.

In fact it is the correlation of formal marking to position in causative pairs that is the basis for our typology. The causative may be derived from the non-causative, or the non-causative may be derived from the causative, or both are derived, or neither is, or the pair is suppletive, etc. Table 1 shows some examples. The 2004 typology sought to classify each verb pair for a single gross type: causativizing, decausativizing, etc. Languages are then typologized based on the preferred or dominant type per language: whichever type represents the plurality among the pairs (or whichever type exceeds the sample mean frequency by one standard deviation) is the dominant type for that language. For instance, causativization is dominant in Estonian, decausativization in Russian, and suppletion and ambitransitivity vie for dominance in English. Nichols et al. 2004 show that language families and branches are often fairly consistent as to dominant type, and Nichols 2014 shows that NeighborNet trees based on these and other derivational pairings can reflect language family history to various extents.

---

[1] Some of the research was funded by NSF 9222294, and work on Uralic is being supported by the Kone Foundation.

[2] The wordlist approach and the focus on the causative alternation were inspired by Nedjalkov 1969, where verbs are typologized. Nichols 1982 proposed a whole-language typology.

Table 1. Examples of some derivational pairing types in causative pairs. Relevant derivational morphology is boldface. The essential typological point is whether the boldfaced morphology is in the non-causative column, the causative column, both, or neither.

Language	Non-causative 'fear'	Causative 'frighten'	Derivation type
Macedonian	plaše **se**	plaši	Decausativization
Russian	boja-t'-**sja**	puga-t'	Decausativization, suppletion
Polish	bać **się**	przestra**sz**-**y**-ć	Double, suppletion
Estonian	hirmu-ma	hirmu-**ta**-ma	Causativization
English	fear; afraid	scare	Suppletion
English	" "	fright**en**	Causativization[3]
	'learn'	'teach'	
Macedonian	nauči	nauči	Ambitransitive
English	learn	teach	Suppletion

2. Method: Data and survey

For each survey language we looked up the members of the same 18-pair verb list used in Nichols, Peterson, and Barnes 2004 (shown here in the Appendix) and determined the formal type of pairing. For the most part we used bilingual dictionaries, first looking up the items in an English-target language dictionary and then verifying the meanings and grammatical behavior in a target language-English dictionary. (In addition to English we used dictionaries in Russian, German, and Finnish.) Where possible we consulted more than one dictionary, and/or conferred with language specialists or native speakers.

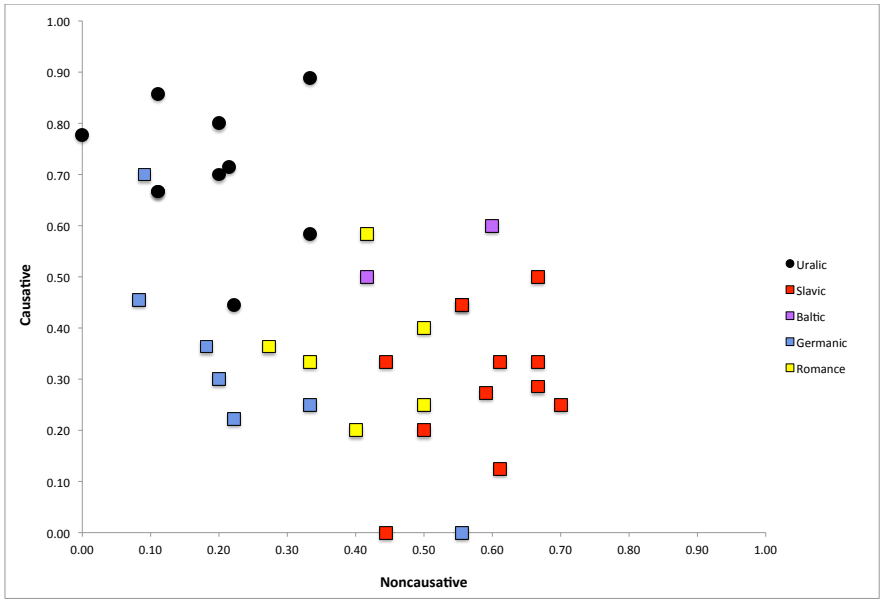
We surveyed several language families of Eurasia as densely as possible: the three old families Indo-European, Uralic, and Nakh-Daghestanian, the younger family Tungusic, and the western Indo-European branches considered separately (Romance, Germanic, Slavic). Previous work using this wordlist has aimed at identifying a single gross type for each pair of verbs (e.g., in Table 1, the pair 'fear'-'scare' is decausativizing in Macedonian, causativizing in Estonian, suppletive in Entlish; etc.). Here, instead of aiming for a single gross type for each pair of verbs, we recode the data to yield three elementary datapoints per pair, based on these questions: (a) is the non-causative derived? (b) is the causative derived? (c) do the two verbs have the same root? The gross types can then be calculated from the elementary types and represent different binnings

[3] English *frighten* and Polish *przestraszyć* are both factitive, i.e. denominal verbs. Here we use the term *causativization* for any pairing where the semantic causative is derived and the non-causative is not, although in these cases the derivational morphology is not generally termed causative.

of them.[4] Table 2 shows the kinds of formal pairings and how they are represented in gross and elementary types.

This new typology makes for clearer plots and other visualizations of the language data. Figure 1 is a plot of the percent of pairs in which the semantic causative is derived against the percent in which the noncausative is derived. Decausativization as a gross type is relatively infrequent, but as an elementary type it is more frequent (it figures in double derivation). Therefore in plots based on gross types the languages are more bunched up toward the left side of the graph, while in Figure 1 they are spread out along both axes. In Figure 1, languages of the same branch or family tend to form fairly discrete clusters.

Figure 1. Percent of verb pairs with derived noncausative x percent with derived causative. Colors identify language families.



[4] This follows the principle of *late aggregation* of Bickel et al. 2016.

Table 2. Derivational types: gross (early aggregation) and elementary (late aggregation). The gross type is used in previous work. "Varies": This gross type can also be reduced to one of the four elementary types. All of the first four can also combine with suppletion.

Gross type:	Elementary type (used in this paper):		
	Non-causative is derived	Causative is derived	Same root
Causativization	No	Yes	Yes
Decausativization	Yes	No	Yes
Double derivation	Yes	Yes	Yes
Ambitransitive	No	No	Yes
Ablaut	Varies[5]	Varies	Yes
Inflectional class change	Varies	Varies	Yes
Auxiliary/LVC[6]	Varies	Varies	Varies
Adjective	Varies	Varies	Yes
Suppletion	Varies	Varies	No

In what follows we review some principles for interpreting NeighborNet diagrams in typologically-based linguistic phylogeny, then present results from the updated coding showing that purely typological data reflects non-descent language family history fairly well (contacts, areality, linguistic geography), and typology plus cognacy of the verb roots reflects phylogeny quite well, drawing examples from Slavic, Germanic, Romance, Tungusic, Turkic, and Nakh-Daghestanian, for all of which the geography of expansion is fairly well understood. We then discuss some possible interpretations for Uralic, for which a full cognacy survey remains to be done and for which the geography of expansion is less certain.

3. Method: Reading NeighborNet graphs in phylogeny

When NeighborNet diagrams are used to represent phylogeny, one looks for splits – sets of parallel lines, which indicate partly shared developments. These identify events

[5] Nichols et al. 2004 set up ablaut as a distinct type and as undirected, i.e. neither member is derived, because while ablaut is often a directed derivation in origin for many languages its history is unknown. But Plank and Lahiri 2015 show that Germanic ablaut can be synchronically shown to be directed. Therefore we hope, optimistically, to be able to determine a direction of derivation for all cases of ablaut. This is for the long future, however, and for the time being an additional datapoint for ablaut and similar alternations needs to stay in the database.

[6] LVC = light verb construction. For these the auxiliary or light verb and the lexical verb are separately coded. (Most often the lexical verb is ambitransitive, e.g. a nominal or nonfinite form that is identical for both members, and the light verb is derived and/or suppletive.)

and bisect the graph into clades or clade-like groups.[7] The closer the lines, the more clade-like the group of languages they lead to; when they are farther apart they indicate non-tree-like evolution (contact effects, etc.). Consider the Slavic family as an example. Figure 2 is an unrooted version of the traditional classification. There are three branches, of which West Slavic and East Slavic are firm clades identified by shared phonological innovations (and Proto-East Slavic is attested as Old Russian), and South Slavic has lexical but no clear phonological innovations and is known to have formed not by divergence from a single ancestor but by convergence of two different immigration streams (and furthermore Slovene, the northernmost South Slavic language, shares early isoglosses with West Slavic and was evidently more closely connected to Czech before the entry of Hungarian and the eastward spread of German divided West from South Slavic: Shevelov 1965:608-613).

Figure 2. The traditional Slavic family tree. Unrooted representation based on Schenker 1995.

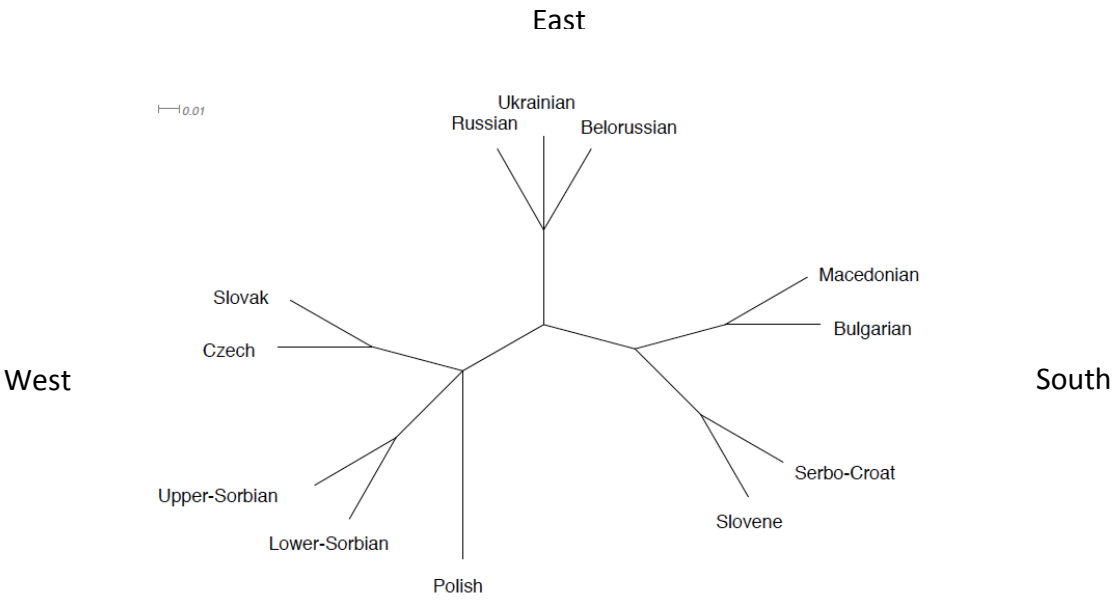


Figure 3 is a NeighborNet[8] diagram drawn with SplitsTree (Huson & Bryant 2006) based on 12 standard sound changes that occurred between late Proto-Slavic and late medieval times; these are shown in Table 3 (for the changes see e.g. Schenker 1995, Carlton 1991, Shevelov 1995). A set of long parallel lines separates West Slavic (Polish through Slovak) from the rest; this is a clear clade with a long period of separate development (the length of the lines shows this), though the lines are far apart, the more distant one reflecting respects in which the northern West Slavic languages pattern

[7] In biology a clade is any descent group. In historical linguistics it is a convenient term for a canonical descent group at any level (subbranch, major branch, highest-level family), i.e. any group with a unique ancestor and one or more diagnostic shared innovations.
[8] For this and other types of computational phylogeny see e.g. Nichols & Warnow 2008.

phonologically with East Slavic. Within West Slavic, the Czechoslovak clade is separate from the rest. Upper and Lower Sorbian do not form a clade, nor do Polish and Sorbian, consistent with traditional scholarship which finds conflicting isoglosses among these. East Slavic is also a true clade, marked by parallel lines quite close together. South Slavic is not a clade: no set of parallel lines separates it either West or East Slavic. It can be described as a cluster, in that the languages are grouped together; this kind of configuration shows that there are some sharings among some of the languages that are probably to be understood as family resemblances but not shared innovations. Bulgarian is drawn toward East Slavic, consistent with phonological similarities between eastern Bulgarian dialects and East Slavic, but otherwise the internal structure of the South Slavic group does not obviously reflect any evolutionary episodes identified by traditional scholarship.

Figure 3. NeighborNet graph based on the 12 sound changes of Table 3.

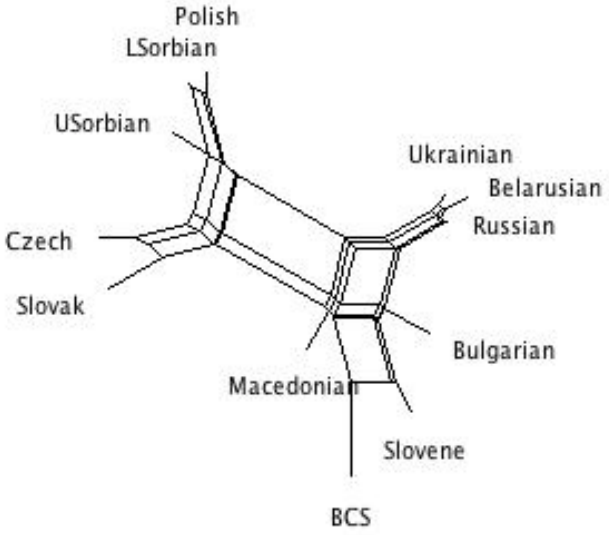


Table 3. 12 sound changes from late Proto-Slavic to late medieval Slavic, in approximate chronological order. Resolution = various adjustments to syllable structure of syllables with internal resonants.

Reflex of *x in the second velar palatalization
*tl, *dl reflexes
*t^{j} reflex
*ORT resolution
*TORT resolution
Reflexes of strong jers *(e/o, e, a,* etc.)
*TuRT resolution
*TRuT resolution
Lenition of *g
Retention/loss of tones

Retention/loss of vowel length
Retention/loss of free stress

In addition to identifying the three branches, it is possible to draw a line on Figure 3 intersecting three widely separated sets of parallel lines: this line goes between Upper Sorbian and Czech, and between Russian and Bulgarian. The two halves of the graph thus bisected are not clades; the fact that the set of parallel lines is much wider than it is long, and that it consists of three separate sets of lines, show that it marks not a clade-defining event (or set of events) but areal factors that have exerted pressure on various sound changes. The areal distinction is between northern and southern Slavic languages, which are recognized as phonological affinity groups in traditional scholarship but not considered clades as there is no innovation or set of innovations that separates north from south.

4. Method: Reading NeighborNet graphs in geography and typology

The north-south split in Slavic discussed for Figure 3 is an example of a split-like configuration that bears a geographical interpretation. A more distinctly geographical shape is shown in Figure 4, a centipede-like shape with a central spine and legs on both sides. This diagram is based on 14 different morphological and lexical contexts in which Proto-Slavic intervocalic *j has been lost or retained, and the languages differ in how many of these preserve or lose *j. A number of different bisections along split lines can be drawn, only one of which is a genuine clade: Czech-Slovak. The centipede configuration is the hallmark of a change or development spreading across a set of languages (Søren Wichmann, p.c.), and this one illustrates the progression of *-j- loss from its beginning in Czech, where intervocalic *-j- is almost always lost, to Russian, where it is never lost. (The set of contexts, the ranking of languages, and the argument that the change began in Czech are from Marvan 1979.) Languages are closer to or farther from the left edge depending on how many losses they share and in which contexts. The 14 contexts form an implicational cline, so the number of contexts in which *j is lost and the specific inventory of contexts largely coincide.

Figure 4. Loss of Slavic intervocalic *-j- in 14 exemplar contexts (Marvan 1979). Belarusian and Macedonian are not shown because they are identical to Ukrainian and Bulgarian respectively.

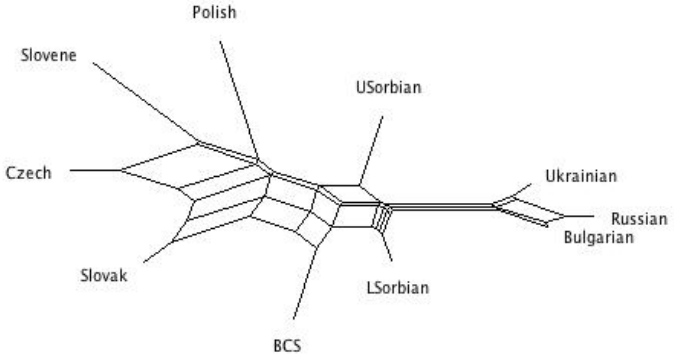
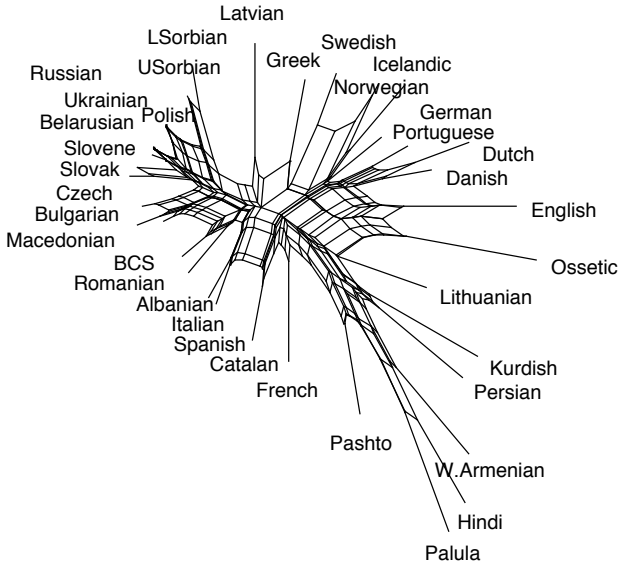


Figure 5 is a purely typological graph, showing the causative alternation in all Indo-European languages in our database. Languages cluster together if they have the same realization of the causative alternation on many of the same verbs. The graph has a centipede-like shape, not because it reflects the progression of some change across the Indo-European languages but because it reflects a gradual decrease in the frequency of decausativization from a peak in Russian at the far left to a low in Hindi and Palula at the lower right. Languages in the left half of the graph use a good deal of decausativization, specifically often reflexivization; those at the right use a good deal of causativization. The Slavic languages all cluster together at the left edge, but do not form a clade (Bosnian-Croatian-Serbian [BCS] is separated from the rest by a number of splits). The other branches cluster together but the clusters include members of other branches. The Romance languages (except for Portuguese) cluster together at the bottom, with Albanian among them and BCS drawn close. The Germanic languages form a protruding cluster at the upper right, and Greek, Ossetic, and Portuguese are among them. At the far right are Indo-Iranian and Armenian, with Lithuanian drawn close. Romance and Germanic cluster on opposite sides of the centipede's midsection; both have intermediate frequencies of decausativization (more in Romance), and Germanic has a fair amount of suppletion. This graph illustrates the essential property of typological NeighborNet graphs: what is important in them is not clades but clusters, created by shared variables and values of variables. Splits, that is sets of parallel lines, when a graph is read as evidence for typology, represent consistencies in the treatment of variables, i.e. the analog to isoglosses or isogloss bundles, but not (or not necessarily) clade-defining events.

Figure 5. The causative alternation in Indo-European languages. Positions of languages reflect which verbs have which realization of the causative alternation. Languages that are close together have the same realization for many of the same verbs. Decausativization dominates at the left end (peaking in Slavic) and causativization at the right (peaking in Hindi and Palula).



5. Results: Typology and cognacy in Slavic

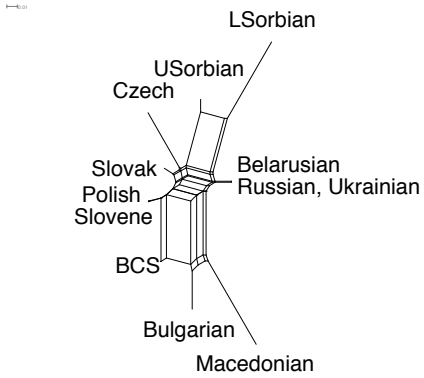
This section uses Slavic data to contrast the results obtained by graphing pure typology, pure cognacy data, and combined typology and cognacy. All use the wordlist for the causative alternation, so we are basing trees on a small number of items: 18 word pairs (or only 9 in some cases), three grammar datapoints for each, and two root cognacy datapoints for each. However, this yields a potentially large total of datapoints: 36 x 3 x 2 = 216. The relatively small investment of time required to survey the 18 verb pairs can thus be quite productive.

Figure 6 shows graphs for the nine inanimate verb pairs. These pairs are mostly high-frequency verbs and fairly resistant to pressures of analogy, pattern copying, etc., and therefore likely to reflect descent history relatively well. Figure 6a, based only on typology, is a poor phylogeny, with East Slavic the only major branch to show up as a clade and Sorbian the only lower-ranking proper clade. It is a centipede graph, with conservative Sorbian and innovative Macedonian at the poles and a prominent north-south split. West Slavic and South Slavic, though not clades, are clusters: going around the graph counterclockwise one encounters all the West Slavic languages together, then all the South Slavic languages, then the East Slavic languages (though for this to be true the statements about West and South Slavic need to assume that Slovene is on the side of its leg toward BCS and Polish is toward Slovak, when in fact they share a single leg and separate positions in their respective clusters cannot actually be ascribed to them). All of this reflects aspects of historical geography and contacts, but not descent. The tree based on cognacy alone (6b) is better, with South and East Slavic emerging as clades and West Slavic as a cluster. Slovene and Czech form a clade, and West Slavic plus Slovene form a clade, a correct reflection of early evolution as discussed above. (6c), combining typology and cognacy, is still better. South and East Slavic are both clades and West Slavic a good cluster. This too is a centipede graph running from conservative Sorbian to innovative Macedonian and Bulgarian, and with a number of east-west splits running

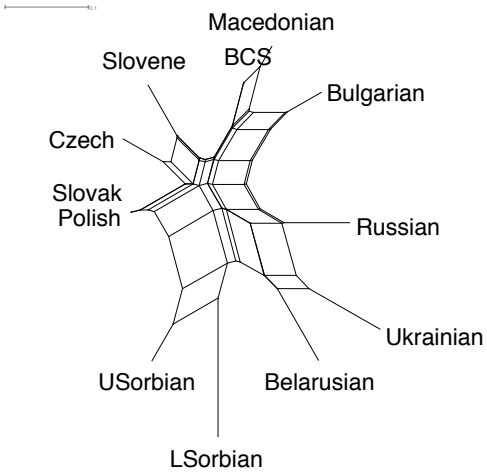
along the vertical axis, reflecting geography. Though Slovene is now firmly in the South Slavic clade, it is drawn toward West Slavic. (6d) combines typology, root cognacy, and cognacy of the valence-related derivational and extensional morphemes, and is a still better tree, with improvements to the internal structure of all three branches: Macedonian and Bulgarian are now more clearly separate from the rest of South Slavic and form a clade, though with a long period of separate evolution; in West Slavic, Czech and Slovak form a clade; and in East Slavic, Russian and Belarusian form a clade. East and South Slavic are clades, and West Slavic comes very close (the magnified inset shows that the dense set of intersecting lines in the middle of the graph contains a small diamond shape that prevents the close double line separating Slovene from Polish from bisecting the graph). Meanwhile the graph is still a centipede running from conservative Lower Sorbian in the north to innovative Macedonian in the south. Czech, Lower Sorbian, and Macedonian have the longest individual branches in the tree, indicating the longest periods of separate evolution and/or the greatest degrees of grammatical and lexical idiosyncrasy. Overall (6d) is a very good fit for both the phylogeny and the geography of contacts in the history of Slavic. Evidently both the typology and the cognacy work together to improve the depiction of evolutionary history.

Figure 6. The causative alternation in Slavic: the 9 animate verb pairs

6a. Typology only.



6b. Root cognacy only.



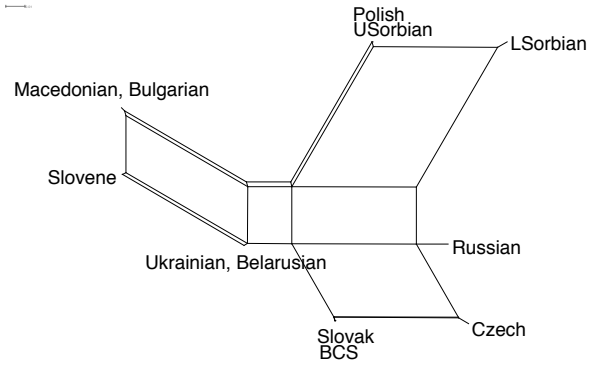
6c. Combined typology and cognacy

6d. Combined typology, root cognacy, and cognacy of derivational morphemes. Inset (at right): magnified view of the central intersection.

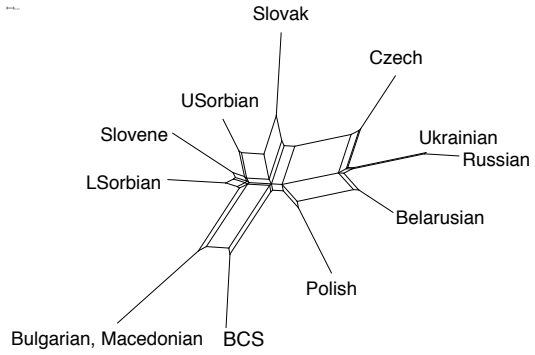
Figure 7 shows the first three graphs for the nine inanimate verb pairs. Cross-linguistically, the inanimate verbs have lower frequency than the animate verbs and are more susceptible to contact influence and universal biases in their evolution. They also seem to have a high frequency of synonyms that are candidates for inclusion in the graph, sometimes making it difficult to pick the best representative of a pair and probably introducing some randomness. (Cognacy of derivational morphemes, analogous to Figure 6c, has not been figured for the inanimate verbs since the incidence of synonymous lexemes and morphological alternatives is high enough to require a separate analysis focused on variation.) 7a is a very poor tree with little interpretable structure, few plausible clusters, and no actual clades. 7b is an improvement, with East Slavic and Czech-Slovak forming clades (albeit Czech-Slovak not at all close, when in reality the pair is quite close-knit). 7c is slightly better, with Sorbian now also forming a clade. In 7b Slovene is out of place in West Slavic; in 7c Slovene is in place but BCS is out of place. All in all the inanimate verbs do not produce trees that reflect either descent or geography very well. Root cognacy makes the greatest contribution to tree quality, and root cognacy and typology do not work together to improve the tree, at least not to the extent we saw with the animate verbs.

Figure 7. The causative alternation in Slavic: the 9 inanimate verb pairs

7a. Typology only.



7b. Root cognacy only.



7c. Combined typology and cognacy

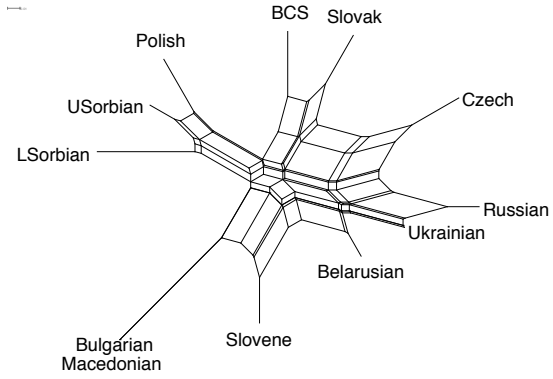
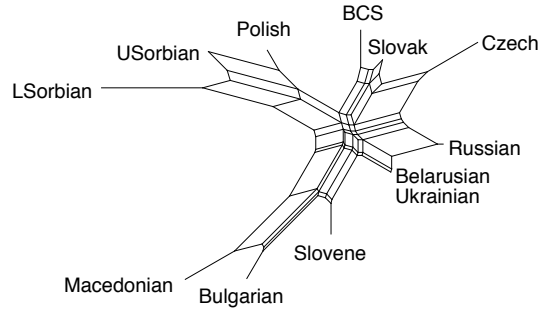


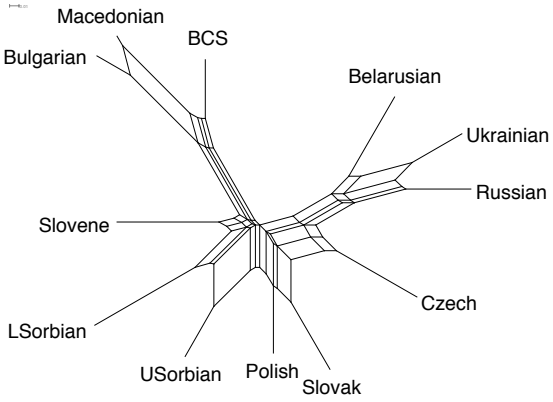
Figure 8 shows the three graphs for all 18 verb pairs. 8a is a weak tree, with Czech-Slovak, East Slavic, and Sorbian correctly shown as clades but BCS out of place disrupting the coherence of both South and West Slavic. 8b is fairly good, with South Slavic and East Slavic clades, West Slavic a cluster, and Czech-Slovak and Sorbian clades. 8c is worse, with South Slavic now a cluster and not a clade and with West Slavic split into two separate pieces. Thus, for all 18 verb pairs as for the inanimate ones, root cognacy seems to make the major contribution to tree quality. It does not work together with typology, as is shown in the decrease in quality from 8b to 8c.

Figure 8. The causative alternation in Slavic: all 18 verb pairs

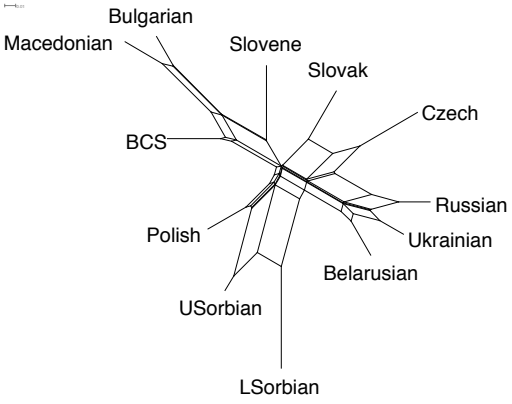
8a. Typology only.



8b. Root cognacy only.



8c. Combined typology and cognacy



To summarize, grammatical information alone gives a reasonable tree shape for the higher-frequency, more stable animate verbs while for the inanimate verbs it does not reflect either descent or geography and contact very well. With the animate verbs, combining different kinds of information – typology, root cognacy, cognacy of derivational elements – improves the graph, and Figure 6d, with all three kinds of

information, gives a remarkably good reflection of evolutionary history, and with only nine pairs of verbs and a total of only 90 datapoints.

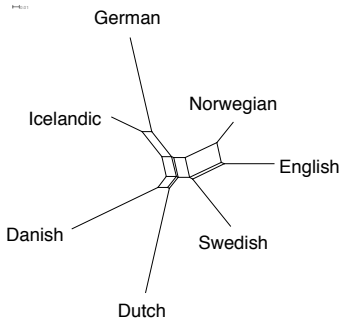
The Slavic graphs have all tended to show East Slavic and sometimes South Slavic as clades but West Slavic as a cluster, in contrast to the graph based on sound changes (Figure 5), in which South Slavic is a cluster and West Slavic a firm clade. This shows that considering grammatical and lexical evidence can make an important contribution to language family history, and it suggests that family trees should not be based only on phonological evidence. On the other hand, the interpretation of the Slavic graphs has required knowledge about history and contacts, which means that one cannot mechanically read evolutionary and contact history off of a tree based on grammatical characters, while in cases like Figures 3 and 4, based on sound changes, graphs can be interpreted more or less on their own.

6. Results: Germanic and others

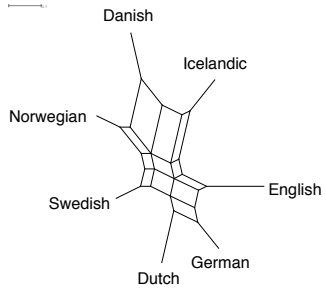
We produced the same three kinds of trees for Germanic, surveying only the animate verbs, with less good results. We surveyed only seven national languages, and the small sample makes the results less firm than for Slavic. In addition, Germanic languages make more use of suppletion than Slavic or Uralic (discussed below), and since in suppletive pairs (like *die* and *kill* or *eat* and *feed*) it is common for both members to be simplex verbs, there are fewer differences of derivational type and grammar contributes less to phylogeny. Figure 9a, based only on the typology of derivations, is a poor tree, with structure not corresponding to phylogeny, contacts, or geography. Figure 9b, based on root cognacy alone, is much better: the North Germanic languages (Icelandic, Danish, Norwegian, Swedish) form a cluster as do the West Germanic languages (English, German, Dutch). The clade-internal structure is less good: we would have expected a Norwegian-Icelandic clade and a closer connection of English to Dutch than to German. The Dutch-German clade, at least, probably reflects preservation of West Germanic vocabulary (while English is lexically more innovative). Figure 9c, based on both typology and cognacy, is little better than 9a though different from it. There is some evidence of geography: the Icelandic-Danish-Norwegian group is more westerly and the Dutch-Swedish-German group more easterly, with English in between (it can belong to either group depending on where one bisects the graph). In all three graphs all languages are at the ends of long individual lines, showing that separate developments have contributed more than shared developments. The main conclusion for now is that the usefulness of grammatical characters in phylogeny depends on the quantity and frequency of the grammatical properties surveyed, and varies from family to family.

Figure 9. The causative alternation in Germanic: the 9 animate verb pairs.

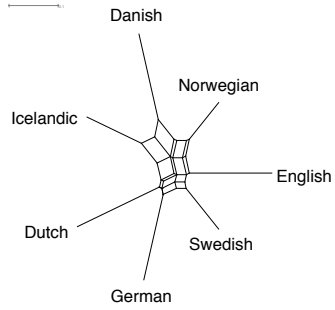
9a. Typology only



9b. Root cognacy only



9c. Typology and root cognacy



Similar results were obtained in trials on Romance and Tungusic (not shown here; some results based on the old coding procedures are shown in Nichols 2014). For Turkic, derivations and cognacy in the verb pairs are so monolithically consistent that there is very little structure in any of the graphs.

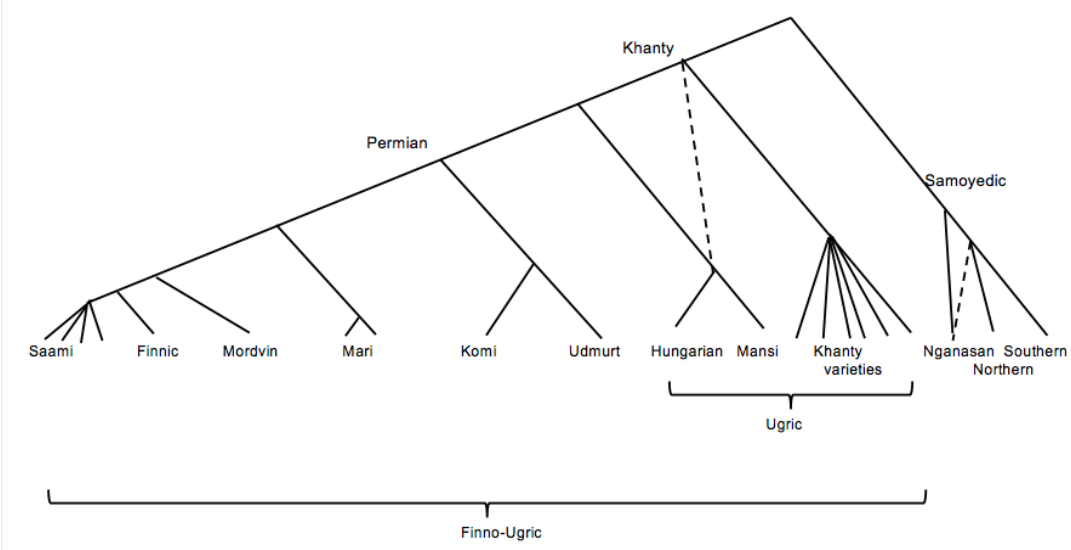
These graphs have shown that typological information can make an important contribution to linguistic phylogeny. When combined with etymological information it yields family trees that approximate actual descent history well, while also revealing evidence of contacts, areal affinities, diffusion of innovations across a speech community, and shared archaisms – and this despite the small wordlist and the weakness of root etymology as evidence. However, separating the phylogenetic and geographical threads requires interpretation based on knowledge of external history and other considerations. Interpreting such trees side by side with ones based on shared sound innovations

improves the analysis; then the typology-cognacy tree can provide valuable confirmation of known or reconstructed history. But even for language families whose history is mostly unknown, typology-cognacy trees can be expected to raise good hypotheses about prehistory. In the long run this may be their most useful contribution to historical linguistics.

7. Discussion: Uralic

Uralic presents a thornier case, as much is unknown about its early history (lacking early written records) and both the higher-level structure of the family tree and the location of the homeland are debated. The family stretches across half of northern Eurasia from western Norway to the Yenisei. The homeland is variously placed anywhere from near the bend of the Volga to just east of the Urals (e.g. Parpola 2012, Janhunen 2009). The traditional family tree is continuously left-branching (west-branching), with an initial bifurcation into Finno-Ugric vs. Samoyedic and subsequent branching of Finno-Ugric as the family spread westward (Figure 10); recently proposed is a star phylogeny with all major branches on an equal footing (Salminen 2001). The star phylogeny is supported by the fact that there are very few sound changes and very few unique lexemes identifying intermediate protolanguages, and even for Proto-Finno-Ugric the sound changes differentiating it from Proto-Uralic are few and fairly minor (Salminen 2001; also Sammallahti 1988, who does not advocate a star phylogeny). Lehtinen et al. 2014, applying computational phylogeny to carefully analyzed lexical data, found results intermediate between a star phylogeny and the traditional tree.

Figure 10. The traditional Uralic family tree. Dotted lines indicate other possible branchings. The tree runs from west (Saami, extending to western Norway) to east (Samoyedic, on the left bank of the Yenisei in Siberia).



We have not yet compiled cognacy data but have done a preliminary survey of the causative alternation in ten Uralic languages. Figure 11 is a graph for the causative alternation in all 18 verb pairs. There is little structure in this tree, and the modest clade-like projections are not correct clades. The closeness of Finnish and Kildin Saami is correct, but Estonian should be even closer to Finnish. Khanty, in the traditional classification, should form a clade with Mansi and Hungarian, but here it clusters with Mansi and forms a false clade with Tundra Nenets. That false clade may have some basis in history, as there is some evidence of early contact or substratal connections between the two. No other projections correspond to membership in traditional branches.

Given the overall east-west distribution of the family one might expect a centipede shape extending from Kildin Saami to Tundra Nenets, but in fact these two languages fall into the same clade-like projection and there is no centipede configuration at all. If there is no proven phylogeny in the tree, however, there is some interesting geography: a clear north-south clustering, with southern languages to the left and northern ones to the right. The southern languages include Erzja, Udmurt, and Mari, which (together with Moksha Mordvin and Komi, not surveyed here) are part of a convergence area also involving Turkic Chuvash and Tatar, so areal effects including Turkic influence may be responsible for much of the southern profile. Neither the Erzja-Udmurt-Mari areal cluster nor the north-south division is clade-like. Overall the tree seems consistent with the geography and what is known about the histories of the languages. However, inclusion of more languages might change the tree considerably.

Figure 11. The causative alternation in Uralic: all 18 verb pairs.

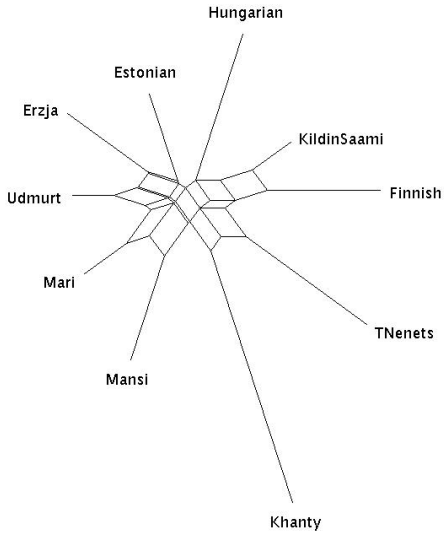
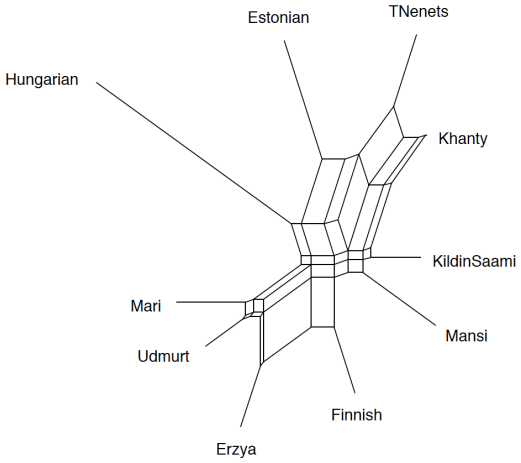


Figure 12 shows the causative alternation only for the nine animate verbs. This graph has more structure, and instead of clusters there are now clade-like structures, none of which correspond well to actual branches. (Finnish and Erzya Mordvin, which form a clade-like pair here, are in the same mid-level branch, but so are Estonian and Kildin

Saami, which are far away in the graph.) They may, however, mark the progress of typological isoglosses across the already-differentiated early Uralic linguistic population. The graph is centipede-like, with eastern languages (Tundra Nenets, Khanty, Mansi, and arguably Hungarian) at the top and western ones (Mari, Udmurt, Erzya, Finnish) at the bottom, but with Estonian and Kildin Saami markedly out of place. There is again a prominent north-south clustering, with southern languages (Estonian, Hungarian, Mari, Udmurt, Erzya) to the left; the membership in southern and northern groups is somewhat different from that for all 18 verbs (Figure 11).

Figure 12. The causative alternation in Uralic: 9 animate verb pairs.



Since some of the value of typologically based NeighborNet graphs lies in their ability to suggest hypotheses, we can raise some possible hypotheses about the early history of Uralic. Despite the obvious east-west distribution of the family and the known relatively recent westward expansion to the westernmost range (e.g. Saarikivi and Lavento 2012, Aikio 2012) vs. the much older presence of Samoyedic in the easternmost parts of its attested range (Janhunen 2012), there is not an obvious signature of east-west spread in the graphs. The mild centipede shape of Figure 12 is the extent of it, and the westernmost languages (Estonian, Finnish, Kildin Saami) are not clustered together at the western end of the diagram but span nearly the whole diagram. This configuration suggests that the basic east-west spread involved migration with little substratum or contact influence, followed by northward expansion mostly by means of language shift, with contact and/or substratal influence. Unless the pre-Uralic linguistic population of northern Eurasia had some overall east-west gradient distribution of typological properties, we would expect northward spreads by shift to have a variety of local typological effects, which is what the graphs seem to suggest.

The north-south areal division visible in both of Figures 11-12 is consistent with the same scenario, and also with Turkic influence in the middle Volga convergence area. A divergent or mixed early history seems to be implied for Khanty, consistent with the conclusions of recent comparative-historical work. In both diagrams Hungarian is out of place relative to either its current location or its origin, as is probably consistent with its complex prehistory and early history (summarized briefly with some lexical evidence by

Abondolo 2009). Estonian is also out of place, an anomalous position for its phylogenetic status and geography, and perhaps to be explained by two developments that have affected only Estonian among the Finnic languages: loss of word-final elements, with consequent morphological restructuring, and strong German influence.

It should be noted that all Uralic languages are predominantly causativizing, unlike the Indo-European languages which range from the heavily decausativizing Russian to the almost entirely causativizing Hindi (Figure 5). In Uralic, therefore, language-to-language differences in causativization must reflect not overall dominant type but shared vs. discrepant patterning of individual verbs. What does it mean if verbs with the same gloss also have the same type of morphology, in languages from entirely different branches and geographical areas? It could reflect a shared archaism, a shared early innovation, or a chance parallel – or it could reflect universals. There are correlations between lexical meaning and propensity to causativize (Haspelmath 1993), and these are probably more visible in the graphs when the overall typological profiles of the languages are similar, as they are for Uralic. If universal preferences are an important factor, then clusters in the graphs may reflect shared susceptibility vs. resistance to universals, and that in turn could reflect contact vs. isolation. Contact presumably favors susceptibility, as languages in contact make choices from wider sets of alternatives than isolated languages, and consistency with universals probably helps some choices emerge as favored. Isolation (in Trudgill's sense of sociolinguistic isolation: 2011) means that the only learners of a language are L1 child learners, who are unfazed by non-consistency with universals.

Many of these uncertainties can be clarified with a detailed survey of cognacy relations, which we are just beginning.

8. Conclusions

Four lessons emerge from these trees. First, the animate verbs – high-frequency, relatively stable words – yield a very good tree for Slavic, while the inanimate ones are more problematic. This suggests that future work using grammatical characters in phylogeny would do well to use high-frequency items. (Nichols 2009 and von Waldenfels & Nichols 2013 got good results from the three Slavic posture verbs 'sit', 'stand', 'lie' in their static, inchoative, and causative forms; these too are high-frequency and stable items.) Second, synonyms (such as English *scare* and *frighten*, or Russian *serdit'sja* and *zlit'sja*, both 'get angry'), where two or more words are contenders for one of the one of the pairs, can introduce randomness into a wordlist procedure that uses one "best" representative per pair. Synonyms seem to be much more in evidence for the inanimate verbs, but their frequency also varies by language (it seems to be more common for Germanic languages than for Slavic), and it is hard to know whether this is linguistic reality or an artifact of lexicography. Our future work will include close synonyms, which we anticipate will improve the usefulness of the inanimate verbs.

Third, as noted above, interpretation of the trees using grammatical characters has required prior knowledge of language history, contacts, migrations, etc. That is, phylogeny can confirm existing knowledge or hypotheses, and the Uralic trees in particular show that it can raise hypotheses about history, contacts, and movements.

Grammatical and cognacy information used together with knowledge of sound changes can move an interpretation from hypothesis to confirmation, or can improve the quality and detail of hypotheses.

Fourth, as we saw with the Germanic trees and the brief discussion of Turkic above, a flawed tree can be quite informative as to how language type shapes language history. We suggested that the poor resolution of the Germanic trees is due in part to the frequency of suppletion in the causative alternation of Germanic, and the lack of usable internal structure in Turkic trees is due to the notable stability and consistency, both lexical and grammatical, of the Turkic family. (Note that Turkic, Germanic, and Slavic – and Romance, which we have examined but not included here – are of roughly similar ages and all have histories of spread, contact, varied contacts, etc., suggesting that it is not sheer passage of time or any aspect of sociolinguistic history that accounts for variation and differentiation of grammar and lexicon.)

These provisos aside, we hope to have shown that typology can be a powerful aid to uncovering language-family history, and the combination of typology, in this case lexical typology, and cognacy information gives a comfortably good approximation to actual phylogeny, which when used together with results of the comparative method on sound changes can provide a very good interpretation of a language family's evolution.

Appendix. The 18 verb pairs used here and in Nichols, Peterson, Barnes 2004.
Animate, inanimate describe prototypical S (of non-causative) or O (of causative).

	Non-causative	Causative
Animate S/O:		
	laugh	make laugh
	die	kill
	sit	seat, have sit
	eat	feed
	learn	teach
	see	show
	be/get angry	anger, make angry
	fear, be afraid	scare, frighten
	hide	hide
Inanimate S/O:		
	boil	boil
	burn	burn
	break	break
	open	open
	be/get dry, dry out	dry, dry out
	straighten (out)	straighten (out)
	hang	hang (up)
	turn over, overturn	turn over, overturn
	fall	drop, let fall

References

Abondolo, Daniel. 2009. Hungarian. Bernard Comrie, ed., *The World's Major Languages (2nd ed.)*, 484-496. London: Routledge.

Aikio, Ante. 2012. An essay on Saami ethnolinguistic prehistory. Grünthal and Petri Kallio, eds., 63-117.

Bickel, Balthasar, Taras Zakharko, Johanna Nichols. 2016. Better data with late aggregation: Autotyp and beyond. Presented at Poznań Linguistics Meeting 2016, Sept. 15.

Carlton, Terence. 1991. *Introduction to the Phonological History of the Slavic Languages*. Columbus: Slavica.

Grünthal, Riho, and Petri Kallio, eds. 2012. *A Linguistic Map of Prehistoric Northern Europe* Helsinki: Société Finno-Ougrienne.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. Bernard Comrie and Maria Polinsky, ed., *Causatives and Transitivity*, 87-120. Amsterdam: Benjamins.

Huson, Daniel H. and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 34:3: 254-267.

Janhunen, Juha. 2012. Etymological and ethnohistorical aspects of the Yenisei. *Studia Etymologica Cracoviensia* 17:67-87.

Janhunen, Juha. 2009. Proto-Uralic – what, where, and when? *The Quasquicentennial of the Finno-Ugrian Society*, 57-78. (SUST 258.) Helsinki: Suomalais-Ugrilainen Seura.

Lehtinen, Jyri, Terhi Honkola, Kalle Korhonen, Kai Syrjänen, Niklas Wahlberg and Outi Vesakoski. 2014. Behind family trees: Secondary connections in Uralic language networks. *Language Dynamics and Change* 4: 189-221.

Marvan, Jiri. 1979. *Prehistoric Slavic contraction*. University Park: Pennsylvania State University Press.

Nedjalkov, V. P. 1969. Nekotorye verojatnostnye universalii v glagol'nom slovoobrazovanii. F. Vardul', ed., *Jazykovye universalii i lingvističeskaja tipologija*, 106-114. Moscow: Nauka.

Nichols, Johanna. 1982. Ingush transitivization and detransitivization. *BLS* 8: 445-462.

Nichols, Johanna. 2009. Expanding character sets for phylogeny: A Slavic test case. In Bethwyn Evans, ed., *Discovering History through Language: Papers in Honor of Malcolm Ross,* 127-151. Canberra: Pacific Linguistics.

Nichols, Johanna. 2014. Derivational paradigms in diachrony and comparison. Martine Robbeets and Walter Bisang, ed., *Paradigm Change in the Transeurasian Languages and Beyond*, 61-88. Amsterdam: Benjamins.

Nichols, Johanna, David A. Peterson and Jonathan Barnes. 2004. Transitivizing and detransitivizing languages. *Linguistic Typology* 8:2: 149-211.

Nichols, Johanna and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2:5.760-820.

Parpola, Asko. 2012. Formation of the Indo-European and Uralic (Finno-Ugric) language families in the light of archaeology: Revised and integrated "total" correlations. Grünthal and Kallio eds., 119-184.

Saarikivi, Janne, and Mika Lavento. 2012. Linguistics and archaeology: A critical view

of an interdisciplinary approach with reference to the prehistory of northern Scandinavia. In *Networks, Interaction, and Emerging Identities in Fennoscandia and Beyond*, ed. Charlotte Damm and Janne Saarikivi. Helsinki: Suomalais-Ugrilainen Seura.

Salminen, Tapani. 2001. The rise of the Finno-Ugric language family. In *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological considerations*, ed. Christian Carpelan, Asko Parpola, and Petteri Koskikallio, 385-396. Helsinki: Suomalais-Ugrilainen Seura.

Sammallahti, Pekka. 1988. Historical phonology of the Uralic languages. In *The Uralic languages*, ed. Denis Sinor, 478-554. Helsinki: Suomalais-Ugrilainen Seura.

Schenker, Alexander M. 1995. *The Dawn of Slavic*. New Haven: Yale University Press.

Shevelov, George Y. 1965. *A Prehistory of Slavic: The Phonological History of Common Slavic*. New York: Columbia University Press.

Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Structure and Complexity*. Oxford: Oxford University Press.

von Waldenfels, Ruprecht, and Johanna Nichols. 2013. Better characters for better phylogenies. Presented at ICHL 21, Oslo.