Day 2: From Words to Trees

Johann-Mattis List (DLCE, MPI-EVA Leipzig)

1 Lexical Motivation

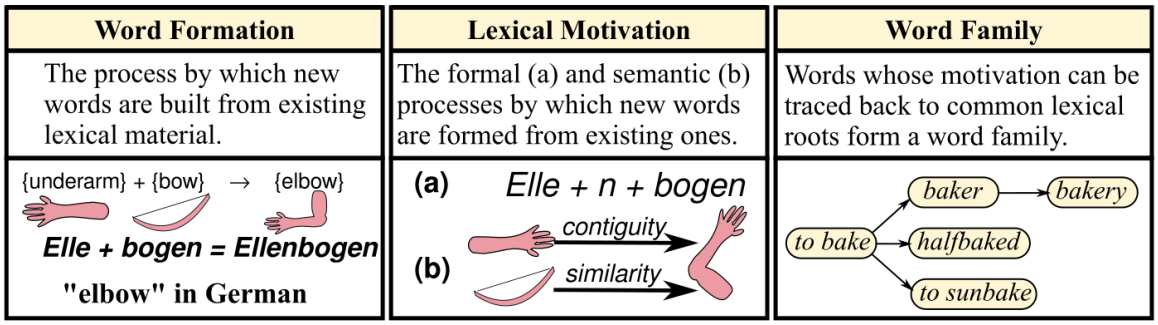
In the first session, we have discussed some major questions regarding cognates, alignments, and sound correspondences. In this session, we will take a rather radical turn and discuss phylogenetic trees and phylogenies in a broader sense. In order to keep the transition smooth, however, we will first look at *lexical motivation* as a major process underlying word formation, which is indispensable in order to understand the major processes of lexical change, which themselves open the doors towards phylogenies and phylogenetic trees.

As we have seen before, in our discussion about alignments and alignability, it can be quite challenging to find examples for regular sound correspondences, not because sound change is not regular, but rather because it has been superseded by numerous morphological processes by which the original shape of word forms has been modified. In order to cope with this problem of drastic information loss when searching for regular sound change processes, scholars have for a long time made use of specific techniques of *internal reconstruction* in order to find the original shape of the word forms which had been changed under the influence of morphological change. These techniques have never been sufficiently formalized, not only, because a formalization might be difficult and language or language-family-specific, but also because the workflows which linguists use often switch back and forth between internal and external reconstructions.

In our group, we have began to work towards a formalization of internal reconstruction by focusing on word formation processes in general and lexical motivation patterns and word families in particular. The results of these efforts have been in part published, but our approaches have evolved with the publications. For this reason, we are still in a stage where we test our methods and try to find a sufficiently large enough number of examples that illustrate how the new techniques can be used. In the following, we will only quickly look at word families, and then show how we try to handle motivation structures with the help of *morpheme glosses*, and how we handle allomorphy with *inline alignments*. I hope that we will be able to publish some more detailed studies presenting our new approaches in the nearer future.

1.1 Word Families

Following Koch (2001), lexical motivation can be defined as summarizing the formal and semantic processes by which new words are formed from existing ones during word formation (Gruaz 2002). Words whose motivation can be traced back to common lexical roots form a *word family* (Hundsnurscher 2002). The following image (taken from a grant application, with the grant to be started in January 2023, which may later be published independently) provides some examples for word formation, lexical motivation, and word families.



Interestingly, scholars do not often talk about word families in historical linguistics. An exception is the field of Sino-Tibetan linguistics, where it enjoys a doubtful reputation, due to the fact that the grouping of words into families is not often done with the help of rigorous principles, see Fellner and Hill 2019). In scholarly practice, however, it is rather the norm than the exception that scholars apply language-specific knowledge about language- or subgroup-internal word formation processes in order to identify those comparanda which substantiate their sound correspondences when comparing words across languages. This internal reconstruction – when carried out rigorously – is often the reason why scholars are convinced of the regularity of sound change, since the "raw" material derived from concrete word forms for concrete terms often does not show regular sound correspondences throughout the whole word without further explanation.

Word families – the result of assembling words in the same language into groups of common descent – are thus central to scholarly practice of historical language comparison, even if there are no common ways to formalize scholars' knowledge on them.

How do compounds fit into the word family schema?

1.2 Morpheme Glosses

When comparing words across languages and inside one and the same language, it is crucial to formally annotate their lexical motivation, that is, the semantic and morphological processes by which they have been derived. As a first attempt to handle these processes – originally tested only on compounds in Burmish languages – Hill and List (2017) introduced *morpheme glosses* as a device to make individual components of words transparent through annotation. Morpheme glosses are similar to *interlinear-glossed text* (Lehmann 2004) but not applied to the sentences in a corpus, but rather to the words in a wordlist or a lexicon. Thus, starting from a morpheme-segmented word form (in our space-segmented representation by which words are modeled as sound sequences), we can annotated individual word parts by providing short glosses which give hints to the original meaning or the grammatical function (inside the complex word). The following table (taken from our grant application) provides some examples of morpheme glosses as represented in the EDICTOR tool (`https://digling.org/edictor`, List 2021a, List et al. 2017).

ID	CONCEPT	TOKENS	MORPHEMES	COGIDS
1794	bow	b o: g + ə n	bow + -:S.m.s/en	11[4] 922[47]
408	elbow	ɛ l + ə + n + b o: g + ə n	underarm + -S.f.s/e + -l/n + bow + -S.m.s/en	11[4] 91[219] 870 922[47] 922[47]
82	rain (noun)	r e: g ə n	rain	190[2]
71	rainbow	r e: g ə n + b o: g + ə n	rain + bow + -:S.m.s/en	190[2] 11[4] 922[47]

After our initial experiments, we successively expanded the idea of using glosses for the annotation of lexical motivation in follow-up studies, trying to show that morpheme glosses are not only apt for the handling of etymological relations between words in the case of compounds but also in the case of derivation (Schweikhard and List 2020). Currently, we are furthermore testing, to which degree these glosses can help us to improve phylogenetic reconstruction and the identification of alignable word parts (Wu and List forthcoming).

Would it be possible to capture universal motivation patterns with the morpheme gloss annotation approach?

1.3 Inline Alignments

So far, we have not fully resolved the problem of handling allomorphy and exceptions in sound correspondence patterns. The problem here is that both allomorphy and (known) exceptions, due to some kinds of irregular sound change, such as, for example, assimilation of frequently used words, analogical processes like contamination, or sandhi phenomena, easily mask the regularity of the correspondences exhibited by the root of a word family. In our formal annotations of linguistic data, we want to capture both the knowledge that certain sounds are unexpected, following our idea of sound laws, but at the same time we want to list them along with the major patterns, showing specifically also what we would *expect* a word to look like if it had been changed by regular processes. A current approach which we are testing on different datasets at the moment are *inline alignments*. An inline alignment is an alignment of a sequence that is represented along with the sequence itself in the same *line* (which is why I call it in-line).

As a very general example, consider German *Eltern* "parents" and its sequential phonetic representation [ɛ l t ɐ n], which we can contrast with its phonological representation [ɛ l t ə ʁ n]. An inline alignment of these two forms, which would take the phonological form as the primary one would look like [ɛ l t Ø/ə ɐ/ʁ n]. So what we do in an inline alignment is we "entangle" both forms in a single sequence in such a way that we list identical segments of the alignment only once, while alternating segments are represented with the help of a dash symbol that separates the "source" of the "target" element. From the perspective of its information, we can generate both the phonological and the phonetic form from this new sequence.

But the principle of representing two sequences in one line does not have to be limited to the representation of phonological and phonetic forms, it can also be used for morphological alternation. Thus [ɛ l t ə ʁ n] can be analyzed as consisting of the comparative of *alt* "old" and the plural ending *-en*, thus referring to "the older (people)". We could represent the original form as [a l t + ə ʁ + ə n], which we can describe with a morpheme gloss as ○LD :comparative :plural, and in an inline alignment as [ɛ/a l t + ə ʁ + Ø/ə n] (this has been outlined in more detail in List 2021b). The advantage of using the analyzed form as the primary form is that we can now directly compare the inline alignment with

other members of the word family *alt* in German. Furthermore, using the root [a l t] as an anchor point, we can also automatically align all kinds of its derivations to this root form and thus provide a multiple alignment of word forms derived from the same root (the use of multiple alignments for language-internal handling of morphology is uncommon, but there are some examples for recent attempts to make use of them, as shown by Beniamine and Guzmán Naranjo 2021 for inflectional morphology).

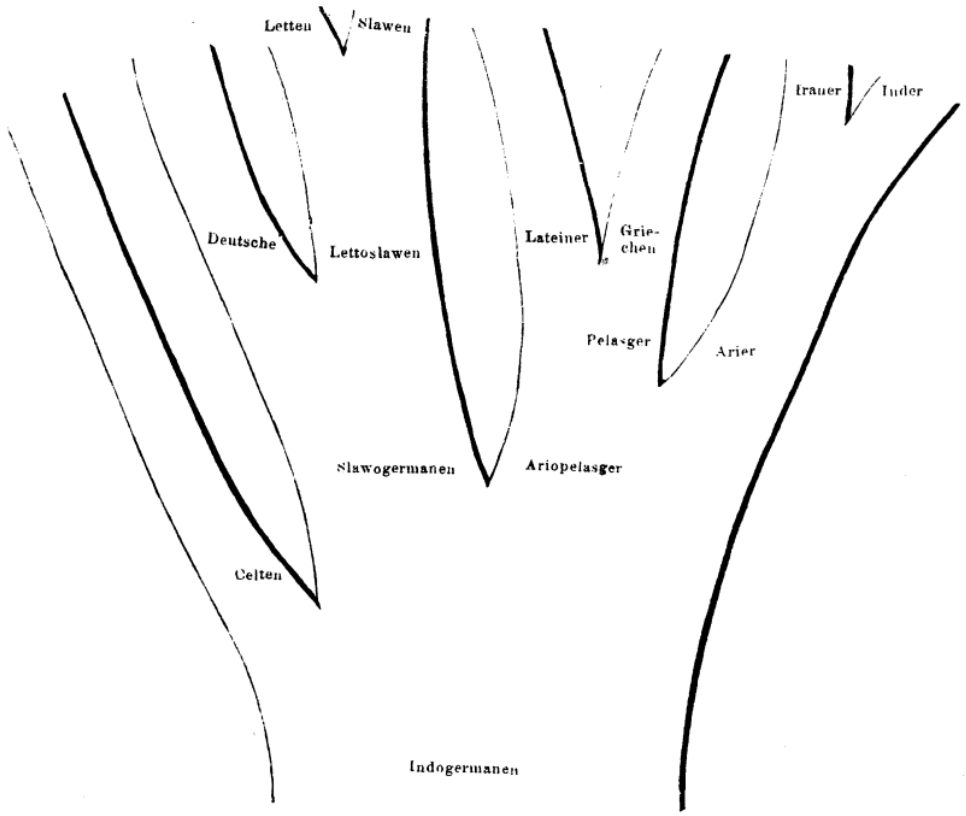
What is the advantage of inline alignments over a listing of all words that belong to one and the same language family, and where are the limits of this approach?

2 Phylogenies

Linguists have known for a long time that languages evolve and that the languages we observe today may stem from common sources which themselves no longer exists. First speculations on the common descent and the tree- or network-like separation of languages can already be found in early studies of the 17th century and thereafter (List et al. 2016). Until the late 18th century, however, the dominant view among scholars in Europe was that all human languages were products of the mythical Confusion of Tongues which prevented the construction of the Tower of Babel (Klein 2004).

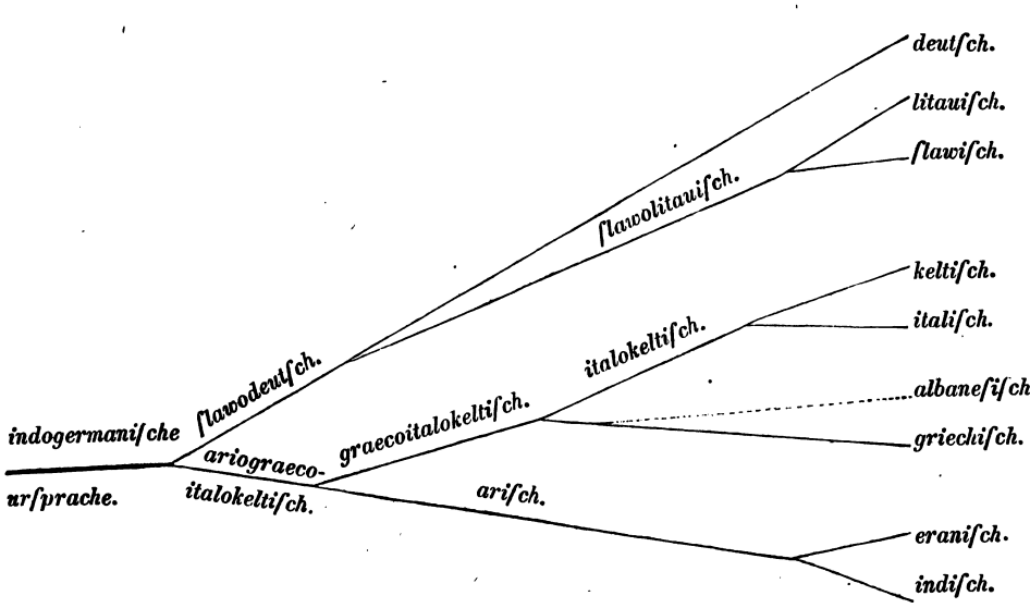
2.1 From Oaks to Tree Diagrams

Unlike modern phylogenetic trees, early linguistic trees were much less formal and systematic, but had the tendency to resemble true trees much more closely. As an example, consider Schleicher's tree from 1853 (Schleicher 1853), which has the appearance of a massive oak with a big trunk. Only later, the family tree visualizations became more schematized, but the interpretation was still far away from being formalized.



As an example for the lack of formalization, consider again a tree by Schleicher, this time from 1861 (Schleicher 1861). While this tree looks much more formalized than the earlier tree from 1853, the description of this tree in the text is interesting, since Schleicher points to branch lengths as representing the supposed time which had elapsed since separation while at the same time emphasizing that the distance between extant languages reflected their synchronic closeness.

> The oldest splits of Indo-European until the development of the funda- mental languages of the language families which constitute the stem of the language [sprachstamm] can be visualized by the following schema. The length of the lines in- dicates the elapsed time, the distance of the lines from each other indicates the degree of relationship. (ibid.: 6f)[1]



While it is difficult to understand the passage by Schleicher completely, it is possible that Schleicher thought of some additional closeness between languages independent of their evolutionary history and tried to mark this in his tree drawing by separating the major subgroups visually from each other in the tree and by placing languages like Albanian and Greek horizontally close to each other while at the same time assigning them a larger divergence time than given for Celtic and Italian. What could such a close placement of languages in a phylogenetic tree reflect from a contemporary perspective on language evolution?

2.2 From Trees to Webs

Not long after Schleicher and some colleagues had propagated their family tree models for the first time, scholars began to contest them. One of the most prominently cited opponents of Schleicher's family trees was Johannes Schmidt (1843-1901), who devoted a complete booklet to contradict Schleicher (Schmidt 1872).

In this study, Schmidt presented concrete data in the form of sets of ho- mologous words ("cognate sets" in linguistic terminology) for the major Indo-European branches. He noted that one could easily find examples for homologs shared exclusively among different possible pairings (Greek vs. Old Indian,

[1] My translation, original text: "Die ältesten teilungen des indogermanischen bis zum entstehen der grundsprachen der den sprachstamm bildenden sprachfamilien laßen sich durch folgendes schema anschaulich machen. Die länge der linien deutet die zeitdauer an, die entfernung derselben von einander den verwantschaftsgrad."

Greek vs. Slavic, Slavic vs. Old Indian) with no residues ("reflexes" in linguistic terminology) in any of the other branches. Based on this finding, Schmidt refuted the family tree hypothesis, arguing that a tree could not explain the observed data.

What Schmidt proposed instead was the rather fuzzy idea of a wave-like expansion of the major branches of the Indo-European languages which contributed to their gradual separation and would explain the specific commonalities between individual pairs which seemed to contradict each other. Unfortunately, Schmidt did not see that the cases he listed could be perfectly explained by the traditional tree model assuming well-known phenomena like differential loss (Geisler and List 2013) or incomplete lineage sorting (Evans et al. 2021, Jacques and List 2019, List et al. 2016). But although his critic was not valid and his alternative model, the "wave theory" (Wellentheorie), as it was called thereafter, did not offer any concrete instructions with respect to the formal modeling of language divergence and spread, many linguists started to present it as a valid alternative to the family tree model.

Nowadays, the wave theory is often presented as some kind of diffusion model in which languages gradually diverge without splitting abruptly. This model of language evolution was already mentioned by Hugo Schuchardt (Schuchardt 1870 [1900]), and family tree models are not capable of modeling the split process in detail. But is it justified to attribute diffusion models to Schmidt's wave theory?

3 Phylogenetic Reconstruction

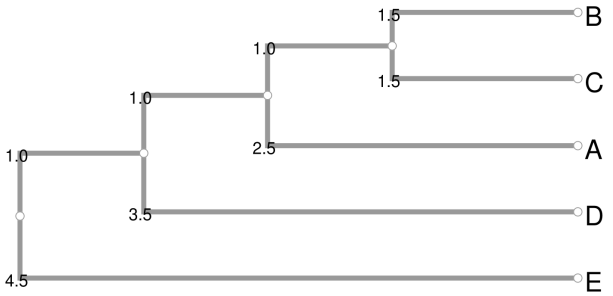
In the following, we will quickly introduce some major concepts of phylogenetic reconstruction as it is practiced in recent approaches to historical language comparison and heavily inspired from evolutionary biology.

3.1 Distance-Based Approaches to Phylogenetic Reconstruction

Distance-based approaches were among the earliest approaches to phylogenetic reconstruction which were proposed by scholars. The basic assumption of distance-based approaches is that one can aggregate certain information about the taxonomic units (languages, species, etc.) in such a way that these aggregated similarities or dissimilarities among the taxonomic units provide enough information to reconstruct a phylogeny of the units in question. While most scholars would contradict this claim nowadays, both in evolutionary biology and in linguistics, distance-based approaches are still useful in order to quickly check a certain dataset, since they are easy to understand and fast and easy to apply.

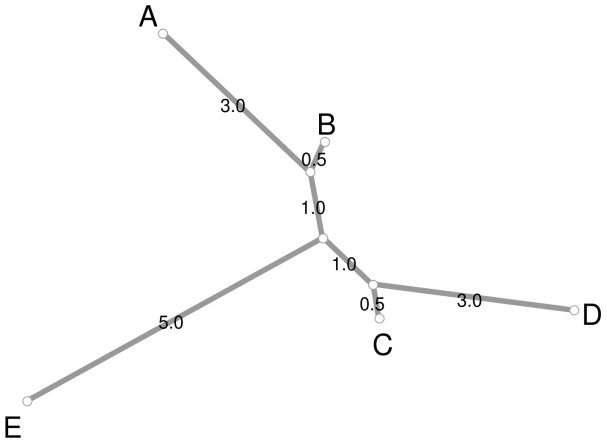
As a first algorithm that can be used to infer phylogenetic trees from the data, there is the algorithm that is nowadays simply called UPGMA by Sokal and Michener (1958). This algorithm is fairly simple, by subsequently merging those languages with each other which show the lowest distance score, and then averaging all distances between the merged languages with the rest, thus creating a new matrix that is then again investigated for the pair with the lowest distance. What is essential about this algorithm is that it yields a rooted tree in which branch lengths are supposed to reflect the true, steady, evolution that occurred. If a distance matrix fulfils the criterion of really reflecting steady evolution (in which change proceeds at the same speed), the distances in the distance matrix are the same as between the branches in the tree. If evolution is not ultra-metric, the UPGMA branch lengths will differ from the distances in the matrix. In such a situation, no rooted tree can display the matrix truthfully. In An unrooted tree, however, can display certain distances, which are called additive. If a matrix shows data which are truly additive, the Neighbor-joining algorithm can be used to find the tree (Saitou and Nei 1987). The difference is illustrated in the following figures.

	A	B	C	D	E
A	0.0	5.0	5.0	7.0	9.0
B	5.0	0.0	3.0	7.0	9.0
C	5.0	3.0	0.0	7.0	9.0
D	7.0	7.0	7.0	0.0	9.0
E	9.0	9.0	9.0	9.0	0.0



(a) *ultrametric distances and corresponding rooted tree*

	A	B	C	D	E
A	0.0	3.5	5.5	8.0	9.0
B	3.5	0.0	3.0	5.5	6.5
C	5.5	3.0	0.0	3.5	6.5
D	8.0	5.5	3.5	0.0	9.0
E	9.0	6.5	6.5	9.0	0.0



(b) *additive distances and corresponding unrooted tree*

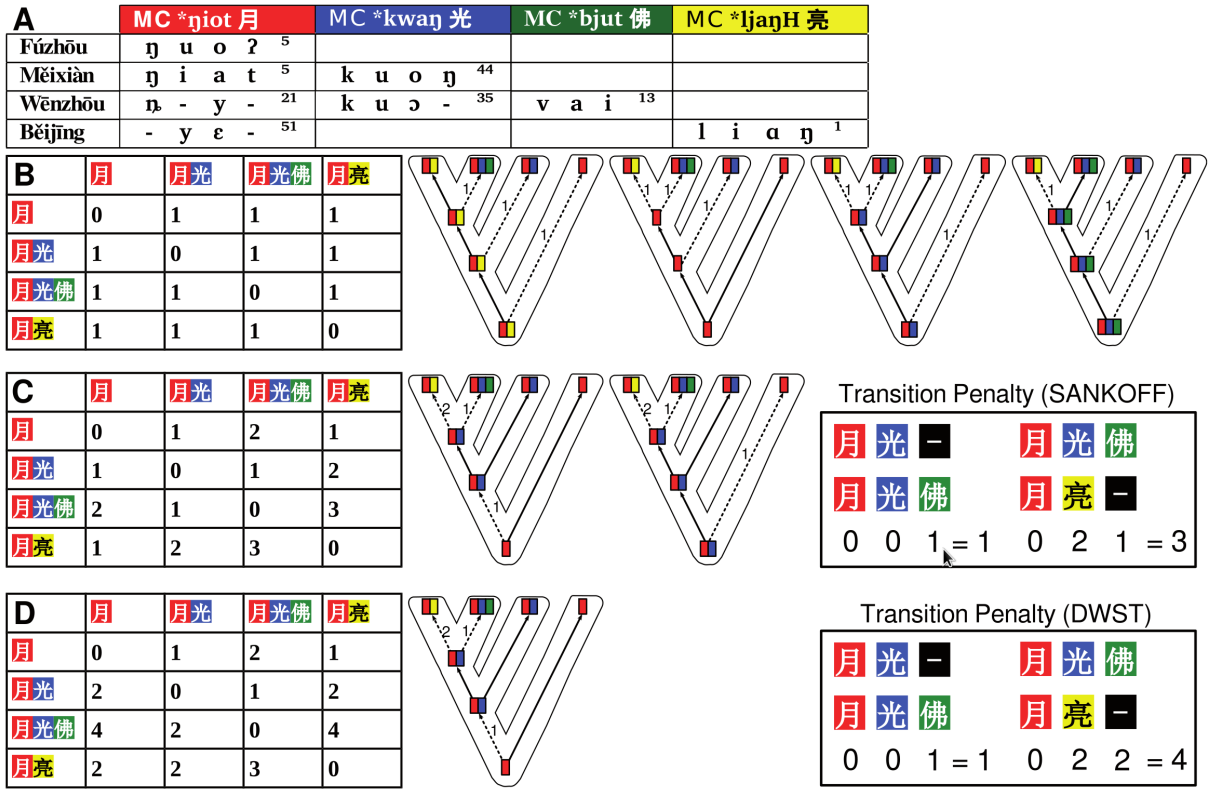
Neighbor-Nets (Bryant and Moulton 2004) show uncertainty in distance-data by representing distance data in the form of a network when data are not additive, but resolving data as a tree, when they are truly additive. Why is it problematic when scholars interpret Neighbor-Nets literally, reflecting concrete evolutionary processes?

3.2 Character-Based Approaches to Phylogenetic Reconstruction

In contrast to distance-based approaches, character-based approaches start from individual features and try to model their evolution along a potential phylogeny. When being given a larger number of features, specific techniques are used in order to infer the phylogeny that explains best how all features evolved under the same phylogeny.

The normal case in which we think about language evolution is by using certain models. Swadesh turned his cognate sets into distance matrices, which means he lost a lot of interesting information on individual processes of lexical replacement, which is a pity, since he had such an interesting model, as we saw last time. Later, in biology, methods were developed to account for the evolution of individual traits, and the most prominent one (at least for a long time) is parsimony or *maximum parsimony*. The idea is that one checks for a certain traits how it evolves along a tree, and while doing so, we penalize certain processes. If the trait does not change, we do not penalize it, but if it turns into something else, we might decide to give it a penalty of 1, 2 or any other value. We could even go so far as to provide specific weights for individual processes. We represent the processes by assuming a finite set of *character states*, and a corresponding *step matrix* in which the transitions by which one character state changes into another state are penalized. The matrix is typically symmetric (in traditional parsimony), but nothing prevents us from using an individually designed matrix. Once this

matrix has been established, we can calculate parsimony in a straightforward way. When working with small trees, one can actually count the cases oneself, without using any complex algorithm. When working with larger trees, using an algorithm is of course a better idea (Sankoff 1975).



The figure above shows different transition matrices (step matrices) and corresponding scenarios for parsimony analyses. What are the major differences between these approaches, and what are the consequences of using a non-symmetric matrix?

3.3 Maximum Likelihood

Maximum parsimony has a great disadvantage which consists in the lack of branch lengths, which are typically not estimated when doing a parsimony analysis. That means, the underlying model assumes that change from an ancestral to a descendant language always occurs in the same *rate*. This is of course not very *realistic* with respect to language change (and biological change), and as a result, alternative approaches were proposed already early, and one of the most important approaches is the so-called *pruning algorithm* by Felsenstein (1973), which provides a new model, based on the likelihood calculation that no longer solves the problem of finding the perfect scenario directly, but instead evaluates all possible scenarios that could happen at once, calculating the *likelihood* of each character state to appear in a certain position in the tree, while the tree is assumed to have branch lengths, which have a direct influence on the likelihood.

What is the advantage of a likelihood model over a parsimony model apart from the possibility to include branch lengths?

References

Beniamine, S. and M. Guzmán Naranjo (2021). "Multiple alignments of inflectional paradigms." *Proceedings of the Society for Computation in Linguistics* 4.21, 1–8.

Bryant, D. and V. Moulton (2004). "Neighbor-Net. An agglomerative method for the construction of phylogenetic networks." *Molecular Biology and Evolution* 21.2, 255–265.

Evans, C. L., S. J. Greenhill, J. Watts, J.-M. List, C. A. Botero, R. D. Gray, and K. R. Kirby (2021). "The uses and abuses of tree thinking in cultural evolution." *Philosophical Transcactions of the Royal Society B* 376.20200056, 1–12.

Fellner, H. A. and N. W. Hill (2019). "Word families, allofams, and the comparative method." *Cahiers de Linguistique Asie Orientale* 48.2, 91 –124.

Felsenstein, J. (1973). "Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters." English. *Systematic Zoology* 22.3, pp. 240–249. JSTOR: 2412304.

Geisler, H. and J.-M. List (2013). "Do languages grow on trees? The tree metaphor in the history of linguistics." In: *Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization*. Ed. by H. Fangerau, H. Geisler, T. Halling, and W. Martin. Stuttgart: Franz Steiner Verlag, 111–124.

Gruaz, C. (2002). "The analysis of word families and their motivational relations." Handbooks of linguistics and communication sciences 1.21. Ed. by A. Cruse, F. Hundsnurscher, M. Job, and P. R. Lutzeier, 700–704.

Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages." *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.

Hundsnurscher, F. (2002). "Das Wortfamilienproblem in der Forschungsdiskussion [Word families in scientific discussion]." Handbooks of linguistics and communication sciences 1.21. Ed. by A. Cruse, F. Hundsnurscher, M. Job, and P. R. Lutzeier, 675–680.

Jacques, G. and J.-M. List (2019). "Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them)." *Journal of Historical Linguistics* 9.1, 128–166.

Klein, W. P. (2004). "Was wurde aus den Wörtern der hebräischen Ursprache? Zur Entstehung der komparativen Linguistik aus dem Geist etymologischer Spekulation." In: *Gottes Sprache in der philologischen Werkstatt. Hebraistik vom 15. bis zum 19. Jahrhundert*. Proceedings of the Symposium "Die Geburt der Philologie aus dem Geist der Hebraistik" (Wittenberg, 10/06–10/06/2002). Ed. by G. Veltri and G. Necker. Studies in European Judaism 11. Leiden: Brill, 3–23.

Koch, P. (2001). "Lexical typology from a cognitive and linguistic point of view." In: *Linguistic typology and language universals*. Handbook of Linguistics and Communication Science 20.2. Berlin and New York: de Gruyter, 1142–1178.

Lehmann, C. (2004). "Interlinear morphemic glossing." In: *Morphology. An international handbook*. Ed. by G. E. Booij, C. Lehmann, J. Mugdan, and S. Skopeteas. Vol. 2. Berlin and New York: De Gruyter, 1834–1857.

List, J.-M. (2021a). *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets. Version 2.0.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://digling.org/edictor.

— (2021b). "Using EDICTOR 2.0 to Annotate Language-Internal Cognates in a German Wordlist." *Computer-Assisted Language Comparison in Practice* 4.4. URL: https://calc.hypotheses.org/2735.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics." *PLOS ONE* 12.1, 1–18.

List, J.-M., J. S. Pathmanathan, P. Lopez, and E. Bapteste (2016). "Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics." *Biology Direct* 11.39, 1–17.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* 4.4, 406–425.

Sankoff, D. (1975). "Minimal mutation trees of sequences." *SIAM Journal on Applied Mathematics* 28.1, 35–42.

Schleicher, A. (1853). "O jazyku litevském, zvlástě ohledem na slovanský. Čteno v posezení sekcí filologické král. České Společnosti Nauk dne 6. června 1853 On the Lithuanian language, with a special focus on Slavic." *Časopis Českého Museum* 27, 320–334. google: cLMDAAAAYAAJ.

— (1861). *Compendium der vergleichenden Grammatik der indogermanischen Sprache*. Vol. 1: *Kurzer Abriss einer Lautlehre der indogermanischen Ursprache*. Weimar: Böhlau.

Schmidt, J. (1872). *Die Verwantschaftsverhältnisse der indogermanischen Sprachen* On the genetic relations among the Indo-European languages. Weimar: Hermann Böhlau.

Schuchardt, H. (1870 [1900]). *Über die Klassifikation der romanischen Mundarten. Probe-Vorlesung, gehalten zu Leipzig am 30. April 1870* On the classification of Romance dialects. Test lecture given in Leipzig on April 30, 1870. Graz. URL: http://schuchardt.uni-graz.at/cgi-bin/print.cgi?action=show\&type=pdf\&id=724.

Schweikhard, N. E. and J.-M. List (2020). "Developing an annotation framework for word formation processes in comparative linguistics." *SKASE Journal of Theoretical Linguistics* 17.1, 2–26.

Sokal, R. R. and C. D. Michener (1958). "A statistical method for evaluating systematic relationships." *University of Kansas Scientific Bulletin* 28, 1409–1438.

Wu, M.-S. and J.-M. List (forthcoming). "Annotating cognates in phylogenetic studies of South-East Asian languages." *Language Dynamics and Change* 0.0, 1–25.