Summary by YIN Yuanhao

Annotating Cognates in Phylogenetic Studies of South-East Asian Languages

Mei-Shin Wu and Johann-Mattis List

Introduction

For the *cognate coding/annotation* of languages where derivation and compounding are frequent, it is usually difficult to find clear criteria to decide the cognacy between *partial cognates*. Should this be done just by identifying the lexical roots shared by complex words or on the basis of the underlying motivations of these words which imply an implicit judgement of the parts of these words.

Increasing the Transparency of Cognate Annotation

As for the cognate coding in South-East Asian languages, there exist two extremes : one takes the morpheme as a basic unit assigned to cognate sets, the other takes the translational words corresponding to concepts instead. One of the two extremes will not take into account the unit of the other level, but the first can be avoided with a careful annotation of partial cognates.

There are two most straightforward approaches to assign words to cognate sets : (1) the *strict* cognate coding and (2) the *loose* cognate coding. In the strict case, only those words sharing all the cognate morphemes within will be assigned to the same cognate sets ; in the loose case, words are linked together whenever two of them share at least one cognate morpheme. Each approach will lead to problems : the strict one increases differences between language varieties, while the loose one risks assigning two words sharing no cognate morphemes to the same cognate set.

Sagart et al. (2019) mitigate the weakness of these two approaches by (3) alignment analyses (Are these analyses the same as the interlinear annotation?) to make sure that there is at least a common cognate morpheme among the words of the same cognate set. But this approach cannot be checked automatically. (Why?)

Following the three approaches above, the authors propose the (4) *morpheme gloss* to specify the *semantic motivations* of those complex words and the *representative* or *salient* parts of those words so as to increase the transparency of the cognate coding. Those partial cognates annotated as "salient" are selected from the original words to undergo then the (1) strict procedure.

Question :

How to understand this sentence : "a certain suffix occurs too frequently in a given dataset to be worthwhile to play a significant enough role to decide if one word that has the suffix should be cognate with another word that lacks the suffix." This is related to another question : what is the criteria to decide if one morpheme is "salient" or not ? One of these criteria proposed in the article is : "salient morphemes, that is, morphemes one deems representative for the whole history of the words", but it seems too general and too abstract.

On the other hand, when it comes to the full word cognacy, there are three major problems to consider or to avoid : "(1) exclude the words suspected of parallel evolution (*avoid homoplasy*), (2) identify the character dependency and re-code the data accordingly (*minimize character*

dependency), and (3) check the variation of the words (*control variation*)."
Note : these three procedure all seem to seek to simplify the full word system reserved for phylogenetic analyses either by excluding those words having the parts which are likely to affect the result of computation of cognacy (*avoid homoplasy*) or by excluding those component parts from the words (*minimize character dependency* and *control variation*).

A Case Study on Chinese Dialect History
Methods
1. Deriving Full Cognates from Partial Cognates (Some doubts)
(1) Construct fuzzy clusters
(2) Order clusters by size (What if the numbers of the words of two clusters are identical ?)
(3) Mark all morphemes of the first cluster as "salient" (The article uses "all words", which seems a typo.)
(4) Remove all the words containing this morpheme from the remaining clusters
(5) Iterate over the remaining clusters

2. Identifying Potential Cases of Homoplasy and Character Dependencies (Badly understood)
-Automated comparison between strict and loose cognate coding
(1) Compute strict cognates from the partial cognates (How to do it ?)
(2) Compute loose cognates from the partial cognates (ibid.)
(3) F-Scores from 0 (completely different clusters) to 1 (identical clusters)

-Consistent identification of cognates across meaning slots
(1) Count in how many concepts a distinct morpheme recurs (1 morpheme > ? concepts)
(2) Average the number of cross-semantic partial cognates for each individual word (1 word > ? partial cognates ; how to calculate ? What's the meaning of "cross-semantic partial cognates" ?)
(3) Average the individual word scores for an entire meaning slot (1 meaning slot > ? words ; how to calculate ?)

3. Annotating Salient Morphemes
-Manually refine the dataset computationally processed (The question still remains : what are the criteria to decide if a morpheme is "salient" or not ?)

The three above procedures finished, all the cognate sets annotated as non-salient are removed, and the remaining cognate sets are computed with the strict cognate coding.

Results
1. Identifying Concepts Susceptible to High Variation
Lowest F-Scores : concepts with high variation which mostly comprise complex nouns, a few complex verbs and demonstrative pronouns. These concepts could cause problems in later phylogenetic analyses.

Highest F-Scores : concepts expressed by monosyllabic words which comprise most adjectives, most basic verbs and some very basic nouns. The monosyllabicity is a tendency.

How to understand this sentence : “The fact that both tests only correlate weakly emphasizes how important it is to use both of them when investigating the potential impact of partial cognates on lexical phylogenies.” What is the logic between these two propositions ?

Ten concepts with the highest average number of colexifications per word and per concept slot. Why is 树皮 “bark” among these concepts ? Isn’t this a little counterintuitive ?

Ten concepts for which no colexifications could be identified throughout all words. There is no clear tendency apart from the monosyllabicity.
How to understand this sentence : “One can be tempted to assume that our concept of “morpheme saliency” might be replaced by some independent principle, such as, for example, the underlying dependency structure of compound words expressing a given concept.” According to the following examples, it seems to argue that it is not always the heads or the modifiers of the words that are “salient”. But for the example of 月亮 “moon”, why is then the unstable part, the head 亮 “shine”/光 “ray”, rather than the modifier 月 “moon”, that is “salient” ? “we therefore find 月, the modifier, as the stable part, while the head of the compound has changed and would therefore be treated as the salient morpheme in our annotation.” Maybe a more basic question is, why is 月 “moon” the modifier and 亮 “shine” the head, rather than the reverse ? Because in some dialects like mine, the structure is 亮月, literally “bright moon”. The most important question still consists in the criteria to decide if a morpheme is “salient” or not, despite this sentence which hedges this question : “the saliency of a morpheme with respect to the history of the word in which the morpheme occurs cannot be determined from the dependency structure alone, although the dependency structure is of crucial importance when it comes to identify the underlying motivation that led to the creation of a compound.” So far the only criterion apparently relevant comes from this sentence aforementioned : “a certain suffix occurs too frequently in a given dataset to be worthwhile to play a significant enough role to decide if one word that has the suffix should be cognate with another word that lacks the suffix.” Maybe this is why 月 “moon” is not taken as “salient”.

2. Cognate Coding and Language Distances
Two sets of distance matrices (one using all the 201 cognate sets, the other using those 59 cognate sets which are more likely to cause problems due to less than 0.8 F-Scores) for four approaches of cognate coding (strict one, loose one, one with common morphemes, one with salient morphemes) are computed. The distance matrices using all the 201 cognate sets for four approaches do not show many differences between one another. When using those 59 cognate sets with less than 0.8 F-Scores, the differences of the distance matrices between four approaches become larger, although they still show quite a correlation between one another : the least correlated pair is between the strict and the loose cognate coding, which is normal because they represent two extremes, and the most similar pair is between the strict cognate coding and that with salient morphemes.

Two heatmaps are made to visualize the distances between the dialects with the strict and the loose cognate coding respectively, which are different especially between the northern and the southern dialects. This can be attributed to the fact that the northern and the southern dialects

are composed of similar morphemes but with different syllable structures, the northern dialects with more multisyllabic words and the southern dialects with more monosyllabic ones. Thus the loose cognate coding will yield much higher similarities between the dialects than the strict cognate coding.

3. Partial Cognates and Language Phylogenies

Four consensus trees by Bayesian phylogenies based on the four approaches of cognate coding all diverge from the traditional accounts of Chinese dialect evolution and differ regarding the degree to which they diverge from the traditional scenarios. This can be attributed to undetected borrowings, dialect convergences and a small number of concepts susceptible to variation. Moreover, the internal age estimates show some remarkable differences.

Discussion and Outlook

Among the four approaches of cognate coding, the *loose conversion scheme* performs worst, leading to mostly star-like phylogenies without much resolution, clearly wrong groupings of varieties and largely inconsistent age estimates, because of the tendency to increase the similarities between varieties by assigning two words sharing no cognate morpheme to the same cognate set.

The *common morpheme conversion scheme* performs better with respect to the resolution but yields inconsistent groupings in comparison with traditional accounts.“

How to understand the reason resulting in this problem proposed by this article : The reason for these problems can be found in the greediness of the approach, which does not further differentiate morphemes with respect to their potential to reflect overall word histories.”

The *strict conversion scheme* and the *salient morpheme conversion scheme* perform best. The strict one leads to higher resolution of the phylogeny at the risk of increasing the distance between varieties by requiring only two words sharing all the cognate morphemes to assign to the same cognate set.

As a conclusion, the authors recommend the *salient conversion scheme*. However, the *strict conversion scheme* is still worthwhile to refer to while the other two approaches should be taken with great care.