An approach to cross-concept cognacy identification

Johannes Wahle
Seminar für Sprachwissenschaft
EVOLAEMP
Universität Tübingen
johannes.wahle@uni-tuebingen.de

***Abstract*—It is a well known phenomenon in historical linguistics, that the meaning of a proto form is different to the meaning of its descendants. This phenomenon of meaning change is often ignored in studies which use tools from statistical phylogenetic analysis to determine language relationships. It has been shown, that the databases currently used in linguistic phylogeny exhibit a considerable amount of the described phenomenon. The current study proposes a method to detect such instances of cross-concept relationships of words. Although the evaluation can not be done by standard means, the results indicate that semantic similarity is a good indicator for cross-concept relationships and that tools from computational biology offer a good framework for this kind of approach.**

I. Introduction

Large datasets of linguistic data recently have given rise to studies investigating language relationships using tools from statistical phylogenetic analysis ([1], [2] among others), which exploit ideas developed in the area of computational biology [3]. These approaches base their analysis in some way or the other on lexical traits. As [4] point out, when working with these traits, the difference of *cognate traits* and *root-meaning traits* has to be taken into account. They illustrate that these two types of traits behave differently for example with respect to homoplasy[1]. Furthermore, they show that the amount of homoplasy in modern languages is considerable. A homoplastic event which includes a semantic drift can lead to the situation that a proto form describes another concept then its descendants, e.g. “reflexes of the PIE (Proto Indo-European) **pod-* ‘foot’ came to mean ‘leg’ independently in Modern Greek and modern Indic and Iranian languages”[4], [5]. These observations suggest that it is necessary to include information about cross-concept relationships into statistical phylogenetic analysis.[2]

The current study uses Profile Hidden Markov models to investigate such extended clusters of related words. Profile Hidden Markov models (ProfHMMs) were developed in the area of computational biology [6]. They are used for several tasks, such as simultaneously aligning multiple related sequences or determining the membership of a new sequence in an existing cluster of sequences. ProfHMMs have already been successfully applied in linguistic research [7].

II. Profile Hidden Markov models

Profile Hidden Markov models are a tool to probabilistically model a multiple alignment. They are a common tool in computational biology to determine the membership of a new sequence in a given family. There are databases such as the PFAM [8] database which provide ProfHMMs for a large amount of protein families. A multiple alignment of a family of sequences is considered to represent a consensus (sequence) of this family. The different states of the ProfHMM give a probabilistic model of the consensus, so aligning a new sequence against such a ProfHMM yields a probabilistic measure of membership of the given sequence in the family.

Figure 1 shows the typical structure of a ProfHMM. There are three types of states, a *match*, an *insertion* and a *deletion* forming a column. Thus, a ProfHMM can be visualized by a series of columns where a column represents a position in the consensus sequence. Match states model the occurrence of a symbol s at position j. The occurrence of s at position j is is based on a set of emission probabilities. An insertion states allows the insertion of a character at position j. This state is used to describe the existence of a symbol which is not present in the consensus. The opposing phenomenon is captured by the deletion state. The transitions are indicated by the arrows in Figure 1. A match state at position j has a transition to the insertion state at the same position as well as to the deletion or match state at position $j+1$. Deletion states has a transition to a deletion state or a match state at the next position or an insertion state at the same position. Insertion states are the only states which allow self transition. This accounts for the possibility of multiple insertions at a given state. The classical algorithms from the Hidden Markov Model literature, such as the Baum-Welch, forward or Viterbi-algorithm can be adapted to the needs of ProfHMMs.

III. Experiments

A. Data

Using the LingPy library ([9], [10]) word lists for 30 languages[3] of the NorthEuraLex database [11] were clustered into cognate sets. The NorthEuraLex database provides translations of 1016 concepts for more than 70 languages of

[1] “Homoplasy is an evolutionary term for independent analogous innovation in parallel lineages.”[4]

[2] Although [4] show the difference between *root meaning* and *cognacy* the term *cognacy* will be used throughout this paper.

[3] Icelandic, Norwegian, Swedish, Danish, German, Dutch, English, French, Spanish, Portuguese, Italian, Romanian, Finnish, North Karelian, Olonets Karelian, Veps, Standard Estonian, Livonian, Russian, Polish, Czech, Slovak, Croatian, Bulgarian, Modern Greek, Standard Albanian, Latvian, Lithuanian, Irish Gaelic, Welsh

Fig. 1. Profile Hidden Markov model; M_j Match state at position j, I_j insertion state at position j and D_j deletion state at position j. The arrows indicate probabilistic transitions. [6], [7]

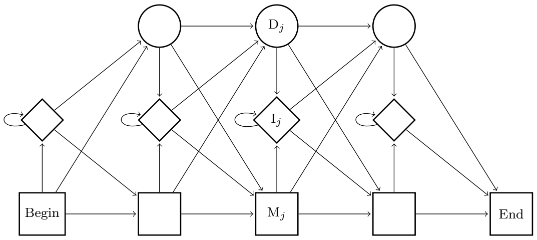


TABLE I
DISTRIBUTION OF MEMBERS PER COGNATE CLASS

members	cognate classes	members	cognate classes
2	3102	9	84
3	1371	10	58
4	901	11	38
5	696	12	25
6	706	13	21
7	338	14	15
8	155	$\geq$ 15	67

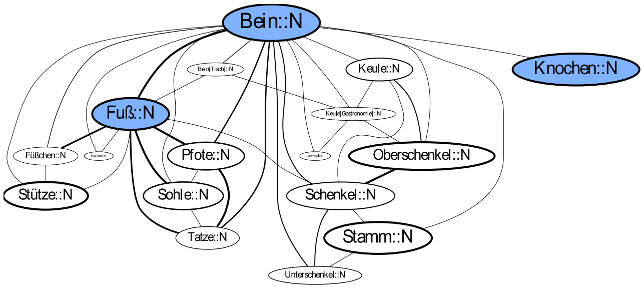
Northern Eurasia. For each of the cognate classes multiple alignments of the phonetic forms were retrieved using the LingPy library. There are in total 5303 cognate classes with just one member and 7577 cognate classes with more than one member. Using notions from set theory, cognate classes with just one member are called *singletons*. Table I shows the distribution of members per cognate class for classes with more than one member.

Relationship beyond standard cognate classes is often caused by *semantic drift* [4]. The 'foot' – 'leg' case from above is an instance of this phenomenon; another instance is the change from 'bone' to 'leg' in German. Polysemy networks have been suggested as a possible model for semantic change [12], [13]. Figure 2 shows a snapshot of a polysemy graph presented by [14], [15]. They constructed a polysemy network which was only generated using dictionary glosses. The thickness of the lines indicates the amount of languages having a word which is polsyemous between the two concepts. As indicated by the blue nodes the above mentioned instances of semantic drift are very closely connteced in this graph. A current version of this network, which was kindly provided by the author, was trimmed to the number of concepts in the NorthEuraLex database. For the current study a node in this network can be seen as a set of cognate classes. Where a cognate class is a set of words which are assumed to descent from the same common ancestor.

B. Model construction & Training

For each of the automatically determined cognate sets with more than one entry a ProfHMM was built and trained on the phonetic forms of the entries using the Baum-Welch algorithm. The number of match states was equal to the number of columns in the multiple alignment with less than 50% gap symbols [6]. The initial transition parameters of the ProfHMMs were derived from a Dirichlet distribution with weight 5.0 for each transition. Since each ProfHMM could only be trained on a small set of items the sound emission parameters were estimated on the basis of pseudo-counts (equation 1). The count of each sound a at each position j (c_{ja}) was weighted by the background frequency of this sound (q_a).

Fig. 2. Snapshot of the polysemy network. ('Bein::N'-leg, 'Fuß::N'- foot, 'Knochen::N' - bone)



$$e_j(a) = \frac{c_{ja} + q_a}{\sum_{a'} c_{ja'}} \quad (1)$$

The background frequency was determined for three different positions; left boundary, right boundary and middle. Depending on the position of the match state, q_a was taken from one of the three distributions, e.g. if $j = 0$ q_a was taken from the distribution for the left boundary. This approach roughly accounts for some distributional characteristics of different sounds.

C. Experiments

Building on the work of [7] for linguistics and [16] for computational biology, the membership of a sequence in a cluster is tested using the forward algorithm. The forward algorithm determines how well a given sequence fits the model. Since the model is a representation of the consensus of the family, the forward algorithm measures the fit of the sequence and the consensus and thus the fit of the sequence and the family. This makes the forward algorithm the optimal algorithm for the given study.

For each singleton c the *proportion of assignment* ($\mathtt{PRA}$) is calculated, i.e. the proportion of nodes in distance n to which this translation can be assigned. A singleton c is considered to be a member of a given cognate class χ if the condition in equation 2 is fulfilled. Membership in a cognate cluster was determined in the following way. For each word in the cluster, the log-odds scores using the forward algorithm were calculated and then averaged ($\bar{\chi}$). If the score for a new sequence c ($sc(c,\chi)$) fell within one standard deviation, the sequence is considered to be a member of this cluster.

$$fit(c,\chi) = \begin{cases} True & \bar{\chi} - \sigma \leq sc(c,\chi) \leq \bar{\chi} + \sigma \\ False & \text{else} \end{cases} \quad (2)$$

If there is a cluster for which equation 2 evaluates to true, the c is assigned to node X for which $\chi \in X$ holds (see 3).

TABLE II
PRA AND AVERAGE NUMBER OF NODES AT DISTANCE n

n	PRA	# neighbours
1	8.03	6.7
2	6.73	37.0
3	6.13	103.1
4	5.73	172.5

TABLE III
COMPARISON OF ACTUAL PRAS WITH THE MEAN PRAS CALCULATED FROM RANDOM NETWORKS

n	PRA	$mean(\texttt{PRA})$	σ	p-value
1	8.03	5.49	0.10	2.2e-16
2	6.73	5.48	0.05	2.2e-16
3	6.13	5.50	0.02	2.2e-16
4	5.73	5.52	0.02	2.2e-16

$$assign(c, X) = \begin{cases} True & \exists\chi : \chi \in X \wedge fit(c, \chi) \\ False & \text{else} \end{cases} \quad (3)$$

The PRA at distance n is then the proportion of nodes X for which 3 evaluates to true (see 4).

$$\texttt{PRA} = \frac{|\{X|assign(c, X) = True \ \wedge \ C \sim X = n\}|}{|\{X'|C \sim X' = n\}|} \quad (4)$$

An evaluation of the resulting cognate clusters is not as straight forward as for standard cognate detection. Since there is neither cognate class information for the NorthEuraLex database, nor is there large scale data for cross-concept cognacy the evaluation of this approach can not be done by standard means. To test these results the PRA was calculated for random network structures.

IV. Results

The results of the PRA measures on the actual data set are shown in Table II. The result show that singletons fit into semantically related clusters. It is important to equate the PRA with the average number of neighbours at distance n to assess the PRA. For each of the simulated random networks the PRA is calculated and based on these results a probability distribution is estimated. The mean of these PRAs ($mean(\texttt{PRA})$) is about 5.5 for each distance. The standard deviation (σ) for each distance is very small. This indicates that the noise level is about 5.5 for the PRA measure in this experiment. As the results from the random networks suggest (Table III) the PRA for nodes at distance 4 come suspiciously close to a level of noise.

As it can be seen by comparing the PRAs derived from the random networks with the results from the actual data, the observed effect of semantic proximity is supported. Although a simple t-test indicates that the results are still significant for higher distances, it is clearly observable that the actual PRA approaches random as the semantic distance increases.

V. Discussion

This study proposes the application of an attested framework from computational biology, Profile Hidden Markov models, to the task of cognate identification across concept boundaries. Classical cognate identification is normally done just within one concept. The evaluation of the method presented here, has to be different than the evaluation of standard cognacy detection tasks. Current databases for phylogenetic analyses in linguistics do not encode cross-concept cognacy. Therefore, the evaluation of the current study works as a proof of concept by showing, that the obtained results are not random effects. The results of this study are also supported by the homoplasy observations made in [4].

The results indicate that the current method offers a way to determine cross-concept similarities which are important for computational historical linguistics. This method can give suggestions for the assignment of words to cross-concept lexical traits. As several studies suggest target cognate classes are very often found in the close semantic neighbourhood of a singleton [12], [15], [13].

VI. Future Work

Since the results of this approach could not be evaluated directly, there observed tendency needs to be checked more carefully. This can be done in different ways in future research; including an improved training methodology which can also deal with the small amount of training data, using other alignment methods or the development of a probabilistic measure of membership.

Acknowledgment

The author has been supported by the ERC Advanced Grant 324246 EVOLAEMP - Language Evolution: the Empirical Turn, which is gratefully acknowledged. The author especially thanks Gerhard Jäger, Armin Buch, Igor Yanovich, Christian Bentz, Taraka Rama, Johannes Dellert and Marisa Köllner for insightful discussions.

References

[1] R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson, "Mapping the origins and expansion of the indo-european language family," *Science*, vol. 337, no. 6097, pp. 957–960, 2012.

[2] G. Jäger, "Phylogenetic inference from word lists using weighted alignment with empirically determined weights," *Language Dynamics and Change*, vol. 3, no. 2, pp. 245–291, 2013.

[3] J. Felsenstein, *Inffering Phylogenies*. Sunderland: Sinauer Associates, 2004.

[4] W. Chang, C. Cathcart, D. Hall, and A. Garrett, "Ancestry-constrained phylogenetic analysis supports the indo-european steppe hypothesis," *Language*, vol. 91, no. 1, pp. 194–244, 2015.

[5] M. Urban, *Lexical semantic change and semantic reconstruction*. London: Routledge, 2015, ch. 16.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, repr. ed. Cambridge Univ. Press, 2001.

[7] A. Bhargava and G. Kondrak, "Multiple word alignment with profile hidden markov models," in *Proceedings of NAACL HLT Student Research Workshop and Doctoral Consortium*. Boulder, Colorardo: Association for Computational Linguistics, 2009, pp. 43–48.

[8] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: the protein families database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D222–D230, 2014.

[9] J.-M. List, S. Moran, P. Bouda, and J. Dellert, "LingPy. Python Library for Automatic Tasks in Historical Linguistics," 2013. [Online]. Available: http://www.lingpy.org

[10] J.-M. List and S. Moran, "An Open Source Toolkit for Quantitative Historical Linguistics," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 13–18.

[11] J. Dellert, "Evaluating cross-linguistic polysemies as a model of semantic change for cognate finding," in *Workshop on semantic technologies for research in the humanities and social sciences (STRiX)*, 2014.

[12] H. Youn, L. Sutton, E. Smith, C. Moore, J. F. Wilkins, I. Maddieson, W. Croft, and T. Bhattacharya, "On the universal structure of human lexical semantics," vol. 113, no. 7, pp. 1766–1771, 2016.

[13] J.-M. List, A. Terhalle, and M. Urban, "Using network approaches to enhance the analysis of cross-linguistic polysemies," in *Proceedings of the 10th International Conference on Computational Semantics - Short Papers*, Stroudsburg, 2013, pp. 347–353.

[14] J. Dellert, "Compiling the uralic dataset for northeuralex, a lexicostatistical database of northern eurasia," in *First International Workshop on Computational Linguistics for Uralic Languages*, 2015.

[15] A. Münch and J. Dellert, "Evaluating the Potential of a Large-Scale Polysemy Network as a Model of Plausible Semantic Shifts," in *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, J. Wahle, M. Köllner, H. Baayen, G. Jäger, and T. Baayen-Oudshoorn, Eds., 2015. [Online]. Available: http://dx.doi.org/10.15496/publikation-8626

[16] A. Krogh, M. Brown, I. Mian, K. Sjölander, and D. Haussler, "Hidden markov models in computational biology: Applications to protein modeling," *Journal of Molecular Biology*, vol. 235, no. 5, pp. 1501 – 1531, 1994.