Analytical Database Design: Approaches in the Mapping between Cognate and Semantic Sets*

Tyler Peterson

1 Project Overview and Objectives

This paper outlines the third of three phases of the analytical database project *The Experimental and Historical Phonology Reconstruction Database - Tupí*, which involves the development of a research program that explores the potential correspondences and correlations between sets of cognates in a language family and their meanings or 'semantic sets'.[1] Several fundamental questions immediately arise in approaching this task, including: is it necessarily the case that a cognate set C also maps to a corresponding semantic set S? At what point do 'distinct' cognate sets also reflect a diffusion in meanings, where the cognate sets A, B and C correspond with the denotations d_1, d_2 and d_3, respectively? Are there any systematic overlaps between cognate and semantic sets? This project phase, taking the data and generalizations collected from the previous two phases of the project, seeks to address these questions through developing a research program that tests the correspondences and correlations, if any, between cognate sets and semantic sets. The ultimate goal of this program is a formal model or mapping function that captures any and all correspondences between sets of cognates and sets of meanings, and its practical implementation within the EHPRD-T database. This would offer a concrete metric that could be used to measure such correspondences, and with this, the possibility of reconstructing the proto-meanings in a language family to formally complement a reconstructed proto-lexicon.

2 Methodology

In the main user interface of the database (which is currently implemented in MS Access), a gloss is indexed to a number. Every language has a variety of

*This paper appears in *Proceedings of the Seventh International Workshop on Computational Semantics IWCS-7*, J. Goertzen, E. Thijsse, H. Bunt, A. Schiffrin (eds). Tilburg: Tilburg University Dept. of Communication and Information Sciences. pp. 359-361

[1] The Portuguese title is the *Programa de Fonologia Experimental e Histórica - Base de Dados para Reconstrução Fonologica: Tupí*, which is currently being developed as joint project between Tyler Peterson at the University of British Columbia and Gessiane Picanço at the Museu Paraense Emílio Goeldi (MPEG).

fields also indexed to that gloss and number. In these individual language sub-interfaces, the user segments the phonemes of that word in that language of that number. There are separate fields are set up for each segment, and well as a variety of sub-segmental features that can be associated with that segment. Various relational operations can be done (through the use of SQL queries) by the user to track phonemic correspondences across the family for any given word. The second phase developed a separate module in the database where users take the words and segmented data associated with them, and organize them into groups or sets based on the features that minimally distinguish them. It is this second project phase that interacts with the program described here.

Consider a very simple hypothetical case: a gloss x is translated in a language L_n as *tik* in L_1^x, *dik* in L_2^x, *dis* in L_3^x, and *nis* in L_4^x. Each of these words form an ordered set of segments: $L_1^x = \langle t, i, k \rangle$, $L_2^x = \langle d, i, k \rangle$, $L_3^x = \langle d, i, s \rangle$, $L_4^x = \langle n, i, s \rangle$. We can then group these three sets by what minimally distinguishes these words: *tik* and *dik* differ only in their first segment, t and d, thus forming set A^2, or $L_1^x \cap L_2^x = A^2$. *dik* and *dis* differ only in their last segment, k and s, thus forming set B^2 ($= L_2^x \cap L_3^x$), and *dis* and *nis* differ in their first segments, forming set C^2 ($= L_3^x \cap L_4^x$). The superscript numbers indicate the *cognate strata*: $L_1^x \cap L_2^x$ and $L_2^x \cap L_3^x$ are considered second-level cognate strata, A^2 and B^2 respectively, because their intersection is a non-empty set, meaning they share at least one segment in common (i.e. $A^2 = \{i, k\}$).[2] Additional cognate strata are formed by successive intersection with other cognate strata: $\bigcap\{A^2, B^2, C^2\} = A^3$. What this is intended to show is that cognate sets are not discrete: they form a continuum, which is expressed through successive intersections of these sets to form supersets, or the different cognate strata.

We speculate that cognate set continua correlate with semantic set continua. As a first pass at this hypothesis, take an example comparing the Tupí languages Karitiana and Karo. In Karo, the word for 'fire' is *cán* while 'firewood' is *cat*, and in Karitiana 'fire', *iso*, and 'firewood', *sõn* (Gabas 1999; Storto 1999). Next, consider that meanings may also group into the same types of strata as outlined above. Under this view, $S^1 = \{fire\}$ and $T^1 = \{firewood\}$. Assuming that we have arrived at three cognate sets at some stratum n: $A^n = \{$*cat, sõn*$\}$, $B^n = \{$*cán*$\}$, and $C^n = \{$*iso*$\}$, the next step involves formulating a mapping relation from the members of these cognate sets into their corresponding meanings. A straightforward mapping would be a relation $R : A^n \to S^1$, and $R : B^n \wedge C^n \to T^1$. However, in Mundurukú, the word *dashá* can mean both 'fire' and 'firewood' (Picanço 2005). We may be able to model this as follows: because of their relational similarity, we can intersect S^1 and T^1, forming a second semantic stratum: $S^1 \cap T^1 = S^2$. Thus, the cognate set $D^n = \{$*dashá*$\}$ maps not to S^1 or T^1, but $R : D^n \to S^2$. This illustrates that at different cognate strata the mappings between them and the semantic strata are going converge or diverge.

Before this model can be implemented and tested within EHPRD-T database, several practical issues must be attended to. The most important of these in-

[2] $L_1^x \cap L_4^x = \{i\} = D^2$ would not be as 'strong' a cognate set as A^2, B^2, or C^2 because their intersection contains only one segment. We hypothesize that this is measured in terms of cardinality: i.e. $|A^2|, |B^2|, |C^2| = 2 \succ |D^2| = 1$. See Peterson 2006 for more details.

volves the design and implementation of a semantics module within the database, and endowing it with the appropriate sets of fields and features that can interface with the cognate set data. We plan to proceed with this in two steps beginning with developing a semantics module and stipulating the semantic sets and denotations contained in it based on the translated data. These fields will be coupled with the data fields from the cognate sets. One possibility is tagging, where cognate sets and their supersets are tagged with the semantic features that can then be related and evaluated. Once the results of this are tested, the semantics sets will be replaced and enriched with featural information. A compatible model would be a *qualia*-type structure, which allows for the specification of the different aspects of a word's meaning through the use of subtyping features in a matrix (Pustejovsky, Bergler, and Anick 1994). A feature matrix of this type would also be easily implementable within the database architecture.

3 Expected Outcomes

While still in the early experimentation and design stages, we believe this research program offers the possibility of developing a practical metric for measuring and tracking systematic cognate-semantic correspondences - or the lack of - between cognate and semantic sets (strata) across a wide variety of semantic and grammatical categories. Preliminary research has also shown that we may also be able to systematically track relationships which may hold between polysemous senses (i.e. hyponymy and metonymy) across branches of the family.

In sum, we believe the EHPRD-T project offers a practical, innovative approach to analytical database design and their uses in comparative linguistics, while offering a computational approach to tracking and mapping any potential, incremental correlations between cognate sets and their corresponding meanings for any individual word across a language family.

References

Gabbas Jr., N. 1999. *A Grammar of Karo, Tupí (Brazil)* Doctoral Dissertation. University of California, Santa Barbara.

Pustejovsky, J., Bergler, S., and Anick, P.G. 1994. “Lexical Semantic Techniques for Corpus Analysis” in *Computational Linguistics* 19, 331-358

Peterson, T. 2006. “Evaluating Cognate Sets”. UBC Ms. draft

Picanço, G. 2005. “Mundurukú: Phonetics, Phonology, Synchrony, Diachrony” Doctoral Dissertation. UBC

Storto, L. 1999. *Aspects of a Karitiana grammar* Doctoral Dissertation. MIT.