Creating a Comparative Dictionary of Totonac-Tepehua

Grzegorz Kondrak
Department of Computing Science
University of Alberta
kondrak@cs.ualberta.ca

David Beck
Department of Linguistics
University of Alberta
dbeck@ualberta.ca

Philip Dilts
Department of Linguistics
University of Alberta
pdilts@ualberta.ca

Abstract

We apply algorithms for the identification of cognates and recurrent sound correspondences proposed by Kondrak (2002) to the Totonac-Tepehua family of indigenous languages in Mexico. We show that by combining expert linguistic knowledge with computational analysis, it is possible to quickly identify a large number of cognate sets within the family. Our objective is to provide tools for rapid construction of comparative dictionaries for relatively unfamiliar language families.

1 Introduction

Identification of cognates and recurrent sound correspondences is a component of two principal tasks of historical linguistics: demonstrating the relatedness of languages, and reconstructing the histories of language families. Manually compiling the list of cognates is an error-prone and time-consuming task. Several methods for constructing comparative dictionaries have been proposed and applied to specific language families: Algonquian (Hewson, 1974), Yuman (Johnson, 1985), Tamang (Lowe and Mazaudon, 1994), and Malayo-Javanic (Oakes, 2000). Most of those methods crucially depend on previously determined regular sound correspondences; each of them was both developed and tested on a single language family.

Kondrak (2002) proposes a number of algorithms for automatically detecting and quantifying three characteristics of cognates: recurrent sound correspondences, phonetic similarity, and semantic affinity. The algorithms were tested on two well-studied language families: Indo-European and Algonquian. In this paper, we apply them instead to a set of languages whose mutual relationship is still being investigated. This is consistent with the original research goal of providing tools for the analysis of relatively unfamiliar languages represented by word lists. We show that by combining expert linguistic knowledge with computational analysis, it is possible to quickly identify a large number of cognate sets within a relatively little-studied language family.

The experiments reported in this paper were performed in the context of the Upper Necaxa Totonac Project (Beck, 2005), of which one of the authors is the principal investigator. Upper Necaxa is a seriously endangered language spoken by around 3,400 indigenous people in Puebla State, Mexico. The primary goal of the project is to document the language through the compilation of an extensive dictionary and other resources, which may aid revitalization efforts. One aim of the project is the investigation of the relationship between Upper Necaxa Totonac and the other languages of the Totonac-Tepehua language family, whose family tree is not yet well-understood.

The paper is organized as follows. In Section 2, we provide background on the Totonac-Tepehua family. Section 3 describes our data sets. In Section 4, we outline our algorithms. In Section 5, we report on a pilot study involving only two languages. In Section 6, we present the details of our system that generates a comparative dictionary involving five languages. Section 7 discusses the practical significance of our project.

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 134–141,
Prague, June 2007.

2 Totonac-Tepehua Language Family

The Totonac-Tepehua language family is an isolate group of languages spoken by around 200,000 people in the northern part of Puebla State and the adjacent areas of Veracruz and Hidalgo in East-Central Mexico (Figure 1). Although individual languages have begun to receive some attention from linguists, relatively little is known about the family as whole: recent estimates put the number of languages in the group between 14 and 20, but the phylo-genetic relations between languages remains a subject of some controversy. The family has traditionally been divided into two coordinate branches: Tepehua, consisting of three languages (Pisa Flores, Tlachichilco, and Huehuetla), and Totonacan. The Totonacan branch has in turn been divided into four sub-branches: Misantla, Lowlands or Papantla, Sierra, and Northern (Ichon, 1973; Reid, 1991), largely on the impressions of missionaries working in the area. Some dialectological work has cast doubt on the division between Northern and Sierra (Arana, 1953; Rojas, 1978), and groups them together into a rather heterogeneous Highland Totonac, suggesting that this split may be more recent than the others. However, the experience of linguists working in Totonacan communities, including one of the authors, indicates that – judged by the criterion of mutual intelligibility – there are likely to be more, rather than fewer, divisions needed within the Totonacan branch of the family.

Although Totonac-Tepehua shows a good deal of internal diversity, the languages that make it up are easily recognizable as a family. Speakers of Totonacan languages are aware of having a common historical and linguistic background, and there are large numbers of easily recognizable cognates and grammatical similarities. A typical Totonacan consonantal inventory, that of the Papantla variant (Levy, 1987), is given in Table 1. Most languages of the family share this inventory, though one of the languages used for this study, Upper Necaxa, has undergone a number of phonological shifts that have affected its consonantal system, most notably the collapse of the voiceless lateral affricate with the voiceless lateral fricative (both are now fricatives) and the lenition of the uvular stop to a glottal stop, a process that has also affected at least some of the

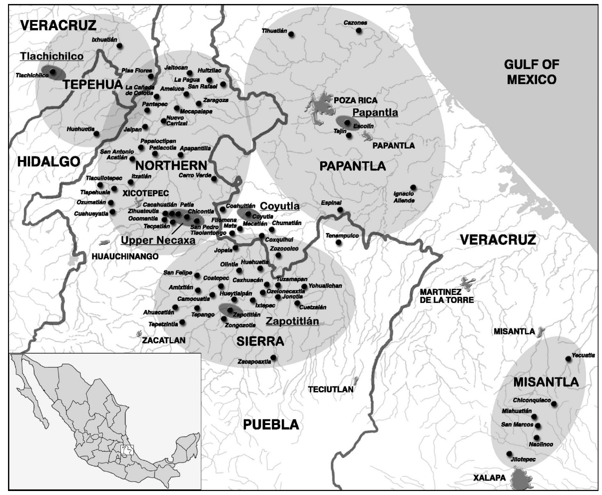

Figure 1: Totonac-Tepehua language area indicating traditional taxonomic divisions.

Tepehua languages. In Upper Necaxa, this lenition has also resulted in the creation of ejective fricatives from historical stop-uvular stop clusters (Beck, 2006). Languages also differ as to whether the back-fricative consonant is /h/ or /x/, and some languages have evolved voiceless /w/ and/or voiceless /y/ phonemes in word-final position. The phonemic status of the glottal stop is an open question in several of the languages.

Plosive	p	t			k	q	
Affricate		ʦ	tɬ	ʧ			
Fricative		s	ɬ	ʃ			h
Approximant	w		l	j			
Nasal	m	n				ɴ	

Table 1: Illustrative Totonac-Tepehua consonantal inventory.

In terms of vocalic inventory, it was previously thought that all Totonacan languages had three-vowel systems (/a/, /i/, /u/), and that they also made distinctions for each vowel quality in vowel length and laryngealization. It has since come to light that at least some languages in the Sierra group do not make length distinctions (in at least one of these, Olintla, it appears that short vowels have developed into a phonemic schwa), and that others do not distinguish laryngealized vowels. A number of languages, including Upper Necaxa and some of the languages adjacent to it, have developed a five-vowel system; the sounds /e/ and /o/ are recognized in the orthographies of several languages of the family even where their phonemic status is in doubt.

3 Data

There are five languages included in this study: Tlachichilco (abbreviated **T**), Upper Necaxa (**U**), Papantla (**P**), Coyutla (**C**), and Zapotitlán (**S**). Tlachichilco belongs to the Tepehua branch; the other four are from the Totonacan branch. Zapotitlán is traditionally considered to belong to the Sierra group of Totonacan, whereas the status of Coyutla is uncertain. The location of each language is indicated by grey lozenges on Figure 1.

The data comes from several diverse sources. The Tlachichilco Tepehua data are drawn from an electronic lexical database provided to the authors by James Watters of the Summer Institute of Linguistics. The data on Upper Necaxa was collected by Beck in the communities of Patla and Chicontla – located in the so-called Northern Totonac area – and data from the Papantla area was provided by Paulette Levy of the National Autonomous University of Mexico based on her field work in the vicinity of the city of Papantla. Data on the remaining two languages were provided by Herman Aschmann. The material from Coyutla was drawn from a word list compiled for Bible translation and the Zapotitlán material has been published in dictionary form (Aschmann, 1983). The glosses of Totonac forms for all the languages are in Spanish.

The dictionaries differ significantly in format and character encoding. The Tepehua and Coyutla dictionaries are in a file format and character encoding used by the *Shoebox* program. The Upper Necaxa and the Zapotitlán dictionaries are in their own formats and character encodings. The Papantla dictionary is in the RTF format. The dictionaries also differ in orthographies used. For example, while most dictionaries use k to represent a voiceless velar stop, the Coyutla dictionary uses c.

4 Methods

In this section, we briefly outline the algorithms employed for computing three similarity scores: phonetic, semantic and correspondence-based. Our cognate identification program integrates the three types of evidence using a linear combination of scores. The algorithms are described in detail in (Kondrak, 2002).

The phonetic similarity of lexemes is computed using the ALINE algorithm, which assigns a similarity score to pairs of phonetically-transcribed words on the basis of the decomposition of phonemes into elementary phonetic features. The principal component of ALINE is a function that calculates the similarity of two phonemes that are expressed in terms of about a dozen multi-valued phonetic features. For example, the phoneme n, which is usually described as a *voiced alveolar nasal stop*, has the following feature values: *Place = 0.85*, *Manner = 0.6*, *Voice = 1*, and *Nasal = 1*, with the remaining features set to 0. The numerical feature values reflect the distances between vocal organs during speech production, and are based on experimental measurements. The phonetic features are assigned *salience* weights that express their relative importance. The default salience values were tuned manually on a development set of phoneme-aligned cognate pairs from various related languages. The overall similarity score is the sum of individual similarity scores between pairs of phonemes in an optimal alignment of two words. The similarity value is normalized by the length of the longer word.[1]

For the determination of recurrent sound correspondences we employ the method of inducing a *translation model* between phonemes in two word lists. The idea is to relate recurrent sound correspondences in word lists to translational equivalences in bitexts. The translation model is induced by combining the maximum similarity alignment with the competitive linking algorithm of Melamed (2000). Melamed's approach is based on the *one-to-one* assumption, which implies that every word in the bitext is aligned with at most one word on the other side of the bitext. In the context of the bilingual word lists, the correspondences determined under the *one-to-one* assumption are restricted to link single phonemes to single phonemes. Nevertheless, the method is powerful enough to determine valid correspondences in word lists in which the fraction of cognate pairs is well below 50%.

[1]Another possibility is normalization by the length of the longest alignment (Heeringa et al., 2006).

Because of the lack of a Totonac gold standard, the approach to computing semantic similarity of glosses was much simpler than in (Kondrak, 2002). The keyword selection heuristic was simply to pick the first word of the gloss, which in Spanish glosses is often a noun followed by modifiers. A complete gloss match was given double the weight of a keyword match. More complex semantic relations were not considered. In the future, we plan to utilize a Spanish part-of-speech tagger, and the Spanish portion of the EuroWordNet in order to improve the accuracy of the semantic module.

5 Pairwise Comparison

The first experiment was designed to test the effectiveness of our approach in identifying recurrent correspondences and cognates across a single pair of related languages. The data for the experiment was limited to two noun lists representing Upper Necaxa (2110 lexemes) and Zapotitlán (763 lexemes), which were extracted from the corresponding dictionaries. Both correspondences and cognates were evaluated by one of the authors (Beck), who is an expert on the Totonac-Tepehua language family.

5.1 Identification of correspondences

In the first experiment, our correspondence identification program was applied to Upper Necaxa and Zapotitlán. Simple correspondences were targeted, as complex correspondences do not seem to be very frequent among the Totonac languages. The input for the program was created by extracting all pairs of noun lexemes with identical glosses from the two dictionaries. The resulting list of 865 word pairs was likely to contain more unrelated word pairs than actual cognates.[2]

The results of the experiment were very encouraging. Of the 24 correspondences posited by the program, 22 were judged as completely correct, while the remaining two (**ʧ:ts** and **tɬ:ts**). were judged as "plausible but surprising". Since the program explicitly list the word pairs from which it extracts correspondences, they were available for a more detailed analysis. Of the five pairs containing **ʧ:ts**, one was judged as possibly cognate: Upper Necaxa [**ʧ**astun] and Zapotitlán [aʔa**ts**astun] '*rincón, esquina*'. Both word pairs containing **tɬ:ts** were judged as possibly cognate: [li**tɬ**an]/[li**ts**eχ] '*favor*', and [**tɬ**aq**tɬ**a]/[**ts**a**ts**a] '*elote*'. Both unexpected correspondences were deemed to merit further linguistic investigation.

[2]Some lexemes have multiple glosses, and therefore may participate in several word pairs.

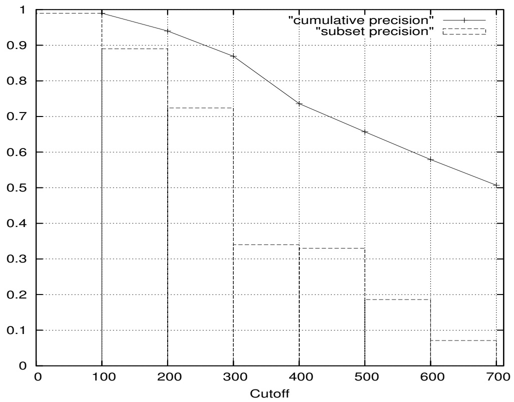

Figure 2: Cognate identification precision on the Totonac test set.

5.2 Identification of cognates

In the second experiment, our cognate identification program was run on the vocabulary lists containing the Upper Necaxa and Zapotitlán nouns. A large list of the candidate word pairs with their glosses was sorted by the total similarity score and evaluated by Beck. The cognation judgments were performed in order, starting from the top of the list, until the proportion of false positives became too high to justify further effort. At any point of the list, we can compute *precision*, which is the ratio of true positives (in this case, cognates) to the sum of true positives and false positives (all word pairs up to that point).

The cognate decisions were based on the following principles. The pairs could be judged as true positives only if the word roots were cognate; sharing an affix was not deemed sufficient. Compound words were counted as cognates if any of the multiple roots were related; for example, both *snowstorm/storm* and *snowstorm/snow* would be acceptable. The rationale is that a person compiling an etymological dictionary would still want to know about such pairs whether or not they are eventually included as entries in the dictionary.

In total, 711 pairs were evaluated, of which 350 were classified as cognate, 351 as unrelated, and 10 as doubtful. 18 of the positive judgments were marked as loans from Spanish. In Figure 2, the boxes correspond to the precision values for the seven sets of 100 candidate pairs each, sorted by score; the curve represents the cumulative precision. For example, the percentage of actual cognates was 86.9% among the first 300 word pairs, and 72.4% among the word pairs numbered 201–300. As can be seen, almost all the pairs in the beginning of the file were cognates, but then the number of false positives increases steadily. In terms of semantic similarity, 30% of the evaluated pairs had at least one gloss in common, and further 7% shared a keyword. Among the pairs judged as cognate, the respective percentages were 49% and 11%.

6 Multiwise comparison

When data from several related languages is available, the challenge is to identify cognate sets across all languages. Our goal was to take a set of diversely formatted dictionaries as input, and generate from them, as automatically as possible, a basic comparative dictionary.

Our system is presented graphically in Figure 3. This system is a suite of Perl scripts and C++ programs. With the exception of the input dictionary converters, the system is language-family independent. With little change, it could be used to determine cognate sets from another language family. In this section, we describe the four stages of the process: preprocessing, identification of cognate pairs, extraction of cognate sets, and postprocessing.

6.1 Preprocessing

The first step is to convert each input dictionary from its original form into a word list in a standardized format. Because of the differences between dictionaries, separate conversion scripts are required for each language. The conversion scripts call on a number of utilities that are maintained in a shared library of functions, which allows for the relatively easy development of new conversion scripts should additional dictionaries become available.

Each line in the resulting language files contains the phonetic form of the lexeme expressed in a uniform encoding, followed a gloss representing the

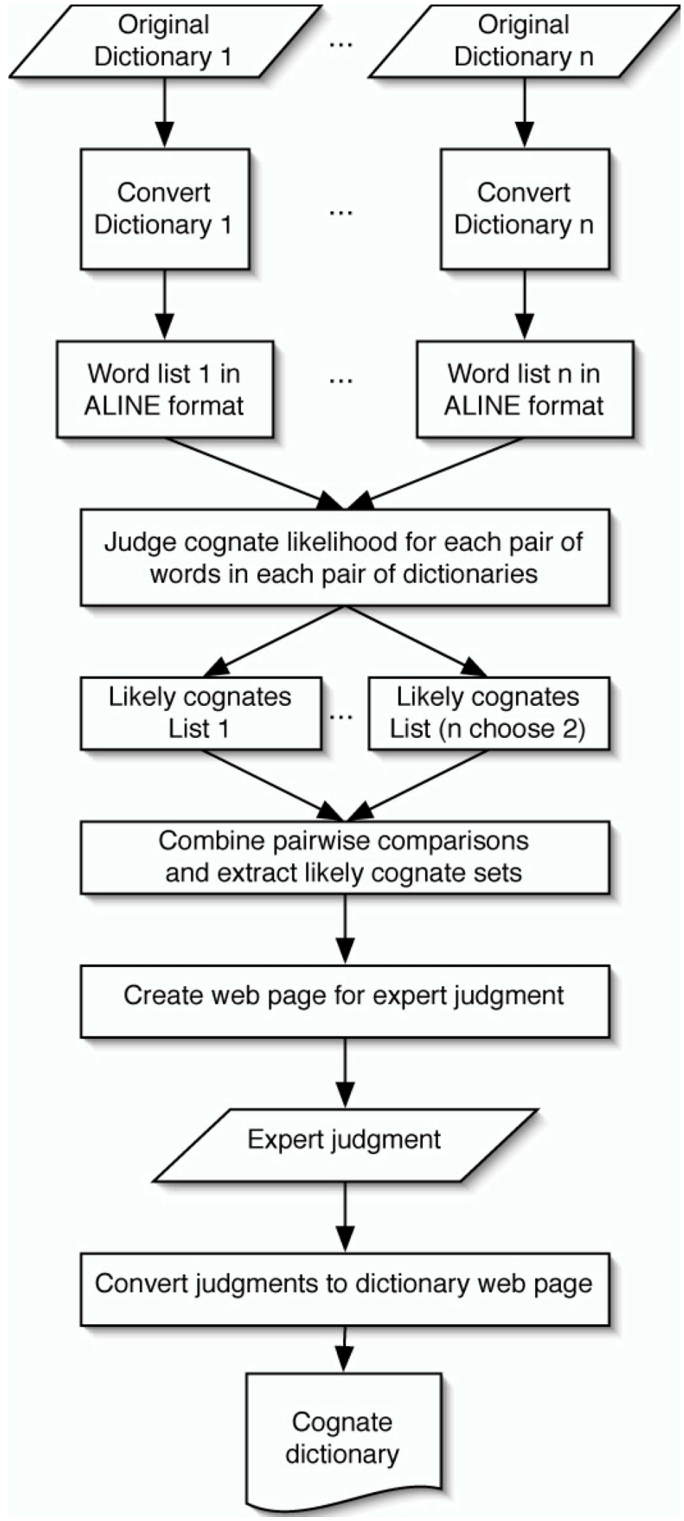

Figure 3: Flowchart illustrating conversion system

meaning of the lexeme. Long glosses are truncated to thirty characters, with sub-glosses separated by semicolons. For the present study, the conversion scripts also removed all dictionary entries that were known not to be nouns.

For the purpose of uniform encoding of phonetic symbols, we adopted the ALINE scheme (Kondrak, 2002), in which every phonetic symbol is represented by a single lowercase letter followed by zero or more uppercase letters. The initial lowercase letter is the base letter most similar to the sound represented by the phonetic symbol. The remaining uppercase letters stand for the phonetic features in which the represented sound differs from the sound defined by the base letter. For example, the phoneme [ʃ], which occurs at the beginning of the word *shy*, is represented by 'sV', where V stands for *palato-alveolar*.

6.2 Identification of cognate pairs

The main C++ program computes the similarity of each pair of words across the two languages using the methods described in Section 4. A batch script runs the comparison program on each pair of the dictionary lists. With n input dictionaries, this entails $\binom{n}{2}$ pairwise comparisons each resulting in a separate list of possible cognate pairs. These lists are then sorted and trimmed to include only those pairs that exceeded a certain similarity threshold.

The batch script has an option of selecting a subset of dictionary pairs to process, which was found useful in several cases. For example, when we discover a newer version of a dictionary, or update an individual dictionary conversion script, only 4, rather than all 10 lists need to be re-generated.

6.3 Extraction of cognate sets

The output from processing individual pairs of word lists must be combined in order to extract cognate sets across all languages. The combination script generates an undirected weighted graph in which each vertex represents a single lexeme. The source language of each lexeme is also stored in each vertex. Links between vertices correspond to possible cognate relationships identified in the previous stage, with the link weights set according to the similarity scores computed by the comparison program.

The algorithm for extracting cognate sets from the graph is the following. First, we find the connected components within the graph by applying the breadth-first search algorithm. The components are added to a queue. For each component in the queue, we exhaustively generate a list of connected subgraphs in which each vertex corresponds to a different source language. (In the present study, the minimum size of a subgraph was set to three, and the maximum size was five, the total number of languages.) If no such subgraphs exist, we discard the component, and process the next component from the queue. Otherwise, the subgraph with the maximum cumulative weight is selected as the most likely cognate set. We remove from the component the vertices corresponding to that cognate set, together with their incident edges, which may cause the component to lose its connectivity. We identify the resulting connected component(s) by breadth-first search, and place them at the end of the queue. We repeat the process until the queue is empty.

Figure 4: A sample judgment screen.

6.4 Postprocessing

The candidate cognate sets extracted in the previous stage are rendered into an HTML page designed to allow an expert linguist to verify their correctness (Figure 4). After the verification, a dictionary composed of the confirmed cognate sets is automatically generated in HTML format, with the glosses restored to their original, untruncated form. Additional cognate sets can be incorporated seamlessly into the existing list. A sample entry in the gener-

317	C	li:qama:n	el juguete; hace burla de el
	T	laaqamaan	el juguete
	S	li:qama:n	el juego; el juguete; lo maltrata; le hace burla
	U	le:ha:ma:n	juguete
	P	li:qama:n	el juguete

Table 2: A sample entry in the generated dictionary.

ated dictionary is shown in Table 2.[3]

6.5 Results

In our initial attempt to extract cognate sets from the graph, we extracted from the graph only those connected components that were complete cliques (i.e., fully connected subgraphs). Of the resulting 120 candidate cognate sets, all but one were confirmed by Beck. The only false positive involved two words that were true cognates, and one word that was morphologically related to the other two. However, although this method was characterized by a very high precision, the overly restrictive clique condition excluded a large number of interesting cognate sets.

In order to improve recall, the method described in Section 6.3 was adopted. 430 possible cognate sets of 3, 4, or 5 words were discovered in this manner. 384 (89%) of these sets were judged to be true cognate sets. Of the remaining 46 sets, 45 contained partial cognate sets. The set that contained no cognate words was composed of three words that share a cognate root, but have different prefixes.

7 Discussion

From a practical standpoint, the procedures used in these experiments provide a powerful tool for the identification of cognate sets and sound correspondences. The identification of these correspondences by traditional means is cumbersome and time-consuming, given the large amounts of data that require processing. The Upper Necaxa dictionary, for instance, contains nearly 9,000 entries, from which a list of about 2,000 nouns would have to be extracted by hand, and then compared pairwise to lists drawn from dictionaries of potentially comparable length of each of the other languages, each of which would also have to be compared to the other. Lists of potential correspondences from each pairwise comparison would then have to be compared, and so on. The algorithms described here accomplish in mere minutes what would take man-hours (perhaps years) of expert labour to accomplish manually, outputting the results in a format that is easily accessed and shared with other researchers as an HTML-format list of cognates that can be made available on the World Wide Web.

The results obtained from a study of this type have important implications for linguists, as well as anthropologists and archeologists interested in the history and migratory patterns of peoples speaking Totonacan languages. Presented with extensive and robust cognate sets and lists of sound changes, linguists gain insight into the patterns of historical phonological change and can verify or disconfirm models of phonological and typological development. These data can also give rough indications of the time-depth of the linguistic family and, potentially, suggest geographical origins of populations. At present, Totonac-Tepehua has not been demonstrably linked to any other language family in Mesoamerica. Careful reconstruction of a protolanguage might reveal such links and, possibly, shed some light on the early movements and origins of Mesoamerican peoples.

These experiments have also allowed us to create the beginnings of an etymological dictionary which will, in turn, allow us to reconstruct a more accurate Totonac-Tepehua family tree. By comparing the relative numbers of shared cognates amongst languages and the number of regular sound changes shared by individual subsets of languages in each cognate set, we hope to be able to determine relative proximity of languages and the order in which the family divided itself into branches, sub-branches, and individual languages. This will shed light on the problem of Totonac-Tepehua origins and migratory patterns, and may help to answer questions about potential links of Totonacan peoples to archeological sites in East-Central Mexico, including the pyramids of Teotihuacán. Accurate determination of distance between variants of Totonacan will also help inform social policy decisions about bilingual education and government funding for language revitalization pro-

[3]The entire dictionary in its current state can be viewed at http://www.cs.ualberta.ca/∼pdilts.

grams, as well as debates about orthographies and language standardization.

Acknowledgements

Thanks to Paulette Levy, James Watters, and Herman Aschmann for sharing their dictionary data. The fieldwork of David Beck was funded by the Social Sciences and Humanities Research Council of Canada and the Wenner-Gren Foundation. Philip Dilts was supported by a scholarship provided by the Government of the Province of Alberta. Grzegorz Kondrak was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Evangelina Arana. 1953. Reconstrucción del prototonaco. Huastecos, totonacos y sus vecinos. *Revista mexicana de estudios antropológicos*, 23:123–130.

Herman P. Aschmann. 1983. *Vocabulario totonaco de la Sierra*. Summer Institute of Linguistics, Mexico.

David Beck. 2005. The Upper Necaxa field project II: the structure and acquisition of an endangered language. Available from http://www.arts.ualberta.ca/˜totonaco.

David Beck. 2006. The emergence of ejective fricatives in Upper Necaxa Totonac. In Robert Kirchner, editor, *University of Alberta Working Papers in Linguistics 1*.

Wilbert Heeringa, Peter Kleiwig, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Proceedings of the COLING-ACL Workshop on Linguistic Distances*, pages 51–62.

John Hewson. 1974. Comparative reconstruction on the computer. In *Proceedings of the 1st International Conference on Historical Linguistics*, pages 191–197.

Alain Ichon. 1973. *La religión de los totonacos de la Sierra*. Instituto Nacional Indigenista, Mexico City.

Mark Johnson. 1985. Computer aids for comparative dictionaries. *Linguistics*, 23(2):285–302.

Grzegorz Kondrak. 2002. *Algorithms for Language Reconstruction*. Ph.D. thesis, University of Toronto.

Paulette Levy. 1987. *Fonología del totonaco de Papantla*. Universidad Nacional Autónoma de México, Veracruz, Mexico.

John B. Lowe and Martine Mazaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics*, 20:381–417.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Michael P. Oakes. 2000. Computer estimation of vocabulary in protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7(3):233–243.

Aileen A. Reid. 1991. *Gramática totonaca de Xicotepec de Juárez, Puebla*. Summer Institute of Linguistics, Mexico City.

García Rojas. 1978. *Dialectología de la zona totonaco-tepehua*. Ph.D. thesis, National School of Anthropology and History, Mexico. Honours thesis.