Dynamic Correspondences: An Object-Oriented Approach to Tracking Sound Reconstructions

Tyler Peterson
University of British Columbia
E270-1866 Main Mall
Vancouver, BC, Canada V6T-1Z1
tylerrp@interchange.ubc.ca

Gessiane Picanço
Universidade Federal do Pará
Belém – Pará – Brasil
CEP 66075-110
picanco.g@hotmail.com

Abstract

This paper reports the results of a research project that experiments with cross-tabulation in aiding phonemic reconstruction. Data from the Tupí stock was used, and three tests were conducted in order to determine the efficacy of this application: the confirmation and challenging of a previously established reconstruction in the family; testing a new reconstruction generated by our model; and testing the upper limit of simultaneous, multiple correspondences across several languages. Our conclusion is that the use of cross tabulations (implemented within a database as *pivot tables*) offers an innovative and effective tool in comparative study and sound reconstruction.

1 Introduction

In the past decade databases have transitioned from a useful resource as a searchable repository of linguistic tokens of some type, to an actual tool capable of not only organising vast amounts of data, but executing complex statistical functions and queries on the data it stores. These advances in database technology complement those made in computational linguistics, and both have recently begun to converge on the domain of comparative and historical linguistic research.

This paper contributes to this line of research through describing the database project *Base de Dados para Estudos Comparativos – Tupí* (BDEC-T) (Database for Comparative Studies – Tupí), which is part of a larger research program investigating the phonemic reconstruction of the Tupí languages. The database component of the BDEC-T is designed to capitalise on the functionality of cross-tabulation tables, commonly known as *pivot tables*, a recent innovation in the implementation SQL queries in many database and spreadsheet applications. Pivot tables can be described as an 'object-oriented' representation of SQL statements in the sense that columns of data are treated as objects, which allow the user to create multidimensional views of the data by 'dragging and dropping' columns into various sorting arrangements. We have found that this dynamic, multidimensional manipulation of the data can greatly aid the researcher in identifying relationships and correspondences that are otherwise difficult to summarize by other query types.

In this paper we report on the results of an experiment that tests the applicability of pivot tables to language data, in particular, the comparative and historical reconstruction of the proto-phonemes in a language family. In doing this, three tests were conducted:

1. The confirmation and challenging of a 'manual' and/or previously established reconstruction of a proto-language, Proto-Tupí;

2. The testing of a new reconstruction generated by our model, and checking it against a manual reconstruction;

3. The testing the upper limit of simultaneous, multiple correspondences across several languages.

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 126–133,
Prague, June 2007.

It is argued that this type of object-oriented implementation of SQL statements using pivot tables, offers two unique features: the first is the ability to check several one-to-one and one-to-many correspondences simultaneously across several languages; and secondly, the ability to dynamically survey the language-internal distribution of segments and their features.

The former feature represents a notable advantage over other 'manual' methods, as the reconstructed forms may be entered in the database as proto-languages, which can be continually revised and tested against all other languages. The latter feature offers the ability to check the language-internal distribution of the (proto-)segments which will aid in preventing possible cases of skewed occurrences, as is shown below. Basic statistical analyses, such as numbers of occurrences, can also be reported, graphed and plotted by the pivot tables, thus providing further details of individual languages and proto-languages, and, ultimately, a more quantitatively reliable analysis.

The net outcome of this is the presentation of a practical methodology that is easily and quickly implementable, and that makes use of a function that many people already have with their database or spreadsheet.

1.1 The Data

The Tupí stock of language families is concentrated in the Amazon river basin of Brazil (and areas of neighbouring countries) and comprises 10 families of languages: Arikém, Awetí, Juruna, Mawé, Mondé, Mundurukú, Puruborá, Ramarama, Tuparí, and Tupí-Guaraní (Rodrigues 1958; revised in Rodrigues 1985), totaling approximately 64 languages. Tupí-Guaraní is the largest family with more than 40 languages, while the other families range from one language (e.g. Awetí, Puruborá) to six languages (e.g. Mondé). From these, the Tupí-Guaraní family is the only family that has been mostly analized from a historical point of view (e.g. Lemle 1971, Jensen 1989, Schleicher 1998, Mello 2000, etc.); there is also a proposal for Proto-Tuparí (Tuparí family), by Moore and Galúcio (1993), and Proto-Mundurukú (Mundurukú family), by Picanço (2005). A preliminary reconstruction at level of the Tupí stock was proposed by Rodrigues (1995), in which he reconstructs a list of 67 items for Proto-Tupí (see further details below). The BDEC-T also includes these reconstructed languages, as they allow us to compare the results obtained from the database with the results of previous, manual historical-comparative studies.

2 The Application: Design and Method

The BDEC-T was initially developed as repository database for language data from various Tupí languages described above, with the purpose of allowing the user to generate lists of word and phoneme correspondences through standard boolean search queries or SQL statements. These lists aided the researcher in exploring different correspondences in the course of a proto-phoneme or word reconstruction. The BDEC-T is implemented within MS Access 2003, which provides the user an interface for entering language data that is then externally linked to tab-delimited text files in order to preserve its declarative format.[1] This also allowed flexibility in accessing the data for whatever purpose in the platform or program of the researcher's choosing.

At present, the BDEC-T for the Tupí stock contains a glossary of 813 words and up to 3,785 entries distributed across 15 Tupían languages. Approximately 18% of this 813-word list appear to have cognates in the majority of languages entered so far, and which can be used as reference for a reliable set of robust cognates across the entire Tupí stock.[2] This number is continually increasing as more languages are entered in the database, and at least 50% of the glossary is filled up for all languages. The average number of entries for each language varies considerably as it depends largely on available sources; yet, in general, the average is of approximately 250 words per language (i.e. about 30%).

[1] The choice of using a proprietary database such as MS Access is mostly a practical one: after considering various factors such as programming, maintenance, distribution and other practical issues, we decided that a database of this type should be useable by researchers with little or no programming experience, as it is fairly easy to learn and modify (see also Bredenkamp, Sadler and Spencer (1998: 149) for similar arguments). It should also be noted that all the procedures outlined here are implementable in open source database and spreadsheet programs such as OpenOffice Calc and Base (vers. 2.3).

[2] There is a separate function in the BDEC-T for assessing and tracking cognates and how they map to semantic sets (see Peterson 2007a for details).

2.1 Data entry and Segmentation

Each of the 65 languages and 4 proto-languages in BDEC-T is associated with its own data entry form. Each data entry form is divided into three main parts:

1. The *word entry fields* where the word for that language is entered (along with two other optional features);
2. The *comparison viewer* that contains fields which simultaneously display that same word in the all the other languages in the database;
3. The *segmentation section* which contains an arrangement fields for recording segment data.

The structure of the stored data is straightforward: the data entered in these forms is stored in a master table where all of the languages are represented as columns. Glosses are the rows, where each gloss is assigned a unique, autogenerated number in the master record when it is entered into the database. This serves as the primary key for all the translations of that gloss across all of the languages.

The third component of the language data entry form, the segmentation section (Fig. 1), contains a linear arrangement of ten columns, S1 to S10, and three rows, each cell of which corresponds to a field. The first row of the ten columns are fields where the user can enter in the segmentation of that particular word, which contains the segments themselves. The second and third rows correspond to optional features (F) that are associated with that segment. In this particular version F1 is unused, while F2 encodes syllable structure (i.e. ‘O’ onset, ‘N’ nucleus).[3]

For example, Figure 1 is a screenshot of a portion of the segmentation section in the language data entry form for Mundurukú. The word being entered is ‘moon’, and the word in Mundurukú is *káʃi*. Segment slots S3 to S6 are used to segment the word.

As a convention, a word will typically be segmented starting with the S3 slot, and not with S1. The reason for this is to allow for at least two segment fields (S1 and S2) to accommodate cognates in

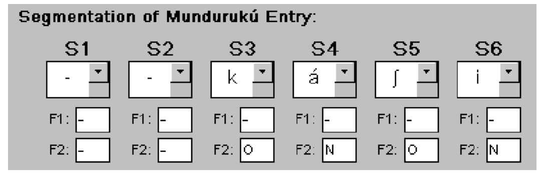

Figure 1: Screenshot of a portion of the Segmentation section in the Mundurukú Data entry form.

Segmentation slot	S1	S2	S3	S4	S5
Avá-Canoeiro			ɨ	t	ɨ
Guajá		w	ɨ	t	ɨ
Araweté	i	w	i	t	ɨ

Table 1: Segmentation of *‘wind’*

other languages that have segments that occur before S3, but are entered into the database at a later time. This is done in order to maintain a consistency between correspondences, regardless of what slot they are in the data base. In other words, we need to be prepared to handle cases that are shown in Tables 1 and 2 above. If the Avá Canoeiro word for ‘wind’ is entered first in Table 1, it is prudent to have segment slots available for languages that are entered later that may have additional segments occurring before. Guajá and Araweté were entered into the database after Avá Canoeiro, and both have additional segments. Keeping S1 and S2 available as a general rule can accommodate these cases.

Our purpose in designing the segmentation component of the form this way was to give the researcher complete control over how words are segmented. This also allows the researcher to cross-check their segmentations in real time with those in the other languages already in the database, which can be done in the comparison viewer (not shown due to space limitations). This is essential for more complicated cases, such as those in Table 2, where there are not only word edge mismatches, but also gaps and (grammaticalized) morphological boundaries that need to be properly corresponded. The significance of this will be demonstrated below.[4]

[3]There is no restriction on the kind of information that can be stored in the two Feature fields. However, in order for them to be useful, they would need to contain a limited set of comparable features across all the languages.

[4]Cases where gaps result in languages already entered would require the user to go back to the other languages entered and re-segment them to include the corresponding gap. This would be the case if *ɨap* was entered without the gap in S3 before the other languages in Table 2. This is facilitated within the database: multiple language forms can be open simultaneously,

Segmentation slot	**S1**	**S2**	**S3**	**S4**	**S5**
Avá-Canoeiro		ɨ		a	p
Guajá		u	ʡ	ɨ	
Mbyá	h –	u	ʔ	ɨ	
Kamayurá	h	ɨ	ʔ	ɨ	p

Table 2: Segmentation of *‘arrow’*

The data entered in the segmentation section of a language’s data entry form is stored in language-specific tables, which has columns for each of the ten segments, and columns recording the two optional features associated with that segment. All of the segment data in the language-specific tables are coordinated by the primary key generated and kept in the master table. The next subsection describes how this segmental data can be used in two specific ways: 1) to track correspondences between languages for a particular cognate or segment slot; and 2), for monitoring the language-internal distribution of segments. We propose that this is achieved through using cross-tabulations of the segment data recorded in each column, and outline a practical implementation of this is using pivot tables.

2.2 Cross-tabulation: ‘Pivot tables’

Access 2003 includes a graphical implementation of SQL statements in the form of cross tabulations, or *pivot tables*, which provide the user an interface with which they can manipulate multiple columns of data to create dynamic, multi-dimensional organizations of the data. There are three basic reasons for organizing data into a pivot table, all of which are relevant to the task at hand: first, to summarize data contained in lengthy lists into a compact format; secondly, to find relationships within that data that are otherwise hard to see because of the amount of detail; and thirdly, to organize the data into a format that is easy to chart. Pivot tables are dynamic because columns of data are treated as objects that can be moved, or literally ‘swapped’ in, out our around in relation to other columns. They are multi-dimensional because column data can be organized along either axis, yielding different ‘snapshots’ of the data. It is this kind of functionality that will be capitalised on in examining correspondences between columns of segment data (S1-10) across any number of languages in the database.

or switched between by the master switchboard.

A cross tabulation displays the joint distribution of two or more variables. They are usually presented as a contingency table which describes the distribution of two or more variables simultaneously. Thus, cross tabulation allows us to examine frequencies of observations that belong to specific categories on more than one variable. By examining these frequencies, we can identify relations between cross-tabulated variables. Typically, only variables with a relatively small number of different meaningful values are cross tabulated. We suggest that phonemes fit this criteria, as there is a finite and relatively low number of total unique phonemes that can ever be potentially cross tabulated.

For example, Figure 2 (below) is a screen shot of a pivot table generated in the BDEC-T that shows the distribution of word and morpheme-initial voiceless stops in Mundurukú in relation to those in the same position for three other languages: Karitiana, Gavião and Karo. This was achieved in the following way: as described above, we assume that the word-initial segment for most words is S3. The S3 column for Mundurukú is then taken to the ‘drop field’ (shaded grey), where all of the values in the S3 of Mundurukú become dependent variables. The S3 columns for Karitiana, Gavião and Karo become independent variables, which allow us to monitor the distribution of voiceless stops in these languages in relation to the S3 segments in Mundurukú. In essence, Mundurukú S3 becomes a sort function on any other S3 columns to the right of it.[5]

Where this method becomes effective is when we ‘swap’ out Mundurukú S3 and replace it with Gavião S3, which is done by pulling the column header and placing it into the grey ‘drop field’. This is shown in Figure 3 below. What Figure 3 immediately demonstrates is the asymmetric correspondence between Mundurukú and Gavião for S3: broadly speaking, the correspondences between Mundurukú and Karitiana, Gavião and Karo are more general, whereas the same correspondences for

[5]Given space considerations, the data in these Tables are just samples - the voiceless stop series was picked from a separate list which acts as a filter on the segments in the Mundurukú S3. Cells where there is a gap ‘-’ do not represent a gap or lack of correspondence, but rather the word for that language possibly hasn’t been segmented yet (gaps are represented by ‘∅’)

M	K	G	K	Mund	Karitia	Gavião	Karo	#	Gloss
k	ŋ	g	n	kíp	ɨita ŋep	git	nãp?	15	piolho
	ø	g	–	káʃi	oti	gát ti	-	44	lua
	ʔ	–	–	kã̰j	eɲ / ʔej	-	ɨ-ganã	69	terra-1
	–	k	–	kədá	-	gakoráá	-	71	procurar
	g/ŋ	g	n	ikopí	ŋop / gop	gap	nãp	12	caba
	ŋ	g	–	ikopí	ŋɨp / gɨp	góôv-aá	-	13	cupim
p	–	b	m	pa̰jbə́	moroɲa	baj	mãjgãra	27	cobra
	–	–	p	pa̰	-	bâbe / ci-pab	pã	37	braço
	–	b	p	pík	peɲetet	õ̋õ̋-baa	pək	60	queimar
	p	–	–	pə	pɨ	-	i-pábe	36	mão
	p	b	p	poʃi	pɨti	õ̋-batĩi	piʔti-rem	6	pesado
t	s	z/s	c	tap	sop	zêp / sép	aʔ-cap	17	pêlo
	–	c	–	tap	-	cap	naʔjop cɨʔ	20	folha-2
	s	z	c	tajʃi	sooɲ	õ̋-zaj / ci-zaj	aʔ-cej	32	esposa
	s	c	c	ti	se	ci	-ci/ɨ	24	líquido
	s	z/s	j	top	ɨ-sɨp	õ̋-zop / ci-sop	ijõ̋m	25	pai-1
	ɲ	j	j	tãj	ɲõɲ	ẽ̋ẽ̋-jẽ̋ẽ̋j	jãj	33	dente
	–	s	c	tap	-	sép	aʔcəp	18	pena
	s	z/s	c	patét	sat	zêt / ci-set	cet	11	nome

Figure 2: Screenshot of a pivot table for voiceless stops in Mundurukú (shaded) corresponding with Karitiana, Gavião and Karo in BDEC-T.

G	M	K	K	Mund	Karitiana	Gavião	Karo	#	Gloss
k	k	–	–	kədá	-	gakoráá	-	71	procurar
	ʃ	k	k	ʃét	kat	két	i-ket	9	dormir
p	–	p	–	ŋə̰bá	paʔep / papɨ	pepó-tẽ̋ẽ̋	-	38	asa
t	tʃ	–	t	tʃó / dʒó	-	ma-tóó	top	47	ver

Figure 3: Screenshot of a pivot table for voiceless stops in Gavião (shaded) corresponding with Mundurukú, Karitiana and Karo in BDEC-T.

Gavião are more restricted.

There is no restriction on the number of independent or dependent variables, and this can be used to investigate the language-internal distribution of segments. Figure 4 shows how the segment data in S3 and S4 from the same language can be used in a pivot table, allowing the user to track the distribution of certain word or morpheme-initial segments and the segments that follow them. This arrangement gives us a snapshot of consonant-vowel patterns in Karo, where S3 has been additionally filtered to show the distribution of vowels that follow the palatals [c] and [j].

One important advantage to this arrangement of data and the use of pivot tables is the potential for tracking multiple correspondences across several languages simultaneously. So far, this is only limited by processor speed and viewing space. We have tested up to five segment correspondences (i.e. S3-8) across three languages, or one correspondence (i.e.

3	4	Karo	Número	Gloss
c	a	aʔ-cap	17	pêlo
		a-capóp	59	cauda-1
	á	cán	49	fogo/lenha
	e	cet	11	nome
	ə	aʔcəp	18	pena
	ej	aʔ-cej	32	esposa
	ĩ	aʔ-cĩn	53	flor
	ɨ	cɨʔ	19	folha-1
	i/ɨ	-ci/ɨ	24	líquido
	Total			
j	a	jate	55	queixada-1
		jaʔo	62	calango-1
		jajo	28	tatu-1
	ã	jãj	33	dente
	o	jokãn	22	tucano
	õ̋	ijõ̋m	25	pai-1

Figure 4: Screenshot of a pivot table for language-internal distribution of [c] and [j] morpheme and syllable-initially in Karo.

S3) for as many as ten languages simultaneously. Given that most words in the Tupí language family have on average three to five segments, the former of these amounts to the ability of corresponding the segments of entire words simultaneously. Considering that any segment column can be swapped in and out dynamically, this adds a substantial amount of power in tracking single correspondences simultaneously across a variety of languages, proto-languages, and potentially even entire families.

Various statistics can be applied to these pivot tables, where the results can be graphed and exported. The analyst may now take these results and proceed with the appropriate detailed investigation, an example of which is presented in the following sections.

3 Proto-Tupí and Mundurukú

To demonstrate the efficacy of this approach, we show now the results obtained with the BDEC-T and the use of pivot tables, and compare them with the results of a previously established set of sound correspondences and reconstructed proto-phonemes. For this, we chose Proto-Tupí, for which Rodrigues (1995) reconstructed 67 lexical proto-forms and established a consonant inventory composed of four complex series of stops, divided into plain, labialized (still uncertain), palatalized, and glottalized (ejectives), shown Table 3.

Rodrigues based his analysis on various synchronic reflexes found in several Tupían languages,

Plain	p	t, ts	ʧ	k
Labialized	(p^{w})	w		(k^{w})
Palatalized		t^{j}		k^{j}
Glottalized	p^{ʔ}, (p^{ʔw})	t^{ʔ}, ts^{ʔ}	ʧ^{ʔ}	k^{ʔ}, (k^{ʔw})

Table 3: Proto-Tupí stop series (Rodrigues 1995)

Rodrigues		**BDEC-T**		**Rodrigues**		**BDEC-T**	
P-T	Mund.	P-T	Mund.	P-T	Mund.	P-T	Mund.
*p	p	*p	p	*ʧ	ʃ	*ʧ	ʃ
			∅		ʧ		ʧ
			ps		ʤ		ʤ
			p/b				
*p^{ʔ}	b	*p^{ʔ}	b	*ʧ^{ʔ}	t	*ʧ^{ʔ}	t
			p		d		d
*t	n	*t	n	*ts	ʤ	*ts	ʤ
			s				ʧ, ʤ
			ʧ		ʃ		ʃ
			t/n				
*t^{ʔ}	d	*t^{ʔ}	d	*ʔ	ʔ	*ʔ	ʔ
			ʤ		∅	*VʔV	V̰
			t/d				
*k	k	*k	k	*k^{ʔ}	ʔ	*k^{ʔ}	ʔ
			ʃ				

Table 4: The correspondence sets as proposed by Rodrigues (1995) compared with those generated by the BDEC-T.

including Mundurukú. Here we compare the correspondence sets postulated by Rodrigues and compare them to those generated by the BDEC-T. The results of the pivot table analysis are shown in Table 4. Note that the BDEC-T predicts a larger set of correspondences than those posited by Rodrigues. However, there are a few cases where both lists agree; for example, for Proto-Tupí *ʧ which corresponds to ʃ, ʧ and ʤ in both cases.

Another important result obtained with the BDEC-T is the possibility of relating other types of segmental information. For example, Mundurukú exhibits a feature that makes it distinct from any other Tupían language: it is the only Tupían language known to make a phonological contrast between modal and creaky (laryngealised) vowels (Picanço 2005). Mundurukú phonation types are crucial for any reconstruction at the stock level – especially in the case of the ejectives proposed by Rodrigues – but this was completely ignored in his proposal. As shown in Table 5 (on the following page), some Proto-Tupí sequences *(C)VʔV yielded (C)V̰ sequences (where the tilde underneath a vowel marks creaky voice on the vowel).

	S1	S2	S3	S4	S5	S6
Proto-Tupí: **upiʔa*	∅	u	p	i	**ʔ**	**a**
Mundurukú: *topsa̰*	t	o	ps	∅	∅	**a̰**
Mekéns: *upia*	∅	u	p	i	∅	**a**

Table 5: *(C)VʔV corresponding with (C)V̰

A comparison that considers only a segment-to-segment correspondence will mistakenly posit the correspondence set *ʔ/∅ for both Mundurukú and Sakirabiá (Mekéns, Tuparí family), when the correspondence is in fact *ʔ/∅ for Sakirabiá but *(C)VʔV/(C)V̰ for Mundurukú. This is true for Rodrigues' analysis, which mistakenly established that "in Mundurukú [the glottal stop] has dropped" (1995: 6). The BDEC-T, on the other hand, allows us to compare features to segments, and to examine various correspondences of segments in a sequence. This is a particular advantage as there will be no missing information. With this, this unique property of Mundurukú, specifically creaky voice, can be explained historically in a principled way.

3.1 Language-internal distribution

A major feature offered by the BDEC-T is the possibility of examining the distribution of segments within the same language, which allow us to better capture the proper environment for correspondences between languages. As Picanço (2005) notes, phonotactic restrictions may, in many cases, be gaps left behind by historical changes. Table 6 provides an example of the distribution of the pairs plain-glottalized stops. At least in the case of *p versus *p^{ʔ}, the only occurrences of the latter is preceding the high central vowel *ɨ; in this environment, both consonants appear to contrast as *p also occurs before *ɨ. In the case of the coronal pairs *t/*t^{ʔ} and *ʧ/*ʧ^{ʔ}, there is no occurrence of the first pair before *ɨ, whereas *ʧ/*ʧ^{ʔ} occur mostly in this environment. As for *ts versus *ts^{ʔ}, these also appear to be in complementary distribution. By using

p	e	pˀ	ɨ	t	ã	tˀ	a
	ɨ				a		i
	i				ĩ		u
	o				ũ		
					u		
tʃ	ɨ	tʃˀ	ɨ	ts	u	tsˀ	a
			a		ɨ		

Table 6: Language-internal distribution of segments

pivot tables, the analyst is able to easily monitor and track distributional gaps or contrasts and so provide a more systematic diachronic analysis.

Another case which illustrates the applicability of pivot tables in arranging segment data concerns the vowels. Rodrigues' comparison produced vowel correspondences between Proto-Tupí and Mundurukú. Again we compare his findings with those detected by the database: Table 7 compares the oral vowel correspondences as in Rodrigues (1995) with those obtained by the pivot tables in the BDEC-T, supplemented by the total of words with the respective correspondence.

In Rodrigues' analysis, the correspondences between proto-Tupí oral vowels and their reflexes in Mundurukú are straightforward: it is a one-to-one correspondence. BDEC-T, however, challenges this analysis as there appear to be other correspondences that have not been observed, with the exception of the correspondence set *e/e, where both methods achieved the same results. Rodrigues' intuitions are, nonetheless, relatively close to what the database produced: the largest number of correspondences match the ones posited by Rodrigues, indicating that a 'manual' analysis, although valid, still has the potential to miss details that the database captures.

In sum, we employed the function of cross tabulations in the form of pivot tables to arrange segmented data. The object oriented function of pivot tables allowed us to dynamically arrange segment data which aided in tracking phonemic and featural correspondences. This was tested against a manual analysis of the data and it was shown to confirm, revise and produce new results.

Rodrigues		**BDEC-T**		
P-T	Mundurukú	P-T	Mundurukú	Total
		∅	a	1
*a	a	*a	∅	1
			a	11
			ə̃	1
			õ	1
			a̰	2
*e	e	*e	e	5
*i	i	*i	i	2
			∅	2
*ɨ	i	*ɨ	ə	1
			i	19
			ḭ	3
			j	1
*o	ɨ	*o	∅	1
			ə́/ə	1
			o	2
*u	o	*u	o	7
			õ	1
			i	1

Table 7: Rodrigues' (1995) oral vowel correspondence sets compared with those generated by the BDEC-T.

4 Conclusion

The use of spreadsheets and databases is well-established in linguistic research. However, as far as we know, the BDEC-T represents the first attempt at harnessing the functionality of pivot tables and cross-tabulation in historical linguistics. On this note, the application computational procedures in the study of sound change and comparison have made notable advances in the past decade. Relevant to this study, systems such ALINE, a feature-based algorithm for measuring phonetic similarity, are capable of automating segmentation and quantitatively calculating cognate probabilities without resorting to a table of systematic sound correspondences (Kondrak 2002). These are valuable models which test many long-standing hypotheses on the nature of sound change and methods for investigating this. While not offering an automated algorithm of this type, we chose to keep segmentation manual in order to maintain accuracy and to make adjustments where needed in the S1-S10 segmentations made in the languages. This also offers a measure of accuracy, as the pivot tables will only yield invalid results if the segments aren't aligned properly.[6]

Although not discussed in this paper, we have promising results from using the optional feature fields (F1 and F2) to generate syllable template to accompany the phonemic correspondences generated by the pivot tables. Also, the application of pivot tables in the BDEC-T has also had success in tabulating mappings between cognate and semantic sets in the Tupían languages (Peterson 2007a). Ultimately, we would like to explore innovative visualizing techniques to display the interdependent relationships between phonemes at various stages of reconstruction (through the proto-languages in the database), and the languages whose inventories they belong to. Conceptually, this would give us a (scalable) two- or three-dimensional plots or 'webs' of correspondences across the languages, perhaps implemented by recent visualization techniques such as treemaps or ConeTrees (Fekete & Plaisant 2002).

The purpose of the BDEC-T is ultimately to complement other current computational approaches to the domain of historical and comparative research by offering a practical level of interactivity and productivity in a research tool. Where automation is not necessary, the BDEC-T offers a database model that effectively enhances the functionality of the kinds of databases that are already widely used.

References

Andrew Bredenkamp, Louisa Sadler and Andrew Spencer. 1998. Investigating Argument Structure: The Russian Nominalization Database. *Linguistic Databases*, John Nerbonne, (ed.) CSLI Publications

Jean-Daniel Fekete and Catherine Plaisant. 2002. Interactive Information Visualization of a Million Items. *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Wash., DC

Cheryl Jensen. 1989. O desenvolvimento histórico da língua Wayampí. Master's Thesis. Campinas: Universidade Estadual de Campinas.

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

[6]We have developed a set of 'diagnostic' pivot tables to help control against improperly aligned segmentations.

Mirian Lemle. 1971. Internal classification of the Tup-Guaran linguistic family. *Tupi Studies I.*, David Bendor-Samuel (ed.), pp. 107-129. Norman: SIL

Augusto S. Mello. 2000. Estudo Histórico da Família lingüística Tup-Guaraní: Aspectos Fonológicos e Lexicais. PhD Dissertation. Santa Catarina: UFSC

Denny Moore and Vilacy Galúcio. 2005. Reconstruction of Proto-Tupari consonants and vowels. in *Survey of California and Other Indian Languages, Report 8*, M. Langdon and L. Hinton (eds.), pp. 119-137.

John Nerbonne. 1998. *Linguistic Databases: Introduction.* John Nerbonne, (ed.) CSLI Publications

Tyler Peterson. 2007a. Analytical Database Design: Approaches in the Mapping between Cognate and Semantic Sets. *Proceedings of the 7th Intl. Workshop on Computational Semantics*, J. Goertzen *et al* (eds). Tilburg: Tilburg University, pp. 359–361.

Gessiane L. Picanço. 2005. Mundurukú: Phonetics, Phonology, Synchrony, Diachrony. PhD Dissertation. Vancouver: University of British Columbia.

Aryon D. Rodrigues. 1958. Die Klassifikation des Tupi-Sprachstammes. *Proceedings of the 32nd International Congress of Americanists*, Copenhagen, 1956; pp. 679-684.

Aryon D. Rodrigues. 1985. Relações internas na família lingüística Tup-Guaraní. *Revista de Antropologia 27/28*, São Paulo, 1956 pp. 33-53.

Aryon D. Rodrigues. 1995. Glottalized stops in Proto-Tupí. Paper presented at the SSILA Summer Meeting, University of New Mexico, Albuquerque, NM.

Charles O. Schleicher. 1998. Comparative and Internal Reconstruction of Proto-Tupí-Guaraní. PhD Dissertation. Madison: University of Wisconsin.