August 30, 2007

AUTOMATED CLASSIFICATION OF THE WORLD'S LANGUAGES: A DESCRIPTION OF THE METHOD AND PRELIMINARY RESULTS

Cecil H. Brown
Northern Illinois University

Eric W. Holman
University of California, Los Angeles

Søren Wichmann
Max Planck Institute for Evolutionary Anthropology & Leiden University

Viveka Velupillai
Institut für Anglistik, Justus-Liebig-Universität Gießen

Abstract

An approach to the classification of languages through automated lexical comparison is described. This method produces near-expert classifications. At the core of the approach is the Automated Similarity Judgment Program (ASJP). ASJP is applied to 100-item lists of core vocabulary from 245 globally distributed languages. The output is 29,890 lexical similarity percentages for the same number of paired languages. Percentages are used as a database in a program designed originally for generating phylogenetic trees in biology. This program yields branching structures (ASJP trees) reflecting the lexical similarity of languages. ASJP trees for languages of the sample spoken in Middle America and South America show that the method is capable of grouping together on distinct branches languages of non-controversial genetic groups. In addition, ASJP sub-branching for each of nine respective genetic groups—Mayan, Mixe-Zoque, Otomanguean, Huitotoan-Ocaina, Tacanan, Chocoan, Muskogean, Indo-European, and Austro-Asiatic—agrees substantially with subgrouping for those groups produced by expert historical linguists. ASJP can be applied, among many other uses, to search for possible relationships among languages heretofore not observed or only provisionally recognized. Preliminary ASJP analysis reveals several such possible relationships for languages of Middle America and South America. Expanding the ASJP database to all of the world's languages for which 100-word lists can be assembled is a realistic goal that could be achieved in a relatively short period of time, maybe one year or even less.

1. Introduction.

We have developed a relatively uncomplicated computerized method for producing near-expert classification of the world's languages. It entails an Automated Similarity Judgment Program (ASJP) which compares languages by pairs for lexical similarity. ASJP yields for each pair of compared languages a Lexical Similarity Percentage (LSP). LSP is the number of items on a list of meanings (Swadesh's 100-item list) for which two compared languages have words that are judged phonologically similar by ASJP, divided by the number of meanings on the list for which both of the languages have words, the result multiplied by 100. LSP is then corrected for factors extraneous to the meaning of the words, and the result is called SSP (see section 3.4).

SSPs generated for a set of languages by ASJP constitute a database for producing computer-generated branching structures for languages similar to language-family trees manually produced by historical linguists.

ASJP requires only minutes to yield SSPs for large numbers of paired languages. At present we have developed 100-item word lists for 245 globally distributed languages. Using this dataset, ASJP compares approximately 3,000,000 word pairs from 29,890 pairs of languages, yielding 29,890 SSPs. From a technological point of view, there is no limit on the number of languages that could be rapidly compared through use of ASJP. An ultimate goal of this project is nothing less than comparison of all the world's languages for which it is possible to obtain 100-item lists. A conservative estimate is that at least 2,500 of the estimated 6,000 languages of the world are well-enough recorded that such lists may be readily assembled for them. ASJP applied to 2,500 languages would produce around 3,125,000 SSPs.

The 29,890 SSPs generated thus far have been used as a database in computer programs designed originally for biologists for the derivation of phylogenetic trees based on genetic data. Trees so produced using SSP data graphically reflect the lexical similarity of languages as judged by ASJP. Languages on the same branch in a tree are more lexically similar than languages on different branches. We have compared ASJP trees to classifications of known genetically related languages produced by expert historical linguists (section 4.). Typically, ASJP trees and expert classifications are in substantial agreement.

The most time-consuming aspect of this project is the assembly of 100-item lists for languages (section 3.1). These must be extracted manually from dictionaries, vocabularies, and word lists. Fortunately, many such lists already exist and are readily available on the internet and from other sources as well. In assembling the 245 lists of our current sample, we have drawn heavily on these pre-existing lists, perhaps coming close to exhausting them. The next phase of data acquisition for this project will mainly entail manual production of new lists from published and unpublished lexical sources.

Once 100-item lists have been assembled, it is necessary to convert words on them into a single, standard orthography. Automated lexical comparison would be impossible if each list employed in comparison were to involve a different or even only a slightly different orthography. We have developed an ASJP orthography into which all original lexical lists are converted (3.2). The ASJP orthography can be viewed as a very simplified version of the International Phonetic Alphabet (IPA). A major feature this standard orthography is that it entails only symbols found on the common QWERTY keyboard for English. Words on lists in a given orthography can typically be converted into the ASJP orthography in a short period of time, usually less than an hour, by a trained transcriber.

As mentioned above and demonstrated presently (4.), ASJP trees agree closely with classifications for known genetic groupings of languages produced by experts. A potential use for ASJP analysis is the identification of language relationships heretofore not observed or only provisionally recognized. In addition, considerable interest has been shown recently in the automated analysis of structural features of language for historical linguistic interpretation (Dunn et al. 2005, Wichmann and Saunders 2007). The automated analysis of the lexicon provided by ASJP, combined with that of structural features, may prove to be an extraordinarily powerful investigatory tool for historical linguistics.

2. Background.

Some preceding proposals are pertinent to the development of computer applications capable of language classification. A method devised by Oswalt (1970) early in the era of accessible electronic computers is similar to our strategy. In Oswalt's approach, languages are compared by pairs and the Swadesh 100-item list is used. Oswalt integrates instructions into his computer program detailing which phonological segments are to be judged similar. If a specific number of segments in a pair of words are judged similar, the words are judged similar. These instructions appear to be considerably more numerous and complex than the instructions pertaining to ASJP (see Appendix D). Oswalt also calculates as a random baseline the similarity between words in different positions on the list (which have different meanings), and then compares the baseline to the similarity score for two languages involving words with the same meaning in order to determine the statistical significance of the similarity score. Oswalt's approach is applied to seven Indo-European languages and to Finnish. Results appear to conform to expert classification. To our knowledge, this approach has not been applied to a larger corpus of languages. Despite the 37 years that have passed since the description of Oswalt's method, it has received scant attention, although Kessler (2001:33) notes that it has been used by Villemin (1983) in a test for connections involving Japanese, Korean, and Ainu.

Oswalt's work could have been inspired by Swadesh himself, who, as early as around 1960, instigated a project for computer-automated comparisons of wordlists. The method, which was presented to students at the Universidad Nacional Autónoma de México and, in 1962, at the Seattle Linguistics Institute, consisted in reducing articulatorily related phonological units to machine-readable symbols in order that a computer could calculate a measure of lexical similarity among languages (Terrence Kaufman and Nicholas Hopkins, personal communication). Apparently this project was never fully realized.

Most research on computerized language classification has followed an approach different from that of Swadesh and Oswalt. The usual method for determining word similarity does not involve programmed instructions setting forth phonological segments

that are to be judged similar. Instead, the computer searches for phonological recurrences involving segments that occur in the same position within words denoting the same Swadesh-list item. A relationship between two languages is inferred when those recurrences exceed chance expectations according to some statistical criterion. Guy (1980) provides the first example of this approach. He uses it to classify as many as 41 languages, but the results are not generally consistent with expert classification except for closely related languages. Ringe (1992) sets a higher standard of statistical rigor, and Kessler (2001) adds further statistical improvements. Goh (2001) increases the power of the test by requiring more than one recurrence in a word. These later authors confine their attention to sets of eight or fewer languages, and the classificatory results appear to match those of experts. To our knowledge, this sort of method has yet to be successfully applied to a sample of languages that exceeds or even approaches the present sample in size. Indeed, it is unclear how amenable this approach may be to dealing with very large language samples.

Kondrak (2003a,b), Inkpen et al. (2005), MacKay and Kondrak (2005), and Kondrak and Sherif (2006) have recently embarked on an ambitious program of applying methods developed in computer science to the discovery of phonological recurrences and the identification of cognates. They have compared a variety of methods, and report better success for some of their methods than for conventional linguistic techniques. They have also achieved progress on some problems not previously attacked, such as finding recurrences that involve several segments at once, and distinguishing cognates from loans. So far, however, they have not taken the next step of applying their methods to language classification.

Another recent approach is used by Nakhleh et al. (2005), which builds on Ringe et al. (2002) to flesh out details of Indo-European language relatedness. The algorithm employed is designed to draw upon findings regarding shared innovations among Indo-European languages which have required years of in-depth study to work out. Because of the algorithm's reliance on highly language-family-specific information, it cannot readily be applied to other language families. Unlike ASJP, the method of Nakhleh et al. (2005), then, is not now appropriate for general language classification.

3. Method.

3.1 100-item list.

The method described here is a lexicostatistical approach. Lexicostatistics developed in conjunction with glottochronology as worked out first by Morris Swadesh in the mid-20th century (Swadesh 1951, 1971). Swadesh devised a list of 100 glosses or meanings, among other longer lists, to be used in comparative analysis to determine how far back in time two genetically related languages began to diverge from one another. This list consists of so-called "core vocabulary," words for things common to

the environments of all humans such as body parts, colors, and natural objects (water, stone, clouds, sun, and so on). Words for core vocabulary are thought to be more resistant to change than words for other vocabulary. In its most well-known embodiment, lexicostatistics uses Swadesh lists as a basis for making lexical comparisons suggesting degrees of language relationship such as, for example, would be illustrated by a family tree model of language affiliation. ASJP uses the Swadesh 100-item list (see Appendix A) for a similar purpose. To date, we have assembled Swadesh 100-item lists for 245 languages for ASJP analysis (see Appendix B for a list of these languages).

3.2 ASJP orthography.

ASJP orthography consists of 41 symbols (representing 7 vowels and 34 consonants), all found on the standard QWERTY keyboard (See Appendix C for a full description of ASJP orthography). Some symbols of ASJP orthography, like those of IPA, represent only one sound, e.g., N = velar nasal (IPA: ŋ). Some single ASJP symbols represent sounds designated by combined symbols in IPA, e.g., C = voiceless palato-alveolar affricate (IPA: tʃ). Unlike IPA, some ASJP symbols can represent more than one sound, e.g., c = both the voiceless alveolar fricative (IPA: ts) and the voiced alveolar fricative (IPA: dz). Some symbols are cover symbols for a relatively broad range of sounds, usually including those occurring rarely in languages. For example, L is used to represent all laterals other than normal l (the voiced alveolar lateral approximate). The symbols used for vocalic sounds cover broad ranges. For example, the symbols a and 3 together can represent all central vowels, with a restricted to the low central vowel and 3 covering all other central vowels.

ASJP orthography is designed to represent all the commonly occurring sounds of the world's languages. Occasionally, rare sounds are encountered in languages not explicitly identified in the orthography. Such a sound is represented by a symbol in the orthography that identifies the sound that is closest to the rare sound in place and manner of articulation. For example, S, which represents the voiceless palato-alveolar fricative (IPA: ʃ), can be used to designate the relatively rarely occurring retroflexed palato-alveolar fricative (IPA: ʂ).

3.3 Instructions for automated similarity judgment.

At the core of ASJP is a set of instructions integrated into the program detailing when two words showing identical symbols (in the ASJP orthography) are judged to be lexically similar. Basically, at least two symbols found in a single syllable of one word for a specific referent must be identical respectively to at least two symbols found in a single syllable of a word for the same referent found in another language, in order for the two words to be judged lexically similar to one another. The following are two examples of ASJP instructions: (1) If language A has a word for "dog" containing the syllable C_1VC_2 (where C = consonant and V = vowel), and language B has a word for

"dog" containing the syllable C_1VC_2, then the words are automatically judged to be lexically similar. Such judgments are order sensitive such that C_1VC_2 would not be regarded similar to C_2VC_1. Note that in this example, the vowels of the syllable nuclei do not have to be judged similar for the respective words to be judged similar. (2) If language A has a monosyllabic word for "moon" with the syllable being C_1V_1, and language B has a word for "moon" (monosyllabic or polysyllabic) containing the syllable C_1V_1, these words are automatically judged to be lexically similar. In this case, identity of the vowels as well as the consonants is required for the words to be judged lexically similar. For full instruction details, see Appendix D.

ASJP differs from previous approaches in requiring symbols to be identical, rather than allowing correspondences between non-identical symbols. We have explored the use of correspondences defined by phonetic similarity such as in Oswalt (1970), and also correspondences inferred from empirical recurrences such as in later work starting with Guy (1980) (see section 2.). These alternatives have not improved the agreement of ASJP classifications with expert classifications.

3.4 The Lexical Similarity Percentage (LSP) and the Subtracted Similarity Percentage (SSP) derived from it.

As noted in 1., LSP is the number of items on the 100-item list for which two compared languages have words that are judged phonologically similar by ASJP, divided by the number of meanings on the list for which both of the languages have words, the result multiplied by 100. For example, languages A and B have words for the same 95 items of the 100 items on the Swadesh list (they do not both have words for five of the items). Of these 95 pairs, 30 are judged to be similar by ASJP. This yields a LSP of 31.6 (30/95 X 100) for the A/B comparison.

LSPs for large numbers of language pairs can be used as a database to create trees graphically representing the relatedness of those languages. We would like to assume that the relatedness represented in such trees reflects mostly the genetic affiliation of languages rather than other factors such as language contact or chance. However, this assumption is clearly not warranted. For example, longer words will provide for more spurious matches and lead to inflated LSPs for language pairs. Another, clearly more serious problem, is that if languages happen to have similar phonologies (phoneme inventories, phonotactics), LSPs will also be affected positively. Similar phonologies may be due to chance, diffusion, or genetic relatedness. Whatever the explanation, we would like for ASJP results to reflect *lexical similarity* rather than *phonological similarity*.

To compensate for the effects of word length and phonological similarity, we compute a Phonological Similarity Percentage (PSP). This is defined as the average similarity (calculated as for LSP) among pairs of words that do not refer to the same concept on the 100-item list. For a full 100-item list this involves comparing the

(100×99/2 =) 4,950 word pairs that are semantically different. PSP is similar to the random baseline used by Oswalt (1970), except that the power of modern computers allows the inclusion of all pairs of semantically different words rather than just a sample as used by Oswalt. An adjusted lexical similarity index is then computed as the Subtracted Similarity Percentage (SSP = LSP - PSP). For example, the PSP calculated electronically for the language A and B comparison mentioned above is 28.1. Subtracting the latter number from the LSP for the pair (31.6) yields a SSP for A/B of 3.46. For generating trees of language relatedness, SSP is used here rather than LSP. In general, we have found that SSPs yield trees closer to expert classification than those produced by LSPs.

3.5 Generating ASJP trees.

SSPs produced by ASJP serve as the database for generating trees of language relatedness through use of computer programs designed originally for producing phylogenetic trees in biology. A number of such programs are available. The software used here is SplitTrees4, specifically, the Neighbor-Joining algorithm (Huson 1998; Huson and Bryant 2006; cf., McMahon and McMahon [2005] for discussion of the algorithm).

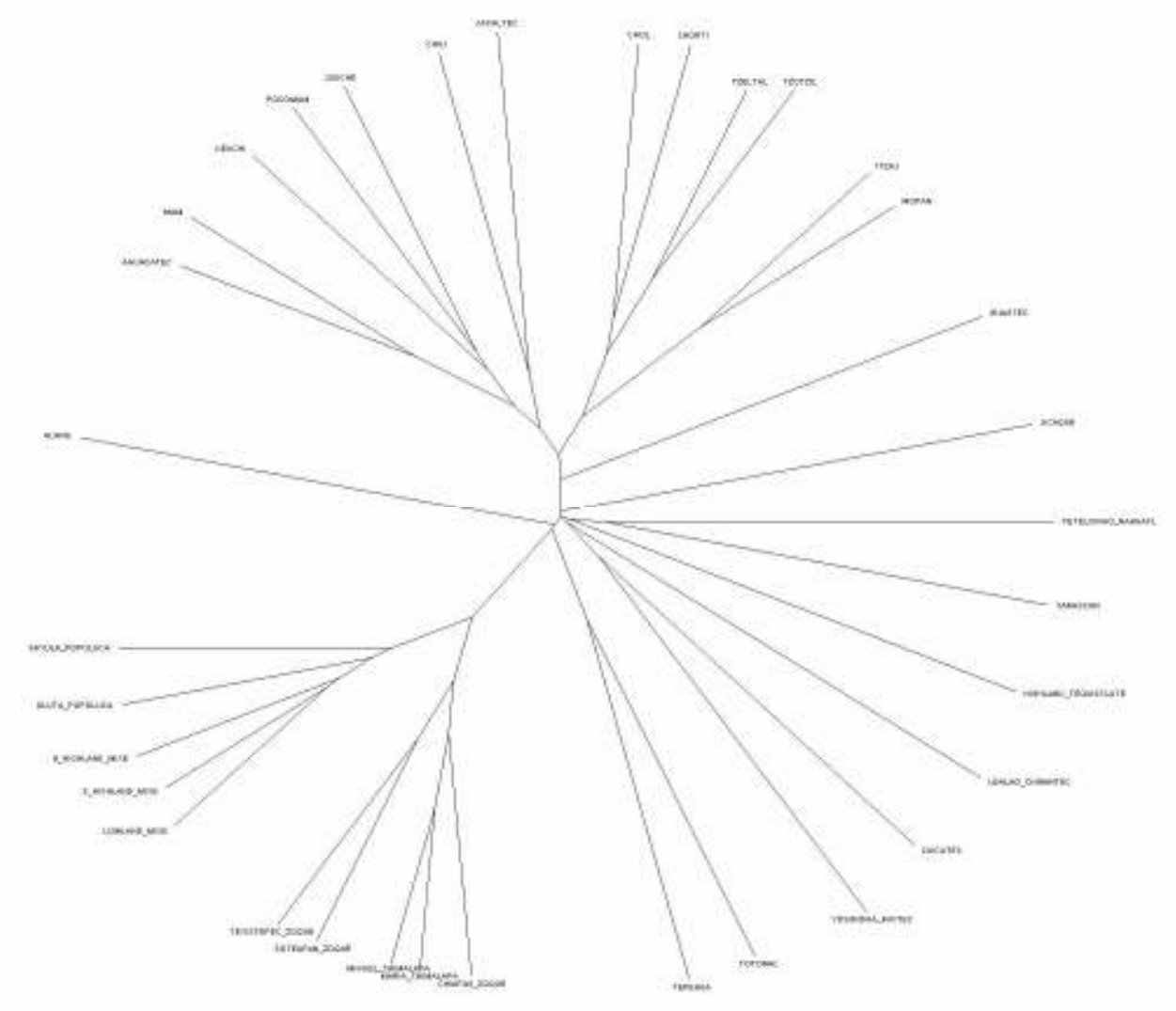
Employing Neighbor-Joining, we have produced a huge language-relatedness tree for all 245 languages of our current sample through use of a database consisting of 29,890 SSPs. Physical constraints on legibility preclude presentation of this tree here (the authors will send an electronic file containing the tree if requested). For illustrative purposes, we provide two smaller trees treating the 34 Middle American languages of our sample and the 52 South American languages. Table 1 shows a portion of the large matrix of SSPs generated by ASJP for all possible pairs of the 34 Middle American languages. Figure 1 shows the resulting tree when the matrix data are processed by the Neighbor-Joining program.

Table 1. A portion of the large SSP matrix for all possible pairs of the 34 Middle American languages of the sample.

SSP	TETE	LEAL	CUIC	YOSO	TARA	XICO	TEPE	CHIA	MARI	MIGU	LOWL	N_HI	S_HI	OLUT	SAYU	SOTE	TEXI...
TETELCINGO_NAHUATL																	
LEALAO_CHINANTEC	.49																
CUICATEC	.82	3.05															
YOSONDUA_MIXTEC	-1.18	-.61	10.90														
TARASCAN	8.85	.03	1.70	-3.96													
XICOTOPEC_TOTONAC	-.11	.32	-4.51	-.08	.49												
TEPEHUA	.40	-.29	2.80	1.50	-.90	23.58											
CHIAPAS_ZOQUE	-.25	.14	.70	-1.24	1.93	3.30	9.97										
MARIA_CHIMALAPA	.79	-1.17	.61	-2.28	1.99	4.78	9.31	51.14									
MIGUEL_CHIMALAPA	-.31	-2.67	-.54	-2.21	.86	3.21	5.84	50.82	67.11								
LOWLAND_MIXE	1.40	-1.29	-1.89	-1.00	1.20	.59	1.62	27.04	28.77	29.93							
N_HIGHLAND_MIXE	1.51	-2.10	-.59	-.88	.83	.14	3.70	25.01	28.96	30.25	52.73						
S_HIGHLAND_MIXE	2.35	-2.04	1.72	.39	1.86	4.11	7.57	28.81	33.26	35.18	59.83	59.87					
OLUTA_POPOLUCA	.71	-.09	-.58	.94	.89	3.10	2.68	24.74	24.88	26.09	43.19	52.25	46.42				
SAYULA_POPOLUCA	.67	.11	1.68	-1.12	-.59	.44	3.02	31.44	33.74	32.88	44.93	49.47	40.39	42.70			
SOTEAPAN_ZOQUE	-1.14	-2.36	.39	.32	3.28	2.46	2.03	38.71	37.37	42.48	28.05	19.64	27.10	24.29	27.31		
TEXISTEPEC_ZOQUE	-.57	-.71	-.36	.21	2.09	.58	3.27	40.71	39.54	40.58	18.39	18.99	25.01	19.91	22.50	53.81	
HUAVE	-.36	-.10	-.25	-.69	1.19	1.48	4.92	-.82	-.59	2.68	.72	.58	.50	1.30	1.73	.52	.73...
AGUACATEC	-1.45	.47	2.23	1.47	1.04	1.97	1.83	.01	-1.34	-1.45	-.48	-.98	-.18	.85	.55	.04	-1.57...
CHOL	-2.50	.84	-.94	-.93	-1.85	2.60	-.17	-1.49	-.07	.91	-1.00	-1.52	.52	-.81	-.86	1.00	-.16...
CHORTI	-1.01	.38	2.80	.23	1.79	.43	1.16	-1.52	-.34	-1.54	-1.90	-.48	-1.70	.32	1.22	-1.21	-1.68...
CHUJ	.33	.56	-.72	1.83	-.02	-.04	-.38	-.43	.73	-.51	-1.32	-.42	2.40	.32	-.87	-.28	.29...

HUASTEC	-.66	-1.05	-.86	.68	-.39	3.00	3.54	-1.31	-1.23	-.31	-.97	-.44	-1.59	.47	.22	.93	-.09...
ITZAJ	-.25	1.30	.73	2.44	.01	-1.07	.98	-.82	.56	-.98	-.36	.28	.00	1.04	-.19	.39	.96...
JACALTEC	-2.48	.60	.18	.15	-1.30	.74	-1.57	-.77	-1.09	-1.06	-.25	-.02	.94	-1.22	-1.37	1.36	-.78...
KEKCHI	-.21	-.36	-.24	3.10	1.21	.23	.74	-.18	.00	-.25	1.48	1.81	1.67	1.47	1.54	1.03	.87...
MAM	2.84	-1.39	-.88	-1.04	-.15	.35	-2.17	-.20	-.70	.28	3.04	1.82	3.39	.51	.97	1.96	-1.57...
MOPAN	.09	2.68	-1.07	-.48	.39	.97	-1.01	-.32	2.08	-.45	.70	.53	.36	.64	.20	.72	.94...
POCOMAM	3.98	1.49	1.23	-.40	1.65	.96	1.43	.19	1.46	1.94	.92	2.10	3.03	3.14	.89	.33	.99...
QUICHE	.43	-.02	-.43	-.53	1.52	2.37	2.26	-.83	-.90	.08	-.10	1.05	2.10	.88	.80	.23	-.96...
TZELTAL	1.05	-.58	-.29	.58	-.13	.90	.54	-.42	.75	1.53	.88	-.52	1.23	2.26	1.09	2.05	-.14...
ZINACANTAN_TZOTZIL	1.29	-1.05	-1.65	-2.98	-.22	-1.77	2.31	.68	2.00	1.85	-.82	-.68	.26	-.93	2.23	-.31	-.52...
HIGHLAND_TEQUISTLATE	2.67	-.59	4.46	-1.21	1.64	.93	-2.67	-1.37	-1.50	.33	-.13	.70	.34	.71	.24	-1.62	-1.36...
JICAQUE	.77	.44	2.34	-1.00	-.50	2.91	.48	.95	-1.27	2.51	-1.28	-1.11	3.01	-2.36	.08	-1.54	.06...
	TETE	LEAL	CUIC	YOSO	TARA	XICO	TEPE	CHIA	MARI	MIGU	LOWL	N_HI	S_HI	OLUT	SAYU	SOTE	TEXI...

Figure 1. Automated Similarity Judgment Program (ASJP) tree for Mesoamerican languages.



The tree of Figure 1 groups together all languages of the Middle American sample belonging to the same language genetic groups. These genetic groups, all of which are *non-controversial*, and languages of the sample affiliated with them are:

OTOMANGUEAN: Chinantec, Cuicatec, Mixtec.
TOTONACAN: Totonac, Tepehua.
MIXE-ZOQUE: Chiapas Zoque, Santa Maria Chimalapa Zoque, San Miguel Chimalapa Zoque, Soteapan Zoque, Texistepec Zoque, Lowland Mixe, South Highland Mixe, North Higland Mixe, Oluta Popoluca, Sayula Popoluca.
MAYAN: Aguacatec, Mam, Kekchi, Pocomam, Quiche, Chuj, Jacaltec, Chol, Chorti, Tzeltal, Tzotzil, Itzaj, Mopan, Huastec.

Figure 2. Automated Similarity Judgment Program (ASJP) tree for South American languages.

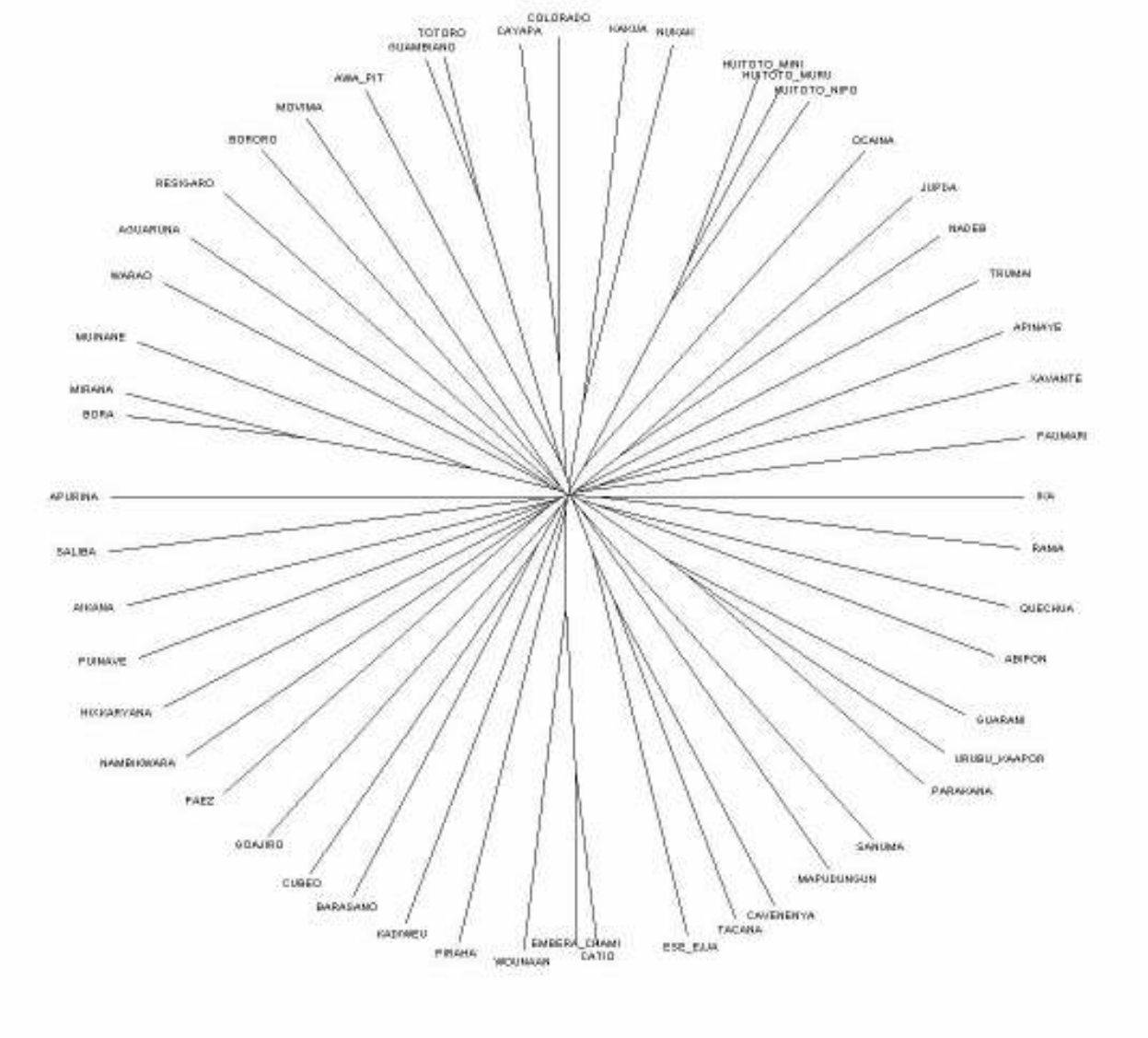


Figure 2 is an ASJP tree for all 52 South American languages of the sample. With a couple of exceptions, the tree of Figure 2 groups together languages of the South American sample belonging to the same language genetic groups. These genetic

groups, all of which are *non-controversial*, and languages of the sample affiliated with them are:

COCONUCO: Totoro, Guambiano.
BARBACOAN: Awa Pit, Cayapa, Colorado.
KAKUA-NUKAK: Kakua, Nukak.
HUITOTO-OCAINA: Huitoto Minica, Huitoto Nipode, Huitoto Murai, Ocaina.
NADAHUP: Jupda, Nadeb.
MACRO-GE: Apinaye, Xavante.
CHIBCHAN: Ika, Rama.
TUPI-GUARANÍ: Guaraní, Urubu-Kaapor, Parakana.
TACANAN: Cavineña, Tacana, Ese-Ejja.
CHOCOAN: Catio, Embera-Chami, Wounaan.
TUCANOAN: Barasano, Cubeo.
BORAN (BORA-MUINANE): Bora, Mirana, Muinane.

There are controversial or provisional genetic groups for South American languages not reflected by the branching of the ASJP tree (Figure 2). Aschmann (1993) groups BORAN (Bora, Mirana, Muinane) and HUITOTO-OCAINA (Huitoto Minica, Huitoto Nipode, Huitoto Murai, Ocaina) languages together in a HUITOTOAN (WITOTOAN) family. In the ASJP tree, BORAN languages are grouped together on a single branch and HUITOTO-OCAINA languages are grouped together on a single branch, but these two branches are not directly connected to one another. The tree, then, is not in agreement with Aschmann's HUITOTOAN proposal. This indicates that languages of BORAN and HUITOTO-OCAINA are not particularly lexically similar to one another. Indeed, Kaufman (1990:43), who regards BORAN and HUITOTO-OCAINA as distinct families, only tentatively recognizes their genetic connection. Aschmann (1993:124) himself comments on the great period of time that must pertain to the separation of these two groups. Referring to glottochronological results, he writes, "Proto-Witotoan would have a time depth of over 7000 years." ASJP does not appear able to register relationships of such great chronological depth.

Kaufman (1994:54) proposes the genetic union of Paez and the COCONUCO languages (Totoro, Guambiano) in a PAEZAN stock. Campbell (1997:173) notes that "[t]here is no consensus upon Paezan, and opinions vary greatly." While the COCONUCO languages are branched together on the ASJP tree (Figure 1), they are not directly connected with Paez. Consequently, the tree lends no support to the proposal.

Kaufman (1994:60) also proposes a PUINAVEAN stock which encompasses Puinave, KAKUA-NUKAK (Kakua, Nukak), and NADAHUP (Jupda, Nadeb). Patience Epps (personal communication) finds no good evidence at this point for grouping together KAKUA-NUKAK and NADAHUP languages and that the case for including Puinave with them is even weaker. While KAKUA-NUKAK and NADAHUP are genetic groups clearly identified by the ASJP tree, their respective branches are not directly linked to one

another nor is either directly connected to Puinave. Consequently, the tree lends no support to the PUINAVEAN proposal.

The ASPJ tree fails to group together the three MAIPURAN languages, Apurina, Goajiro, and Aikana, on a single branch, and Bororo is not grouped with the other two MACRO-GE languages, Apinaye and Xavante. Like the HUITOTOAN situation described above, this arboreal shortcoming perhaps reflects the great chronological depths of both MAIPURAN and MACRO-GE.

4. Comparison of ASJP Trees and Expert Classification.

Figures 1 and 2 demonstrate that ASJP is capable of recognizing known genetic groups of languages from among large numbers of related and unrelated languages. It is also capable of near-expert classification of the languages within genetic groups (subgrouping). In 4.1 we compare ASJP trees for languages of nine different genetic groups with expert classifications for languages within those groups.

4.1 Language genetic groups.

4.1.1 Mayan.

Languages of the MAYAN genetic group are joined together on a single branch of the ASJP tree for Middle American languages (Figure 1). Table 2 is an expert classification for MAYAN languages described by Brown and Wichmann (2004:129-130). The sub-branching of MAYAN languages on the tree is substantially the same as that of the expert classification of Table 2.

Table 2. Expert classification for Mayan languages (Brown and Wichmann 2004:129-130), limited to only those Mayan languages found in Figure 1.

GREATER KANJOBALAN: Chuj, Jacaltec
HUASTECAN: Huastec
EASTERN MAYAN:
    QUICHEAN: Kekchi, Pocomam, Quiche
    MAMEAN: Aguacatec, Mam
GREATER TZELTALAN:
    CHOLAN: Chol, Chorti
    TZELTALAN: Tzeltal, Tzotzil
YUCATECAN: Itzaj, Mopan

(Note: Huastecan, Eastern Mayan, Greater Tzeltalan, and Yucatecan are coordinate branches. The affiliation of Greater Kanjobalan, if any, is unknown [Brown and Wichmann 2004:130].)

4.1.2 Mixe-Zoque.

Languages of the MIXE-ZOQUE genetic group are joined together on a single branch of the ASJP tree for Middle American languages (Figure 1). Table 3 is an expert classification for MIXE-ZOQUE languages provided by Wichmann (1995:9). The sub-branching of MIXE-ZOQUE languages on the tree is substantially the same as that of the expert classification of Table 2.

Table 3. Expert classification for Mixe-Zoque languages (Wichmann 1995:9), limited to only those Mixe-Zoque languages found in Figure 1.

MIXEAN:
- **OAXACA MIXEAN:** Lowland Mixe, South Highland Mixe, North Highland Mixe
- Oluta Popoluca
- Sayula Popoluca

ZOQUEAN:
- **CHIMALAPA ZOQUE:** San Miguel Chimalapa Zoque, Santa Maria Chimalapa Zoque
- **GULF ZOQUEAN:** Texistepec Zoque, Soteapan Zoque
- **CHIAPAS ZOQUE:** Chiapas Zoque

4.1.3 Otomanguean.

Languages of the OTOMANGUEAN genetic group are joined together on a single branch of the ASJP tree for Middle American languages (Figure 1). The sub-branching of OTOMANGUAN languages on the tree is the same as that of the expert classification of Table 4 extracted from Kaufman (1994).

Table 4. Expert classification for Otomanguean languages (Kaufman 1994), limited to only those Otomanguean languages found in Figure 1.

WESTERN OTOMANGUEAN: Chinantec
EASTERN OTOMANGUEAN:
- **MIXTECAN:** Mixtec, Cuicatec

4.1.4 Huitoto-Ocaina.

Languages of the HUITOTO-OCAINA genetic group are joined together on a single branch of the ASJP tree for South American languages (Figure 2). The sub-branching of HUITOTO-OCAINA languages on the tree is the same as that of the expert classification of Table 5 extracted from Aschmann (1993).

Table 5. Expert classification for Huitoto-Ocaina languages (Aschmann 1993), limited to only those Huitoto-Ocaina languages found in Figure 2.

OCAINA: Ocaina
HUITOTO PROPER:
MINICA-MURUI: Huitoto Minica, Huitoto Murui
NIPODE: Huitoto Nipode

4.1.5 Tacanan.

Languages of the TACANAN genetic group are joined together on a single branch of the ASJP tree for South American languages (Figure 2). The sub-branching of TACANAN languages on the tree is the same as that of the expert classification of Table 6 extracted from Gordon (2005).

Table 6. Expert classification of Tacanan languages (Gordon 2005), limited to only those Tacanan languages found in Figure 2.

ARAONA-TACANA:
CAVINEÑA-TACANA: Cavineña, Tacana
TIATINAGUA: Ese Ejja

4.1.6 Chocoan.

Languages of the CHOCOAN genetic group are joined together on a single branch of the ASJP tree for South American languages (Figure 2). The sub-branching of CHOCOAN languages on the tree is the same as that of the expert classification of Table 7 extracted from Gordon (2005).

Table 7. Expert classification of Chocoan languages (Gordon 2005), limited to only those Chocoan languages found in Figure 2.

EMBERA:
NORTHERN: Catio
SOUTHERN: Embera-Chami
Wounaan

4.1.7 Muskogean.

Languages of the Muskogean grouping are spoken historically in the U.S. Southeast. Figure 3 is the ASJP tree for Muskogean languages based on a matrix of SSPs pertaining only to (all possible) pairs of Muskogean languages of our sample.

Table 8 is an expert classification for Muskogean languages provided by Haas (1949, 1979).

Figure 3. ASJP tree for Muskogean languages.

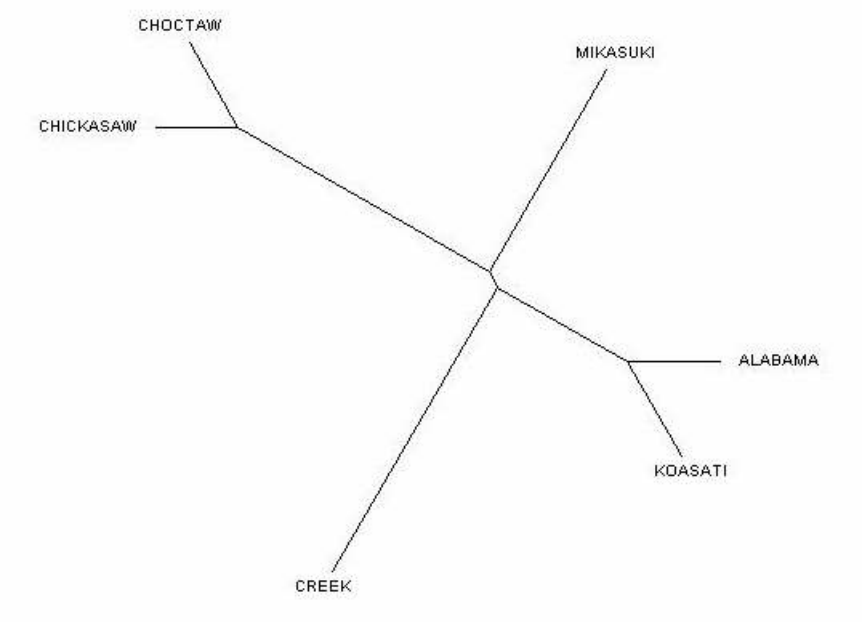


Table 8. Expert classification for Muskogean languages (Haas 1949, 1979), limited to only those languages found in Figure 3.

WESTERN MUSKOGEAN: Choctaw, Chickasaw
EASTERN MUSKOGEAN:
CENTRAL MUSKOGEAN:
ALABAMA-KOASATI: Koasati, Alabama
HITCHITI-MIKASUKI: Mikasuki
CREEK-SEMINOLE: Creek

The ASJP tree follows expert classification in grouping Choctaw with Chickasaw and Koasati with Alabama. The tree departs from expert classification by not associating Mikasuki more closely with the ALABAMA-KOASATI languages.

4.1.8 Indo-European.

Figure 4 is the ASJP tree for Indo-European languages based on a matrix of SSPs pertaining only to (all possible) pairs of Indo-European languages of our sample. Table 9 is the well-known, standard classification for Indo-European languages.

Figure 4. ASJP tree for Indo-European languages.

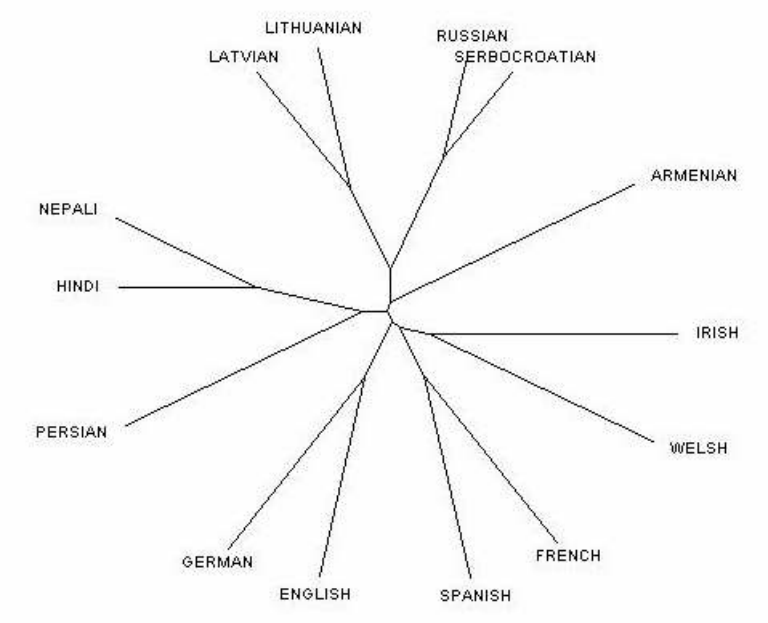


Table 9. Standard classification for Indo-European languages, limited to only those languages found in Figure 4.

INDO-IRANIAN:
IRANIAN: Persian
INDO-ARYAN: Hindi, Nepali
CELTIC: Irish, Welsh
ITALIC: French, Spanish
GERMANIC: English, German
ARMENIAN: Armenian
BALTIC: Latvian, Lithuanian
SLAVONIC: Russian, Serbo-Croatian

The language subgroups of the ASJP tree (Figure 4) and of the standard classification for Indo-European languages (Table 9) are substantially the same.

Probably the most controversial aspect of general Indo-European language classification is whether or not the BALTIC and SLAVONIC languages together constitute a BALTO-SLAVONIC division of the language family. There is little or no debate that BALTIC and SLAVONIC languages share certain linguistic features not found in other Indo-European languages. However, there is discussion over whether or not these common features are due to intensive contact or to genetic affiliation. Whatever the explanation, the ASJP tree for Indo-European (Figure 5) clearly groups BALTIC and SLAVONIC languages together on a branch separate from other languages.

The ASJP tree also shows a split between GERMANIC, ITALIC, and CELTIC languages on the one hand, and Armenian, BALTIC, SLAVONIC, and INDO-IRANIAN

languages on the other. These groups are reminiscent of the Centum/Satem distinction long recognized for Indo-European languages. Centum/Satem is based on the observation that Eastern languages (Satem) show certain phonological features that Western languages (Centum) lack. Since the presence of such features in Satem languages can be explained by diffusion, Centum/Satem is widely regarded as the product of early contact situations. The ASJP tree appears to capture this diffusional result.

4.1.9 Austro-Asiatic.

Figure 5 is the ASJP tree for Austro-Asiatic languages based on a matrix of SSPs pertaining only to (all possible) pairs of Austro-Asiatic languages of our sample. Table 10 is an expert classification of Austro-Asiatic languages provided by Bradley (1994:159).

Figure 5. ASJP tree for Austro-Asiatic languages.

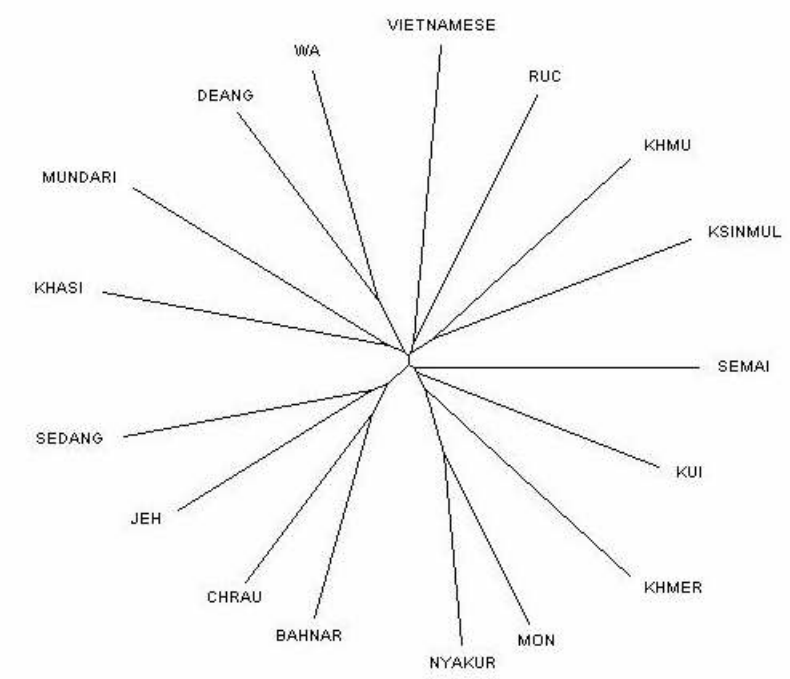


Table 10. Expert classification for Austro-Asiatic languages (Bradley 1994), limited to only those languages found in Figure 5.

MUNDA: Mundari
MON-KHMER:
- **KHASIAN:** Khasi
- **NORTHERN MON-KHMER:**
  - **PALAUNGIC:** Deang, Wa
  - **KHMUIC:** Khmu, Ksinmul
- **VIET-MUONG:** Vietnamese, Ruc

KATUIC-BAHNARIC:
BAHNARIC: Bahnar, Jeh, Sedang, Chrau
KATUIC: Kui
KHMER: Khmer
MONIC: Mon, Nyakur
ASILIAN: Semai

The ASJP tree recognizes these terminal subgroups of the expert classification: PALAUNGIC (Deang, Wa), KHMUIC (Khmu, Ksinmul), VIET-MUONG (Vietnmese, Ruc), BAHNARIC (Bahnar, Chrau, Jeh, Sedang), and MONIC (Mon, Nyakur). The tree fails to group together languages pertaining to KATUIC-BAHNARIC: languages of BAHNARIC and KATUIC respectively do not achieve branch association. However, Bradley (1994:160) notes that while most scholars closely affiliate languages of these two groups, some prefer to keep them separate. The ASJP tree, of course, does not lend support to this union. The tree also fails to group together languages pertaining to NORTHERN MON-KHMER: languages of PALAUNGIC and KHUMUIC respectively do not achieve branch association. Finally, the affiliation of Khasi with Mundari of the MUNDA grouping of Austro-Asiatic departs from expert classification. This probably reflects contact between speakers of MUNDA languages and Khasi, all spoken in eastern India. Mundari and Khasi are both spoken in the northern area of eastern India.

4.1.10 Austronesian.

We acknowledge our most problematic ASJP classification, that for the 43 Austronesian languages of our sample. This classification departs from expert determination with regard to some higher- and mid-level divisions of the family. These discrepancies probably reflect substantial diffusion of lexical items across these languages, resulting for the most part from the long tradition of long-distance seafaring of their speakers. Dyen's (1965) lexicostatistical classification of Austronesian languages apparently is similarly troubled by results reflecting considerable lexical diffusion since he is compelled to organize his classification by geographic region rather than solely by cognate percentages. Despite this strategy, a significant number of languages are left "ungrouped" in these regions, languages whose genetic affiliations with other Austronesian languages are now relatively well understood.

4.2 Evaluation of comparisons.

For the vast majority of cases, ASJP achieves near-expert classification in subgrouping languages of known genetic affiliation. For the most part, when ASJP is not in agreement with expert classification, the ASJP subgrouping reflects lexical diffusion, a phenomenon that expert classification attempts to factor out of consideration. Thus, the essential difference between ASJP and expert classification is that ASJP is sensitive to

both genetic relationship and diffusion, while expert classification, ideally at least, is sensitive only to genetic affiliation.

5. Uncovering heretofore unrecognized language relationships.

ASJP provides the prospect of discovering many language relationships not previously observed. Since it can rapidly produce comparisons of an unlimited number of language pairs, comparisons for many languages never previously compared can be easily achieved. Indeed, it is well within the range of possibility that all of the world's recorded languages can be relatively effortlessly compared with one another. An electronic search of such a huge comparative corpus has the potential to identify many related pairs of languages hitherto not known to be so.

For some world areas, for example, Europe, North America, and Middle America, most languages have been reasonably thoroughly compared for relatedness by historical linguists. Languages of other world areas, for various reasons, are still in need of substantial comparative attention. South America is one of these areas (Campbell 1997:170-171). Many South American languages remain only poorly known and comparative analysis of those that are reasonably well-recorded is not well-advanced. For example, a number of recorded languages of the region appear to be language isolates, languages not known to be genetically related to any other languages. Many of these have not been systematically compared with other languages of South America to which they may turn out to be related in some manner, either genetically or through contact.

The preliminary results of ASJP demonstrate its potential usefulness in expanding comparative analysis of the lexicon to languages of those world areas where the study of language relationship is not particularly well-advanced. Figure 2 is an ASJP tree for the 52 South American languages of the preliminary sample of 245 languages. As noted above, with only a couple of exceptions, known genetically related languages are correctly branched together on the tree (3.5). In addition, other languages not known to be related or only provisionally proposed to be related are also branched together. These are as follows:

1. **COCONUCO** (Totoro, Guambiano) with **BARBACOAN** (Awa Pit, Cayapa, Colorado).
2. **QUECHUAN** (Quechua) with **GUAYKURUAN** (Abipon).
3. **YANOMAM** (Sanuma) with **ARAUCANIAN** (Mapudungun).
4. **MURA** (Piraha) with **CHOCOAN** (Wounaan, Embera-Chami, Catio).
5. **TUCANOAN** (Cubeo, Barasano) with **MAIPURAN** (Goajiro).
6. **ISOLATE** (Paez) with **NAMBIKWARAN** (Nambikwara).
7. **ISOLATE** (Saliba) with **MAIPURAN** (Apurina).

To our knowledge, only one of these seven groupings, i.e., 1, has been previously suggested. Kaufman (1994:54) proposes a PAEZAN STOCK in which COCONUCO and BARBACOAN languages are included along with Paez and Andaki.

We do want to claim that all languages of the seven above groups are genetically related within their respective groups. In some instances, the close lexical similarity motivating such grouping may be due to contact and resulting diffusion. For example, this is probably true of the TUCANOAN/MAIPURAN and QUECHUAN/GUAYKURUAN associations. In addition, coincidental similarity cannot be ruled out.

Our claim is this. ASJP provides information useful for focusing scholars' attention on potential relationships among languages not previously recognized as affiliated or only provisionally so proposed. ASJP constitutes a powerful tool for creating a vast amount of comparative information that can be easily electronically managed and narrowed down to only those sets of comparisons with real potential for revealing language relatedness. These possible relationships can then be explored in detail and evaluated by experts.

There are only 52 South American languages in the current sample of 245. ASJP yields 1,326 pairwise comparisons for these 52 languages. If the number of South American languages were, for example, tripled to 156, the number of pairwise comparisons produced by ASJP for languages of the area would expand to 12,090, thus vastly enhancing the possibility of finding new relationships among the little-studied languages of this world area.

ASJP may also contribute to discovering relationships among languages of world areas whose languages are relatively well-studied. A case in point are the languages of Middle America. Figure 1 is the ASJP tree for the 34 Middle American languages of the sample. As noted above, with no exceptions, known genetically related languages are correctly branched together on the tree (3.5). In addition, other languages not known to be related or only provisionally proposed to be related are also branched together. These are as follows:

1. **TOTONACAN** (Totonac, Tepehua) with **MIXE-ZOQUE** (see Table 3 for languages).
2. **TOTONACAN** and **MIXE-ZOQUE** with **HUAVEAN** (Huave).
3. **UTO-AZTECAN** (Nahuatl) with **TARASCAN** (Tarascan).
4. **UTO-AZTECAN** and **TARASCAN** with T**EQUISTLATECAN** (Highland Tequistlatec).

To our knowledge, no one has seriously proposed the unions reported by 3 and 4. Since these both involve Nahuatl, diffusion in large part probably explains the lexical similarity registered by ASJP since Nahuatl was a widespread lingua franca in Mesoamerica both before European contact and well into colonial times.

As Campbell (1997:323) notes, a number of scholars have considered the possibility that MAYAN, TOTONACAN, and MIXE-ZOQUE languages constitute a MACRO-MAYAN genetic group which may also inclusively extend to HUAVEAN. Campbell (1997:324) believes that eventually the genetic association of MAYAN with MIXE-ZOQUE, and, perhaps, also these groups with TOTONACAN will be proved. The ASJP tree for Middle American languages (Figure 1) lends evidence for the union of TOTONACAN with MIXE-ZOQUE, and the association of these two groups with HUAVEAN (1 and 2), but not the affiliation of any of these three groups with MAYAN. To our knowledge, there are no published or otherwise widely circulated descriptions of a detailed comparison of TOTONACAN and MIXE-ZOQUE languages. We have undertaken a detailed inspection of the ASJP word lists yielding SSPs indicating a TOTONACAN/MIXE-ZOQUE relationship, and conclude that this association definitely deserves further investigation. On the other hand, a preliminary look at the possible linkage of HUAVEAN with TOTONACAN/MIXE-ZOQUE is not nearly so encouraging.

Such results indicate that even in its earliest manifestation, treating only a small percentage of the world's languages, ASJP shows considerable potential for uncovering language relationships heretofore unrecognized or only provisionally proposed. Indeed, if lexical data for all the world's recorded languages were entered into the ASJP database, comparative exploration would achieve a level of comprehensiveness heretofore not known in historical linguistics.

6. Conclusion.

The major conclusion of this report is that ASJP yields near-expert classification of the world's languages. Its classificatory results are not always in perfect agreement with expert classification primarily because the ASJP classification is based on lexical similarity which is influenced by *both* genetic affiliation and diffusion, while expert classification, at least ideally, is based on genetic affiliation alone. Nevertheless, the usually very close agreement between ASJP and expert classification indicates that the influence of diffusion on lexical similarity involving core vocabulary is minimal.

The continuing development and use of ASJP holds out inviting prospects for historical linguistics. For example, ASJP provides for the possibility of discovering language relationships heretofore not apparent. Should we be able to produce 3,125,000 SSPs for 2,500 of the world's language (a very real possibility) we would then be in possession of an enormous corpus of data amenable to electronic searches for yet-to-be-observed language relationships.

Relating to more specific applications, ASJP data, for example, can facilitate the determination of degrees of both lexical and phonological stability, important to selecting the most appropriate data for studying relationships among languages. In addition, an ASJP database enables lines of investigation not necessarily directly

focused on language change and relatedness. For example, ASJP allows measurement of how frequently different phonological segments are associated with different meanings on the 100-item list. Perhaps surprisingly, preliminary results show that some segments are found to be associated with some meanings more often than expected by chance, suggesting that sound symbolism is an influence even on core vocabulary items.

Finally, it was once observed by a long-departed, anonymous historical linguist that it would take "a thousand scholars working a thousand years" to classify all of the world's languages. ASJP now provides the possibility that only nine scholars working for just nine months could assemble the necessary lexical data to achieve this goal.[1] The millions of machine-made comparisons required for such a classification would probably take less than nine hours.

Postscript

We would like to encourage scholars to join us in our effort to automate comparison and classification of all of the world's languages. The project would especially benefit from individuals who are willing to produce and transcribe word lists. If interested in becoming an ASJP project member, please contact Cecil H. Brown at chbrown@niu.edu.

Acknowledgements

We would like to thank Patience Epps for providing comments relating to the classification of languages of South America. Others who responded with comments to a first draft of this paper or in other important ways deserve our gratitude as well. These include Barry Alpher, Gene Anderson, Robert Blust, Michael Cahill, Shobhana Chelliah, William Croft, Michael Dunn, Rob Goedemans, Martin Haspelmath, Nicholas A. Hopkins, Terrence Kaufman, Konstantin Krasukhin, David B. Kronenfeld, Frank Landsbergen, Stephen Levinsohn, Luisa Maffi, Robert Mailhammer, Carolyn Miller, Edith Moravcsik, Maarten Mous, Steve Marlett, Johanna Nichols, Andrew Pawley, Doris Payne, Ger Reesink, Keren Rice, Don Ringe, John Roberts, David S. Rood, Malcolm D. Ross, Fedor Rozhanskiy, Keith Snider, Jae Jung Song, John Stark, Thomas Stolz, Lynn Thomas, J. Marshall Unger, Pieter van Reenen, and Ljuba Veselinova.

[1]Assuming that a single transcriber could reasonably be expected to produce 1.5 100-word lists per day, 9 transcribers could produce 67.5 lists in a five-day work week, and 2,430 lists in 9 months (36 five-day work weeks). 2,500 is the conservative estimate for the number of languages in the world well-enough recorded that 100-word lists could be assembled for them (see 1.).

REFERENCES CITED

Aschmann, Richard P. (1993). *Proto Witotoan.* Arlington: SIL and the University of Texas at Arlington.

Bradley, David (1994). East and South-East Asia. In *Atlas of the World's Languages,* Christopher Moseley and R.E. Asher, eds., pp. 159-192. Routledge: London.

Brown, Cecil H., and Søren Wichmann (2004). Proto-Mayan syllable nuclei. *International Journal of American Linguistics* 70-128-186.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science* 309:2072-2075.

Dyen, Isidore (1965). *A Lexicostatistical Classification of the Austronesian Languages.* Indiana University Publications in Anthropology and Linguistics, Memoir 19 of the *International Journal of American Linguistics*. Baltimore: Waverly Press.

Goh, Gwang-Yoon. (2000). Probabilistic meaning of multiple matchings for language relationship. *Journal of Quantitative Linguistics* 7:53-64.

Gordon, Raymond G., Jr. (ed.) (2005). *Ethnologue: Languages of the World, Fifteenth Edition.* Dallas, Texas: SIL International. Online version: **http://www.ethnologue.com/**.

Guy, J. B. M. (1980). *Glottochronology without Cognate Recognition.* Pacific Linguistics, Series B, No. 79. Canberra: Australian National University.

Haas, Mary R. (1949). The position of Apalachee in the Muskogean familly. *International Journal of American Linguistics* 15:121-127.

Haas, Mary R. (1979). Southeastern languages. In *The Languages of Native America: Historical and Comparative Assessment,* Lyle Campbell and Marrianne Mithun, eds., pp. 299-326. Austin: University of Texas Press.

Huson, D.H. (1998) SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* 14(10):68–73.

Huson, D.H. and D. Bryant (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2):254-267.

Inkpen, Diana, Oana Franza, and Grzegorz Kondrak (2005). Automatic identification of cognates and false friends in French and English. *Proceedings of RANLP-2005*, Bulgaria, Sept. 2005, pp. 251-257.

Kaufman, Terrence (1990). Language history in South America: What we know and how to know more. In *Amazonian Linguistics: Studies in Lowland South American Languages*, Doris L. Payne, ed., pp. 13-73. Austin: University of Texas Press.

Kaufman, Terrence (1994). The native languages of South America. In *Atlas of the World's Languages,* Christopher Moseley and R.E. Asher, eds., pp. 46-76. Routledge: London.

Kaufman, Terrence (1994). The native languages of Meso-America. In *Atlas of the World's Languages,* Christopher Moseley and R.E. Asher, eds., pp. 34-41. Routledge: London.

Kessler, Brett (2001). *The Significance of Word Lists.* Stanford, California: CSLI Publications.

Kondrak, Grzegorz (2003a). Phonetic alignment and similarity. *Computers and the Humanities* 37:273-291.

Kondrak, Grzegorz (2003b). Identifying complex sound correspondences in bilingual wordlists. *Proceedings of the Fourth International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003)*, pp. 432-443, Mexico City, February 2003. (Lecture Notes in Computer Science 2588, Springer-Verlag.)

Kondrak, Grzegorz and Tarek Sherif (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. *Proceedings of the COLING-ACL Workshop on Linguistic Distances*, pp. 43-50, Sydney, Australia, July 2006.

MacKay, Wesley and Grzegorz Kondrak (2005). Computing word similarity and identifying cognates with pair hidden Markov models. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL 2005)*, pp. 40-47, Ann Arbor, Michigan, June 2005.

McMahon, April and Robert McMahon (2005). Language Classification by Numbers. Oxford: Oxford University Press.

Nakhleh, Luay, Don Ringe, and Tandy Warnow (2005). Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81:382-420.

Oswalt, Robert L. (1970). The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3:117-129.

Ringe, Donald A., Jr. 1992. *On Calculating the Factor of Chance in Language Comparison. Transactions of the American Philosophical Society*, 82, pt. 1. Philadelphia: American Philosophical Society.

Ringe, Donald, Tandy Warnow, and Anne Taylor (2002). Indo-European and computational cladistics. *Transactions of the Philological Society* 100:59-129.

Swadesh, Morris (1951). Diffusional cumulation and archaic residue as historical explanations. *Southwestern Journal of Anthropology* 7:1-21.

Swadesh, Morris (1971). What is glottochronology? In *The Origin and Diversification of Language,* Joel Sherzer, ed., pp. 271-284. Chicago: Aldine Atherton.

Villemin, F. (1983). Un essai de détection des origines du japonais â partir de deux méthodes statistiques. In *Historical Linguistics,* B. Brainerd, ed., pp. 116-135. Bochum: N. Brockmeyer.

Wichmann, Søren (1995). *The Relationship among the Mixe-Zoquean Languages of Mexico.* Salt Lake City: University of Utal Press.

Wichmann, Søren and Arpiar Saunders (2007). How to use typological databases in historical linguistic research. *Diachronica* 24(2): ??-??.

APPENDIX A:
Swadesh 100-item list (Swadesh 1971:283).

1. I	21. dog	41. nose	61. die	81. smoke
2. you	22. louse	42. mouth	62. kill	82. fire
3. we	23. tree	43. tooth	63. swim	83. ash
4. this	24. seed	44. tongue	64. fly	84. burn
5. that	25. leaf	45. claw	65. walk	85. path
6. who	26. root	46. foot	66. come	86. mountain
7. what	27. bark	47. knee	67. lie	87. red
8. not	28. skin	48. hand	68. sit	88. green
9. all	29. flesh	49. belly	69. stand	89. yellow
10. many	30. blood	50. neck	70. give	90. white
11. one	31. bone	51. breasts	71. say	91. black
12. two	32. grease	52. heart	72. sun	92. night
13. big	33. egg	53. liver	73. moon	93. hot
14. long	34. horn	54. drink	74. star	94. cold
15. small	35. tail	55. eat	75. water	95. full
16. woman	36. feather	56. bite	76. rain	96. new
17. man	37. hair	57. see	77. stone	97. good
18. person	38. head	58. hear	78. sand	98. round
19. fish	39. ear	59. know	79. earth	99. dry
20. bird	40. eye	60. sleep	80. cloud	100. name

APPENDIX B:
The 245 languages of the current ASJP sample.

Abipon, Aguacatec, Aguaruna, Aikana, Alabama, Alutor, Amharic, Apinaye, Apurina, Arabic, Arikara, Armenian, Auyana, Awa (N. Guinea), Awa Pit, Bahnar, Bali, Bambara, Banoni, Barasano, Basque, Beng, Blackfoot, Bora, Bororo, Buli, Burmese, Cantonese, Carolinian, Catio, Cavineña, Cayapa, Central Amis, Central Carrier, Central Yupik, Chamorro, Chechen, Cherokee, Chickasaw, Chinantec, Choctaw, Chol, Chorti, Chrau, Chuj, Chukchee, Colorado, Comanche, Cowlitz, Creek, Cubeo, Cuicatec, Deang, Diegueno, Eastern Cham, Eastern Pomo, Embera Chami, English, Ese-Ejja, Estonian, Evenki, Favorlang, Fijian, Finnish, French, Gadsup, Gapapaiwa, Gaurang Gaur, Georgian, German, Giman, Goajiro, Guambiano, Guaraní, Hanunoo, Hausa, Hawaiian, Hidatsa, Hindi, Hixkaryana, Hmong, Huastec, Huave, Huitoto Minica, Huitoto Murui, Huitoto Nipode, Hungarian, Ika, Imorod, Indonesian, Ingush, Iraqw, Irish, Itzaj, Jacaltec, Japanese, Jeh, Jicaque, Jupda, Kadiweu, Kairiru, Kakua, Kannada, Kanuri, Kapampangan, Kaulong, Kekchi, Kewa, Khasi, Khmer, Khmu, Kilivila, Klamath, Koasati, Kokota, Korean, Koryak, Ksinmul, Kui, Kusaie, Lahu, Lakhota, Latvian, Lenakel, Lithuanian, Lou, Lowland Mixe, Maidu, Malagasy, Mam, Mandarin, Mapudungan, Mayo, Mikasuki, Mirana, Miwok, Mixtec, Mocha, Mon, Mopan, Movima, Muinane, Mundari, Nadeb, Nahuatl, Nalik, Nambikwara, Nasioi, Navajo, Nepali, Nez Perce, Ngizim, Northern Itelmen, North Highland Mixe, Nukak, Nung, Nunggubuyu, Nyakur, Ocaina, Oluta, Oneida, Oromo, Paez, Parakana, Paumari, Persian, Piraha, Pocomam, Proto-Wintun, Puinave, Quechua, Quiche, Rama, Resigaro, Roviana, Ruc, Russian, Saliba, Samoan, Sanuma, Sayula, Sedang, Sediq, Semai, Serbian-Croatian, Siar, Sika, Sisiqa, Soboyo, Somali, Soteapan, Southern Itelmen, South Highland Mixe, Spanish, Spokane, Sudest, Sundanese, Swahili, Tacana, Tagalog, Taiof, Tairora, Takia, Tamil, Taraon, Tarascan, Telugu, Tepehua, Tequistlatec, Texistepec, Thai, Tigre, Timucua, Tiwi, Toaripi, Totonac, Totoro, Trumai, Tungak, Turkana, Turkish, Tzeltal, Tzotzil, Urubu Kaapor, Vietnamese, Vitu, Wa, Warao, Welsh, Wichita, Wik-Mungkan, Wiyot, Wounaan, Xavante, Yabem, Yagaria, Yamdena, Yapese, Yareba, Yavapai, Zoque (Chiapas), Zoque (San Miguel Chimalapa), Zoque (Santa Maria Chimalapa), Zulu, Zuni.

APPENDIX C:
ASJP orthography

Below are presented the symbols of the standard QWERTY keyboard for English, symbol modifiers, and conventions for symbol presentation that constitute the ASJP orthography.

VOWELS (symbols, modifiers, and conventions):

Symbols:

i = high front vowel, rounded and unrounded [IPA: i, I, y, ɣ]
e = mid front vowel, rounded and unrounded [IPA: e, ø]
E = low front vowel, rounded and unrounded [IPA: a, æ, ɛ, œ, ɶ]
3 = high and mid central vowel, rounded and unrounded [IPA: ɨ, ɘ, ə, ɜ, ʉ, ɵ, ɞ]
a = low central vowel, unrounded [IPA: ɐ]
u = high back vowel, rounded and unrounded [IPA: ɯ, u]
o = mid and low back vowel, rounded and unrounded [IPA: ɤ, ʌ, ɑ, o, ɔ, ɒ]

Modifier:

An asterisk (*) following any one of the above seven vowel symbols indicates vowel nasalization, for example, ta*k. ASJP judges nasalized vowels as being similar to their non-nasalized counterparts.

Conventions:

Long vowels, e.g., uu or u: or u˙, are transcribed as if they were short vowels, e.g., u.

Accents on vowels are not recorded.

Tone is not recorded.

CONSONANTS (symbols, modifiers, and conventions):

Symbols:

p = voiceless bilabial stop and fricative [IPA: p, ɸ]
b = voiced bilabial stop and fricative [IPA: b, β]

m = bilabial nasal [IPA: m]
f = voiceless labiodental fricative [IPA: f]
v = voiced labiodental fricative [IPA: v]
8 = voiceless and voiced dental fricative [IPA: θ, ð]
4 = dental nasal [IPA: n̪]
t = voiceless alveolar stop [IPA: t]
d = voiced alveolar stop [IPA: d]
s = voiceless alveolar fricative [IPA: s]
z = voiced alveolar fricative [IPA: z]
c = voiceless and voiced alveolar affricate [IPA: ts, dz]
n = voiceless and voiced alveolar nasal [IPA: n]
S = voiceless postalveolar fricative [IPA: ʃ]
Z = voiced postalveolar fricative [IPA: ʒ]
C = voiceless palato-alveolar affricate [IPA: tʃ]
j = voiced palato-alveolar affricate [IPA: dʒ]
T = voiceless and voiced palatal stop [IPA: c, ɟ]
5 = palatal nasal [IPA: ɲ]
k = voiceless velar stop [IPA: k]
g = voiced velar stop [IPA: g]
x = voiceless and voiced velar fricative [IPA: x, ɣ]
N = velar nasal [IPA: ŋ]
q = voiceless uvular stop [IPA: q]
G = voiced uvular stop [IPA: G]
X = voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative [IPA: χ, ʁ, ħ, ʕ]
7 = voiceless glottal stop [IPA: ʔ]
h = voiceless and voiced glottal fricative [IPA: h, ɦ]
l = voiced alveolar lateral approximate [IPA: l]
L = all other laterals [IPA: ʟ, ɭ, ʎ]
w = voiced bilabial-velar approximant [IPA: w]
y = palatal approximant [IPA: j]
r = voiced apico-alveolar trill and all varieties of "r-sounds" [IPA: r, R, etc.]
! = all varieties of "click-sounds" [IPA: !, ǀ, ǁ, ǂ]

Modifiers:

The symbol ~ is a modifier that follows two juxtaposed consonants. ASJP regards such consonants as being in the same single position in a syllable. For example, kw~at is an ASJP transcription of a syllable originally transcribed by k^{w}at. ASJP judges syllables such

as kat and wat as both being lexically similar to kw~at. ASJP also judges strings such as kaw and kwi as both being lexically similar to kw~at. Examples of the common use of the modifier involve consonants that are labialized (kw~, tw~), aspirated (kh~, th~), palatalized (ky~, ty~), and pre-nasalized (nd~, mb~).

The symbol $ has a function similar to that of ~ except that it follows three juxtaposed consonants instead of two. ASJP regards the three consonants as being in the same single position in a syllable. For example, ASJP judges nim, dam, and yom as all being similar to ndy$im. ASJP also judges syllables such as nad, niy, and dey as all being similar to ndy$im. This modifier is sparingly used.

The modifier ” immediately follows a consonant that is glottalized (i.e., an ejective), e.g., k”. Glottalized consonants, e.g., t”, k”, C”, are judged similar to their non-glottalized counterparts, e.g., respectively, t, k, C, such that t = t”, k = k”, etc.

Conventions:

Word-initial glottal stops are not recorded.

Certain complex syllable nuclei are reduced to simple syllable nuclei by deleting certain consonants in certain positions in nuclei: CVhC, CV7C, CVxC, CvXC, and CVyC are all reduced to CVC (where C = consonant and V = vowel).

GENERAL CONVENTIONS:

A gloss on the 100-item list denoted by a word containing no consonants is treated as if there were no word for the gloss in a pertinent language. (This is because ASJP is set up only to match compared words that both have at least one consonant.)

If an original transcription of a word shows a symbol for a sound not accounted for by the ASJP orthography (see above), then that sound will be transcribed by that ASJP symbol most closely resembling it in manner and place of articulation.

APPENDIX D:
Instructions for automated similarity judgment

CONDITIONS FOR WHEN TWO WORDS WITH IDENTICAL SYMBOLS ARE JUDGED SIMILAR:

At least two symbols found in a single syllable of one word for a specific referent must be identical respectively to at least two symbols found in a single component of a word for the same referent found in another language, in order for the two words to be judged lexically similar to one another.

Where C = consonant and V = vowel, if there is a C_1VC_2 component in both of two compared words, and there are identical symbols for consonants such that $C_1 = C_1$ and $C_2 = C_2$ in respective words, then the words are judged similar. Note that in this circumstance the symbols for consonants must be identical, but symbols for vowels do not have to be identical. Examples: **b**u**k** // **b**e**k** and **v**3**n** // **v**i**n**ik. Such judgments are order sensitive such that C_1VC_2 of one word would not be regarded similar to C_2VC_1.

If there is a C_1VC_2 or C_1C_2 component in one of two compared words and a C_1C_2 component in the other, and $C_1 = C_1$ and $C_2 = C_2$ in respective words, then the words are judged similar. Examples: yaN**k**a**p**ulan // **kp**sx and ape**8k**w~ // pLa**8k**.

If there are only two or three symbols in one of two compared words, and these symbols are either C_1V_1 or CC_1V_1 or V_1C_1 or V_1C_1C, and the other word has two or more symbols, and a C_1V_1 component or V_1C_1 component of one word agrees with a respective C_1V_1 or V_1C_1 component of the other word such that $C_1 = C_1$ and $V_1 = V_1$, then the two words are judged similar. Note that in this circumstance the symbols for consonants must be identical *and* the symbols for vowels must be identical as well. Examples: ph~**it** // **it** and an**ni**ow // **ni**.

If a word has the shape $\#V_1C_1V_2\#$ (# = word boundary), and a compared word has as a component V_1C_1 or C_1V_2, and where $C_1 = C_1$ and $V_1 = V_1$ or $V_2 = V_2$ in respective words, the words are judged similar. Note that in this circumstance the symbols for consonants must be identical *and* the symbols for vowels must be identical as well. Examples: **ap**i // rew**ap**ga and E**ku** // pa**ku**ni.

A component of the form C_1C_2~, for example, kw~ and ph~, in one of two compared words, is judged similar to either C_1 or C_2 occurring in the other word as long as $C_1 = C_1$ or $C_2 = C_2$ in respective words. Examples: p**h~**o**y**oq // **h**o**y**ol and e8**k**w**~i**wiwa // a**ki**.

When symbols for vowels must be taken into consideration in comparisons (and they are not taken in consideration in CVC/CVC comparisons, see above), if there are in a syllable one vocalic symbol immediately followed by another vocalic symbol, then either

of the two vocalic segments will be considered a match with another string if one of the two vocalic symbols is found in an appropriate place in another word. Such vowel run-ons and/or diphthongs only come into consideration when found in strings of symbols such as #CV_1V_2#, #V_1V_2C#, #V_1V_2CV#, and #VCV_1V_2#. For example, language 1 may have tau 'head' and language 2 may have tudik 'head'. These two words are to be judged similar: **ta****u** // **tu**dik. Another example is language 1 tau and language 2 tamas. These two words are to be judged similar: **ta**u // **ta**mas.

#C_1V_1# is to be regarded as similar to #$C_1V_1V_2C$# or #$C_1V_2V_1C_2$# when $C_1 = C_1$ and $V_1 = V_1$. For examples: **no** // **no**ak and **no** // **n**i**o**p.

C_1C_2~ and C_1VC_2 are judged similar when $C_1 = C_1$ and $C_2 = C_2$. For example: **ch~**ikh~ // a**c**a**h**ua.