Journal of Quantitative Linguistics
2006, Volume 13, Numbers 2–3, pp. 225–264
DOI: 10.1080/09296170600850759

Routledge
Taylor & Francis Group

A Critique of the Separation Base Method for Genealogical Subgrouping, with Data from Mixe-Zoquean*

Michael Cysouw, Søren Wichmann and David Kamholz
Max Planck Institute for Evolutionary Anthropology, Leipzig

ABSTRACT

Holm (2000) proposes the "separation base" method for determining subgroup relationships in a language family. The method is claimed to be superior to most approaches to lexicostatistics because the latter falls victim to the "proportionality trap", that is, the assumption that similarity is proportional to closeness of relationship. The principles underlying Holm's method are innovative and not obviously incorrect. However, his only demonstration of the method is with Indo-European. This makes it difficult to interpret the results, because higher-order Indo-European subgrouping remains controversial. In order to have some basis for verification, we have tested the method on Mixe-Zoquean, a well-studied family of Mesoamerica whose subgrouping has been established by two scholars working independently and using the traditional comparative method. The results of our application of Holm's method are significantly different from the currently accepted family tree of Mixe-Zoquean. We identify two basic sources of problems that arise when Holm's approach is applied to our data. The first is reliance on an etymological dictionary of the proto-language in question, which creates problems of circularity that cannot be overcome. The second is that the method is sensitive to the amount of documentation available for the daughter languages, which has a distorting effect on the computed relationships. We then compare the results of Holm's approach with lexicostatistics and show that the latter actually performs quite well, producing a family tree for Mixe-Zoquean very similar to the one arrived at through the comparative method.

*Address correspondence to: Søren Wichmann, MPI-EVA, Department of Linguistics, Deutscher Platz 6, D-04103 Leipzig. E-mail: soerenw@hum.ku.dk

0929-6174/06/1302–30225$16.00

1. INTRODUCTION[1]

Holm (2000, 2003, 2004, 2005) has recently developed a subgrouping method that is akin to lexicostatistics but whose results, he claims, are superior to it. Holm criticizes more traditional approaches on the grounds that they are vulnerable to the "symplesiomorphy trap" and the "proportionality trap". The symplesiomorphy trap, in fact a widely recognized limitation of the comparative method, refers to the difficulty of distinguishing common retentions (symplesiomorphies in the parlance of phylogenetics) from shared innovations (synapomorphies) (Holm, 2004, p. 9). Proportionality refers to the assumption that a higher percentage of cognates shared between two languages indicates a closer historical relationship between them. Holm correctly points out that languages change in different ways and at different rates due to various factors, both internal and external, which are essentially unpredictable. One must always consider the possibility that later changes have obscured the earlier relationships within a family. Lexicostatistics, argues Holm, fails to take this into account. It produces incorrect results because it conflates similarity with genealogical relatedness, an error he calls the proportionality trap (Holm, 2000, p. 74).

Holm proposes his "separation base method" as an alternative. In practical terms, the method makes use of an etymological dictionary of the proto-language in order to infer a family tree. That is, one must already have an extensive list of reconstructed proto-roots along with their reflexes in the daughter languages, such as Pokorny (1994; 1948–1959) for Indo-European or Wichmann (1995) for Mixe-Zoquean. This is significantly more than is required for lexicostatistics, but there are still quite a few families where an etymological dictionary is available but the subgrouping is unclear. If Holm's method were to prove reliable, it would be very useful in such cases. Because of the relatively complicated assumptions and procedures involved, we have tested the method empirically before making an evaluation. Holm (2000) applies the method to Indo-European, but as there is no consensus on the higher Indo-European subgrouping, there is no way to independently verify the results. Our test case is Mixe-Zoquean, a well-studied family of Mesoamerica. This family has a much lower time-depth compared to

[1]We gratefully acknowledge comments from Sheila Embleton and Hans Holm on an earlier version of this paper. The usual disclaimers apply.

Indo-European, which makes the subgrouping more reliable. The main proposals for the internal subgrouping of Mixe-Zoquean come from two scholars who have independently applied the traditional comparative method and come to very similar conclusions. This makes the family a good candidate to test the performance of a new method.

In Section 2, we give a brief overview of the genealogical relationship, geographical distribution, and recent comparative study of the Mixe-Zoquean languages. In Section 3, we elaborate the "separation base" method proposed by Holm and apply it to Mixe-Zoquean. The results contain a number of errors. In Section 4, we investigate the cause of these errors and conclude that Holm's approach suffers from fundamental weaknesses that would be difficult to overcome. In Section 5, we show that lexicostatistics, imperfect as it might be, clearly outperforms Holm's method. Finally, we show that parsimony analysis yielding a split decomposition tree produces even better results. We conclude that although Holm's criticism of lexicostatistics is in principle correct, its failure in practice (due to the "proportionality trap") is not as likely as he would lead us to believe, and Holm's approach does not overcome the weaknesses.

2. THE MIXE-ZOQUEAN LANGUAGES

Mixe-Zoquean is a family of languages spoken in the general area of the Isthmus of Tehuantepec in Mexico, in the states of Tabasco, Veracruz, Chiapas, and Oaxaca. Figure 1 shows, for each language, the approximate centre of the area in which it is presently spoken. The family also contains the extinct language Tapachultec Mixe, which was spoken in southeast Chiapas near the Pacific coast and the border with Guatemala. There is, however, too little data available for Tapachultec to be of any use for the purposes of this paper.

The history of the classification of Mixe-Zoquean languages up to the early 1990s is discussed in detail in Wichmann (1994, pp. 220–227). Here we shall give only a few highlights and a summary of more recent contributions. In the latter half of the 19th century it was recognized that Chiapas Zoque and Midland Mixe were related. During the following half-century increasingly more languages were recognized as belonging to the family by various scholars. Foster (1943) was the first to recognize that Soteapan Zoque, Texistepec Zoque, and Chiapas Zoque belong to one branch of

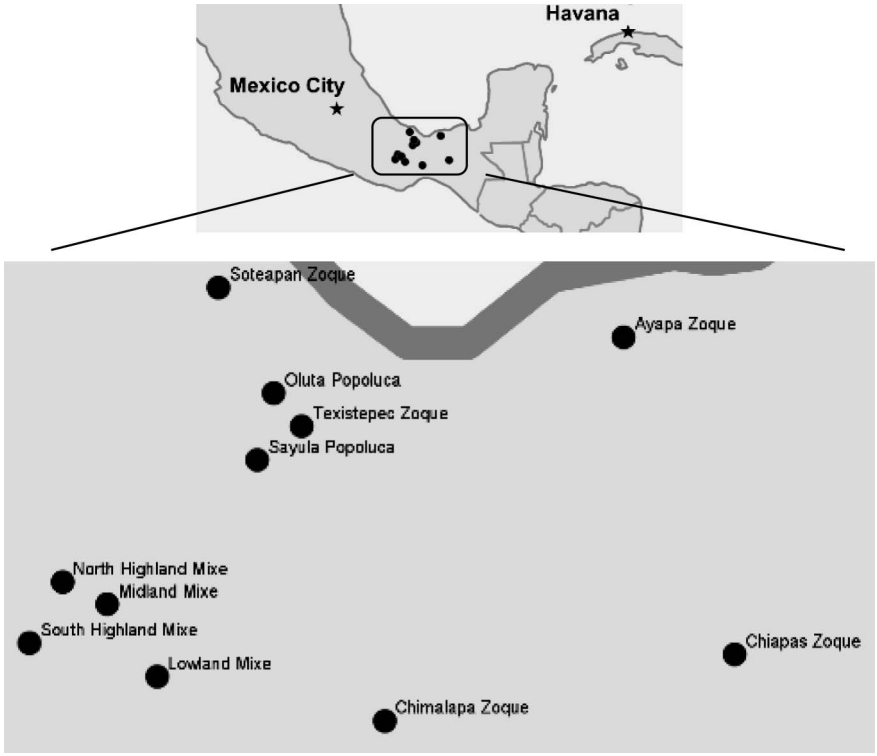


Fig. 1. Geographical distribution of the Mixe-Zoquean languages.

the family, while Oluta Popoluca, Sayula Popoluca, and Oaxaca Mixean belong to another. All subsequent research has affirmed the existence of a main division into a Mixean and a Zoquean branch, including the important contribution of Nordell (1962). Kaufman (1964a) showed that the extinct language Tapachultec belongs to the Mixean branch. Wichmann (1995) was the first scholar to include all Mixe-Zoquean languages in a classification. This classification, which is sustained by detailed phonological evidence, is shown in Figure 2 (for surveys of the evidence see Wichmann, 1994, pp. 227–230; 1995, pp. 8–12). The abbreviations indicated there are used throughout this paper.

Kaufman (1964b; 1974) presents classifications that differ somewhat from each other as well as from that of Wichmann (1995) regarding the internal configurations of the two main branches of the family. Since no arguments are given for these classifications and since the author has

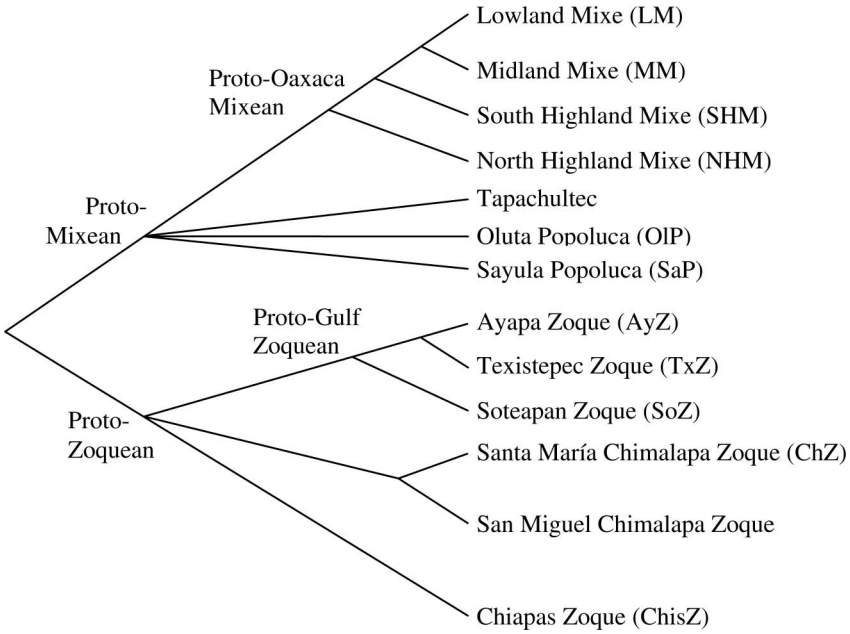


Fig. 2. The classification of Mixe-Zoquean languages of Wichmann (1995).

subsequently revised them, we shall not dwell on them here (see Wichmann, 1994, pp. 222–223 for a summary). Instead, we cite only Kaufman's latest classification, which is published in Kaufman and Justeson (2004), shown in Figure 3. It should be stressed, however, that this classification is not accompanied by explicit arguments either. Although Kaufman in some cases prefers different language designations and spellings, we employ the same names as throughout the rest of the paper in order to avoid confusion.

A comparison of Figures 2 and 3 reveals only minor differences in the classifications of Wichmann and Kaufman. The latter excludes Midland Mixe and South Highland Mixe altogether. However, no major controversy is to be inferred from this difference. Kaufman considers Oaxaca Mixean, including the South Highland and Midland varieties, to be a chain of dialects, whereas Wichmann considers a four-way division to be warranted. (Kaufman may also have excluded these two languages because he did not personally collect data on them.) There is, however, a difference of real importance with regard to the internal classification of the Mixean languages. Kaufman considers Sayula Popoluca and Oaxaca

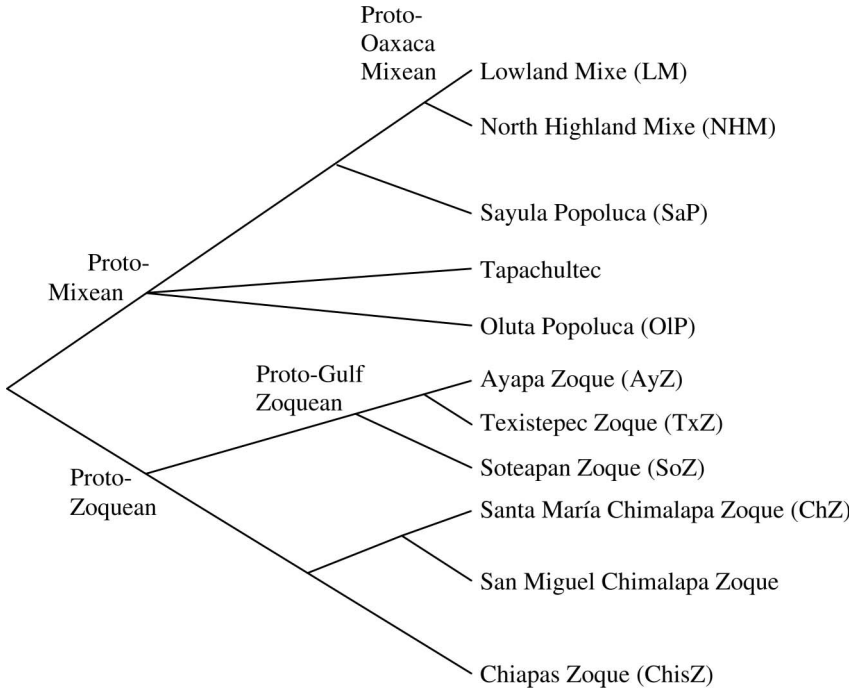


Fig. 3. The current classification of Kaufman (after Kaufman & Justeson, 2004, p. 1072).

Mixean to form a subgroup within Mixean, whereas Wichmann does not consider Sayula Popoluca and Oaxaca Mixean more closely related to each other than to Oluta Popoluca and Tapachultec. While the two authors differ on this point, it is noteworthy that in an earlier proposal Kaufman (1964b) considered Sayula Popoluca and Oluta Popoluca to form a subgroup, but now agrees with Wichmann that no such subgroup should be posited. With regard to the two internal classifications of the Zoquean branch, the only difference is that Kaufman considers Chiapas Zoque and Chimalapa Zoquean to form a subgroup, whereas Wichmann does not. Wichmann (1995, p. 12) does consider this as a "very likely" hypothesis, but prefers not to propose it in a definite way because of the lack of sufficient supporting data.

The Mixe-Zoquean family is one of the best investigated families of the Americas from a diachronic point of view. Two scholars, Kaufman and Wichmann, have independently carried out extensive studies, each drawing upon data that have also, to a certain extent, been gathered

independently. Nevertheless, there is a high level of agreement between their respective classifications. To get from Wichmann's to Kaufman's classification one needs only to set up two extra nodes at shallow levels of the classification. All nodes in Wichmann's tree are also in Kaufman's; the latter simply includes two additional ones embedded under the Mixean and Zoquean nodes, respectively.

The general point of this section is to stress that the application of the traditional comparative method by independent scholars may lead to very similar results. Additionally, we wish to make it evident that the classification used in this paper as the standard by which the results of other methods of establishing sub-grouping are to be compared and judged, viz. Wichmann's, is indeed well established and uncontroversial. We could have also used the one given in Kaufman and Justeson (2004), but prefer Wichmann (1995) since it is backed up by explicit arguments. Nevertheless, if a sub-grouping method yields results more similar to Kaufman's we shall not dismiss it for that reason. We do, however, consider a method suspect if it leads to positing nodes that are found neither in Kaufman's nor in Wichmann's family trees.

The possible external affiliations of the Mixe-Zoquean languages are not of much relevance to this paper, as relatively little is known. Wichmann (1994, pp. 238–242) summarizes 23 different proposals that have been made in the literature over the past century and a half. Many of these have been made in the absence of supporting evidence. Some families that are repeatedly mentioned as possibly related to Mixe-Zoquean are Mayan, Huavean, Penutian, Totonacan, and Uto-Aztecan. Wichmann (1999; 2003) believes that a Uto-Aztecan connection is a very real possibility and provides some lexical comparisons in support of this. Nevertheless, the relationship is so far removed as to be on the margin of what might be captured by the traditional comparative method, and this would no doubt also be true of any other external genetic link. In order to compare the results of different methods of classification against the comparative method it is necessary to stay within the limits of which the latter applies, which, in this case, is defined by the Proto-Mixe-Zoquean node.

3. HOLM'S "SEPARATION BASE" METHOD

It is a basic assumption of historical linguistics that all languages change over time. The *Stammbaum* model of linguistic change states

that every language is a direct descendant of an earlier language, from which it has preserved some features and lost or changed others. Thus, every language exhibits both innovations and retentions in comparison to some earlier stage. Both Holm's method and the traditional comparative method rest on these basic, uncontroversial assumptions.[2] But whereas the comparative method, when properly applied, relies on shared innovations to establish subgrouping within a family, Holm's method is based exclusively on shared retentions, specifically, shared lexical items.[3] He claims that it is possible to estimate subgroup relationships by modelling the loss of inherited vocabulary over time as a case of the hypergeometric distribution (this approach was first suggested by Kendall, 1950 in a reply to Ross, 1950, and a first attempt to apply this method can be found in Davies and Ross, 1975, pp. 40–48).

In practice, the method uses an etymological dictionary of the proto-language as its starting point. It requires a list of proto-roots, and for each root, a list of the daughter languages in which a reflex of the root is known to have survived.[4] Then, for every pair of languages in the family, a value N is calculated from the number of proto-roots that have survived in each language individually and the number that are shared between the two. This N value is an estimate of the number of proto-etyma that were present when the two languages diverged. As proto-etyma are lost over time, the smaller the estimate, the later

[2]It is now accepted that some languages, such as creoles, do not fit this model. But these are rather special cases, and we have no reason to presume anything other than ordinary genetic descent for Mixe-Zoquean.

[3]Holm (2003, p. 43) argues that "in real families" it is very difficult to isolate shared innovations from other phenomena such as borrowing. It is also never known if a language had an innovating feature in the past and later lost it. Shared retentions are thus more reliable data. We see no reason to believe that shared innovations should be inherently more difficult to recover than shared retentions. In fact, Holm's belief in the linguist's infallible ability to reconstruct the proto-lexicon is one of the major weaknesses of his approach (see Section 4).

[4]Thus, the basic data used in Holm's method are very different from the word lists typically used in lexicostatistics. In traditional ("Swadesh") word lists, the meanings are shared, but not necessary the form. In contrast, Holm's method uses cognates, which, though related in form, might have different meanings in different languages.

the divergence. After these estimates are calculated for all pairs of languages in the family, they can be used to reconstruct its history.[5]

3.1 The Hypergeometric Distribution

Holm's method is analogous to the "capture-recapture" method of estimating population size, which is based on the hypergeometric distribution. This distribution applies to a sample taken without replacement from a population consisting of two kinds of objects (Johnson et al., 1993, p. 237). (Without replacement means that members are not placed back into the population after they are drawn. Each drawing thus comes from a population one member smaller than the previous one.) The capture-recapture method involves taking two samples and using a simple formula to estimate the population size that produced the observed distribution.

To illustrate, suppose one wishes to estimate how many deer are in a forest. The first step is to capture a group of deer and mark them so that they can be identified later; this is sample one. The marked deer are then returned to the forest and time is allowed for them to mix back in evenly with the rest of the population, though not so much time that the total population size might change. A second sample is then taken, and the number of marked ("recaptured") deer are counted. Now, if the assumptions of the model hold (see below), one expects the ratio of marked deer to total deer in the second sample to be proportional to the ratio of marked deer to total deer in the population.

$$\frac{\textit{size of sample one (total deer marked)}}{\textit{num.deer in forest}} = \frac{\textit{recaptured deer (num.marked deer in sample two)}}{\textit{size of sample two}} \qquad (1)$$

[5]In this model, the proto-language is like an initial stock of lexemes of finite size which slowly diminishes over the family's history. It monotonically decreases, because innovations and borrowings are ignored. A problem with this assumption is the existence of recurrent cognation (Brainard, 1970, p. 70; Embleton, 1986, p. 63). A lexeme that is lost might later be reinserted into the language by borrowing, e.g., the borrowing of Latin words into French.

It is then a simple matter to solve for total number of deer:

$$num.deer\ in\ forest = \frac{size\ of\ sample\ one \times size\ of\ sample\ two}{num.marked\ deer\ in\ sample\ two} \quad (2)$$

The process of etymon loss leading from their common ancestor to languages x and y is analogous to the sampling of deer. In the deer example one chooses an arbitrary size for each sample, but in the linguistic case, the sampling is beyond our control. It occurs over time as each language retains certain of its proto-etyma, thus "marking" them. As we have already reconstructed the proto-lexicon, there is no need – nor indeed any possible way – to send the marked etyma from language x back "into the wild" to the proto-stage and see how many are retained in language y. Instead, we simply count the number of shared retentions between them. It is thus irrelevant whether a particular language is considered x as opposed to y; the results of the calculations will be the same. Holm's method uses the equation in (3) to estimate the number of proto-etyma that were present in the last common ancestor of two daughter languages:

$$\frac{num.etyma\ retained\ in\ language\ x}{num.etyma\ in\ last\ common\ ancestor\ of\ x\ and\ y} = \frac{num.etyma\ shared\ between\ x\ and\ y}{num.etyma\ retained\ in\ language\ y} \quad (3)$$

We may then solve for the number of proto-etyma N present at the last common ancestor of x and y, shown here in abbreviated form (read as *r*etentions over *a*greements):[6]

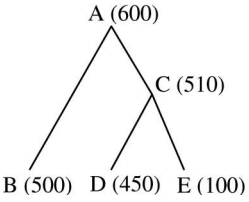
$$N_{xy} = \frac{r_x r_y}{a_{xy}} \quad (4)$$

For this model to be accurate, the sampling must occur in a random fashion (Johnson et al., 1993, p. 269). For languages this means that the likelihood of a given etymon's being lost should be the same for all languages in the family (for discussion of this assumption as it applies to real datasets see 4.3 below). The method essentially infers the unity of a

[6] N should be rounded down to the nearest integer. If N is already an integer, the next lowest integer is an equally accurate estimate. We have taken the average of the two integers in such cases.

subgroup from the unique "signature" left by the specific pattern of etymon loss for each period of common ancestry in a language family. For example, in the Proto-Mixe-Zoquean case, there must have been a period – how long we cannot know – where Proto-Mixean and Proto-Zoquean were distinct from each other, but maintained a certain amount of unity among themselves. During these periods of common development, there must – Holm hypothesizes – have been a characteristic set of proto-etyma retained and lost in each branch. Simply counting shared retentions in the daughter languages cannot recover the two branches, because the *number* of retentions indicates only how relatively innovative or conservative a language is. Rather, we must identify which languages share a particular "signature". This may be accomplished by calculating the *N* estimates for every pair of languages in a family and properly interpreting them. It is worth noting that, unlike glottochronology and unlike older methods of phylogenetics in biology, this method does not assume a constant rate of change. Indeed, there may be any amount of change in any of the branches, so long as a reasonable number of roots from the proto-language survive.

In theory, language relationships can successfully be reconstructed in this fashion. Consider the hypothetical tree in Figure 4 (based on Holm, 2003), with the number of remaining proto-etyma indicated at each node. In this example, the two daughter languages of proto-stage C are very different; language E is very innovative and language D is very conservative. By simply counting the number of retentions, language D would be grouped with the more distantly related language B, which is also conservative. Holm's approach may be evaluated by simulating the loss of etyma at each stage of development in the tree. The resulting number of shared retentions ("agreements") for each pair of languages are given in the table, along with the corresponding *N* estimates. These numbers



	Agreements	*N* estimates	Predetermined *N* values
B-D	378	595	600
B-E	83	602	600
D-E	87	517	510

Fig. 4. Hypothetical tree and results of the separation base method.

were produced by a simulation we wrote and show the results of a typical run. It can be clearly seen that the *N* values point to a closer relationship between D and E, as their value is significantly smaller than the other *N* values. Note also that the *N* values are good estimates of the (predetermined) number of retentions for the two proto-stages in the tree.

3.2 Application to Mixe-Zoquean

In order to test Holm's approach empirically, we have applied it to data taken from all sets of cognates in Wichmann (1995, pp. 235–522) listed as reflexes of Proto-Mixe-Zoquean roots. There is a total of 618 such reconstructed etyma in all.[7] In this list, we counted the total number of retentions *r* for each language and the number of agreements *a* for each pair of languages. We then calculated the *N* estimates for each pair of languages from these counts. The results are given in Table 1.[8]

The manner in which Holm (2000) uses these values to reconstruct the history of a family is only rather sketchily laid out in his article. We attempt to follow him here as closely as possible, although it has sometimes been necessary to guess precisely what he has in mind. The first thing Holm calculates from the values in Table 1 is the relative ordering of "earliest separation" and "latest separation" (Holm, 2000, pp. 82, 85). By this he appears to mean simply ordering the languages by their highest and lowest *N* values in Table 1, respectively. These lists are given in Table 2.[9] Following Holm's logic, ChisZ must have been the first

[7]The total of 699 Proto-Mixe-Zoquean entries mentioned by Wichmann (1995, p. 230) also includes reconstructed numerals and grammatical morphemes. We exclude these from consideration in this paper, however, as Holm's focus is on lexical items. It should not significantly affect the outcome in any case.

[8]We used a program written in `perl` to tabulate Wichmann's data and make the calculations. Three of the *N* estimates in Table 1 are larger than the original number of 618 reconstructed etyma included in this analysis, viz., those for ChisZ-OlP (626), ChisZ-SaP (643), and ChisZ-TxZ (644). The precise reasons for this are unclear, but likely result from the unexpectedly low number of agreements attested for these pairs, even clearly lower than what would be expected by chance.

[9]For the concept of latest separation, Holm refers to the "nearest neighbour method" as described by Embleton (1986, pp. 30–32; 1991, pp. 371–372). This method groups languages together recursively according to lowest dissimilarity. Note that this method need not simply produce a linear ordering, but is also capable of producing trees. Note further that Holm's "earliest separation" cannot be interpreted as the inverse of Embleton's method. It simply makes no sense in terms of reconstruction to group languages together that are highly dissimilar. The earliest separation should thus be considered a different approach, though likely inspired by Embleton's nearest neighbour method.

Table 1. Summary of Mixe-Zoquean etymological data based on Wichmann (1995). Number of agreements among shared retentions *a* in the lower left triangle. Estimates *N* of proto-etyma present in the last shared ancestor in the upper right triangle. Number of attested reflexes *r* on the last line.

	AyZ	ChisZ	ChZ	LM	MM	NHM	OlP	SaP	SHM	SoZ	TxZ
AyZ		577	406	545	511	554	467	444	428.5	357	431
ChisZ	107		575	606	605	595	626	643	589	591	644
ChZ	64	191		555	551	560	507	491	393	411	506
LM	97	368	169		472	540	585	579	491	559	594
MM	88	314	145	344		511	561	550	428	552	575
NHM	91	358	160	337	303		591	587	460	565	593.5
OlP	93	293	152	268	238	253		476	494	492	534
SaP	94	274	151	260	233	245	260		492	473	525
SHM	66	203	128	208	203	212	170	164		473	538
SoZ	85	217	131	196	169	185	183	183	124		447
TxZ	111	314	168	291	256	278	266	260	172	222	
r	121	511	215	437	372	417	359	345	234	251	396

Table 2. Languages ordered according to earliest and latest separation.

Earliest separation		Latest separation	
ChisZ	644 (from TxZ)	ChisZ	575 (from ChZ)
TxZ	644 (from ChisZ)	LM	472 (from MM)
SaP	643 (from ChisZ)	OlP	467 (from AyZ)
OlP	626 (from ChisZ)	NHM	460 (from SHM)
LM	606 (from ChisZ)	SaP	444 (from AyZ)
MM	605 (from ChisZ)	TxZ	431 (from AyZ)
NHM	595 (from ChisZ)	MM	428 (from SHM)
SoZ	591 (from ChisZ)	SHM	393 (from ChZ)
SHM	589 (from ChisZ)	ChZ	393 (from SHM)
AyZ	577 (from ChisZ)	SoZ	357 (from AyZ)
ChZ	575 (from ChisZ)	AyZ	357 (from SoZ)

language to split off from the Mixe-Zoquean ancestral language, because it has both the highest earliest separation and the highest latest separation. What happened next is unclear, because the remaining languages do not occur in the same order in the two columns. Yet there are two clusters of languages (separated by white space in the table), which can be taken to show that TxZ, SaP, OlP, LM, MM, and NHM were the next to split off, followed by SoZ, SHM, AyZ, and ChZ.

A further step proposed by Holm is the calculation of so-called *Bx* values for each pair of languages. The *Bx* values are calculated from the *N* estimates in Table 1. Essentially, they give an indication as to how similar two languages are to each other as opposed to the remaining languages. Specifically, to calculate *Bx* for a pair of languages *a* and *b*, Holm sums the differences between the N_{ay} and N_{by} for all languages *y* not equal to *a* or *b*, and takes the mean of these differences. The formula for *Bx* is given in equation (5), and the resulting values are shown in Table 3.

$$Bx = \frac{1}{n-2} \sum_{y \neq a,b} \left| N_{ay} - N_{by} \right| \tag{5}$$

Holm then considers the lowest *Bx* value for each language, as the lowest value should indicate which language is most similar to it. We also consider here the language with the next lowest value. According to the ordered *Bx* values (see Table 4), the Mixe-Zoquean languages are divided primarily into three groups, shown by the white space in the table. Interestingly, the group of four languages at the bottom of Table 2 (SoZ, SHM, AyZ, and ChZ) also form one group in Table 4. Furthermore, the languages NHM, LM, and MM form a group together with the somewhat more strongly differentiated ChisZ, which, according to Table 2, was the first language to split off.

Table 3. Holm's similarity measure *Bx* for all pairs of Mixe-Zoquean languages.

	ChisZ	ChZ	LM	MM	NHM	OlP	SaP	SHM	SoZ	TxZ
AyZ	147.8	34.4	101.3	92.1	94.0	68.1	59.9	61.6	22.2	74.1
ChisZ		121.8	58.3	81.7	54.9	79.8	87.9	140.5	125.7	73.7
ChZ			85.4	69.4	79.1	45.3	41.0	47.9	27.4	48.1
LM				23.3	12.8	57.7	65.8	82.2	90.7	69.1
MM					28.9	60.3	69.6	63.1	85.3	70.6
NHM						60.4	68.1	85.6	86.9	73.2
OlP							11.9	65.2	45.9	24.3
SaP								59.5	41.8	22.8
SHM									55.6	72.6
SoZ										51.9

Table 4. Lowest two *Bx* values for each language.

Language	Nearest to	*Bx* value	Next Nearest	*Bx* value
ChisZ	NHM	54.9	LM	58.3
MM	LM	23.3	NHM	28.9
NHM	LM	12.8	MM	28.9
LM	NHM	12.8	MM	23.3
OlP	SaP	11.9	TxZ	24.3
SaP	OlP	11.9	TxZ	24.3
TxZ	SaP	22.8	OlP	24.3
SoZ	AyZ	22.2	ChZ	27.4
AyZ	SoZ	22.2	ChZ	34.4
ChZ	SoZ	27.4	AyZ	34.4
SHM	ChZ	47.9	SoZ	55.6

3.3 Constructing a Historical Narrative

Holm suggests combining the information from Tables 2 and 4 and using knowledge of the geographical distribution of the languages to reconstruct the history of the family. From such information, a sequential historical narrative should then be constructed. (Holm believes that a tree cannot adequately express historical divergence among languages). For Mixe-Zoquean, this narrative might run as follows.

A division first starts to appear in Proto-Mixe-Zoquean between a south-eastern group (ChisZ, NHM, LM, MM) and a central group (OlP, SaP, TxZ). The languages within each group are highly similar to each other, according to the *Bx* values, and show the highest values of separation in Table 2. In the south-eastern group, ChisZ is the first to become completely separated from all other Mixe-Zoquean languages (it has the highest earliest and latest separation values of all languages), ending up as the easternmost language in the family. The remaining three languages from this group (NHM, LM, MM) are extremely similar, and are probably best considered a group of close relatives or even dialects in the south. The *N* values among these languages indicate a subgroup of LM and MM splitting off from NHM. The central group (OlP, SaP, TxZ) is also early to separate from the other languages. These three languages are extremely similar, according to their *Bx* values, and should be considered a separate subgroup. The *N* values indicate that OlP and SaP form a further subgroup among the three. The remaining languages (SoZ, ChZ, AyZ and SHM) are the last to branch off, remaining long in contact with the other languages before becoming completely distinct. Among these

four, SHM is the most outlying (according to the *Bx* values) and might be considered the earliest to split off (cf. Table 2), ending up quite close to the southern languages geographically (NHM, LM, MM). SoZ, ChZ, and AyZ are the most similar to each other according to the *Bx* values, but are found at different extremes of the Mixe-Zoquean area geographically. SoZ is located at the north-western extreme of the area in which the Mixe-Zoquean languages are spoken, AyZ to the northeast, and ChZ to the south. From the estimates in Table 1 it can be inferred that AyZ and SoZ shared the most recent historical ancestor, so ChZ split off first towards the south. AyZ and SoZ then split to the northeast and northwest, respectively.[10]

Unfortunately, we know that this narrative is wrong (cf. Fig. 2). Holm's method correctly groups three of the four Mixean languages together (NHM, LM, MM), and even gets their sub-grouping right, but makes a major error in placing SHM. The method correctly groups three of the Zoquean languages together (AyZ, SoZ, ChZ), also in the right order, but wrongly places ChisZ among the Mixean languages. OlP and SaP form an independent branch of Proto-Mixe-Zoquean rather than being grouped with the other Mixean languages, and TxZ is furthermore wrongly grouped together with them. A minor error is having SaP and OlP form a subgroup, although there are no shared innovations to justify such a grouping.

4. WHAT WENT WRONG?

There are several possible explanations for the incorrect results produced by Holm's method. We first investigate whether the results can be improved by interpreting the *N* values differently, concluding that they can to a certain extent, but that significant disagreements with the established history of Mixe-Zoquean still remain. We show that there are serious problems resulting from the basic principles of Holm's approach, viz., the usage of the hypergeometric distribution and of data from etymological dictionaries. Finally, we discuss some specific problems caused by the Mixe-Zoquean data we have used here. However, we argue

[10]After seeing an earlier version of this paper, Holm replied that this reconstruction was not how he would interpret the data; he claims to see a clear separation between Mixean and Zoquean. However, we still fail to see this anywhere in the *N* or *Bx* estimates.

that similar problems would arise in the application of Holm's method to other language families.

4.1 Other Ways to Interpret the *N* Values

It is possible that the *N* values in Table 1 accurately represent the history of Mixe-Zoquean, and it is simply the manner of interpretation that is flawed. We have therefore looked at several other ways to interpret the *N* values, all of which produce hierarchical tree structures. This allows us to judge whether the errors lie in the interpretation of the *N* values or in the *N* values themselves.

We first considered the *Bx* values, which are an aggregate of the *N* values. Applying the "nearest neighbour" method (Embleton, 1986, pp. 30–32; 1991, pp. 371–372) to the *Bx* data produces a tree that is almost identical to the narrative interpretation given above (see Fig. 5). The main disagreement with the Holm-style narrative is that the SaP-OlP-TxZ subgroup splits off after SHM instead of before. The ordering within the LM-MM-NHM subgroup is also slightly different. Generally, however, the agreements with the interpretation in Section 3.3 are striking.

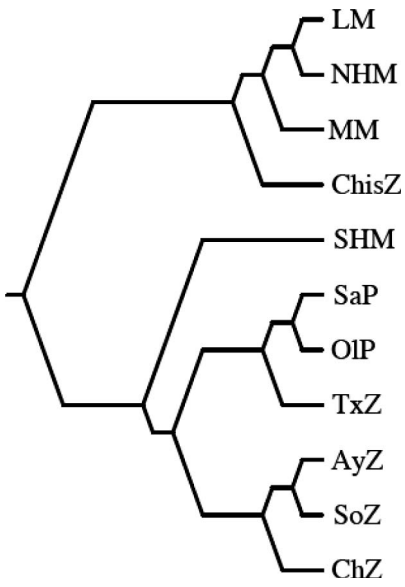


Fig. 5. Mixe-Zoquean tree based on Holm's *Bx* values, using Embleton's nearest neighbour method to infer the tree.

We next considered using the *N* values themselves as a measure of dissimilarity – the higher the *N*, the less closely related the two respective languages. We tested three separate tree-building methods from phylogenetics that take a dissimilarity matrix as input. We used algorithms that do not presuppose a constant rate of change, since assuming a constant rate of change would not be suitable for languages. These algorithms all produce unrooted trees; a dissimilarity matrix on its own provides insufficient evidence to establish a root.[11] Figure 6 shows the tree produced by the Fitch algorithm (Fitch & Margoliash, 1967).[12] We also tried the ADDTREE (Sattath & Tversky, 1977)[13] and NeighborJoining (Saitou & Nei, 1987)[14] algorithms, which produced almost exactly the same results. In the Fitch tree, the splits are all relatively close together, reflecting the fact that the *N* estimates are all relatively close together. Note that there is not much evidence for the various subgroups of Mixe-Zoquean, which can be inferred from the small branches between the various internal nodes and the long branches from the last shared ancestor to the individual languages.

Focusing only on the order of the splits, the tree produced by the Fitch algorithm shows a rather different structure than the narrative in Section 3.3 above or the tree in Figure 5. The Fitch tree is closer to the established history of the family, but still contains serious mistakes. The tree with the fewest errors results from fixing the root between SHM and ChisZ, which creates a Mixean group on one side (SHM, NHM, MM and LM, though SHM and NHM are grouped in the wrong order) and a Zoquean group on the other (ChisZ, ChZ, TxZ, SoZ and AyZ, though TxZ and SoZ are grouped in the wrong order). But SaP and OlP are unavoidably placed on the Zoquean side of the tree, and additionally are incorrectly grouped together.

To summarize, the different interpretations of the *N* estimates come to somewhat different conclusions. However, all contain serious mistakes as

[11]In order to establish a root, and hence the historical direction of these splits, we would need an outgroup language for comparison. That is, we would need a language that is historically related to Mixe-Zoquean but not a member of the family. For this purpose we might use Proto-Uto-Aztecan (cf. the end of Section 2). However, too few of the items in the list of 618 Proto-Mixe-Zoquean etyma have potential Uto-Aztecan cognates for us determine the root with any level of certainty.

[12]For this analysis, we used the `fitch` program from the `phylip` package.

[13]For this analysis, we used the `T-Rex` program.

[14]For this analysis, we used the `neighbor` program from the `phylip` package.

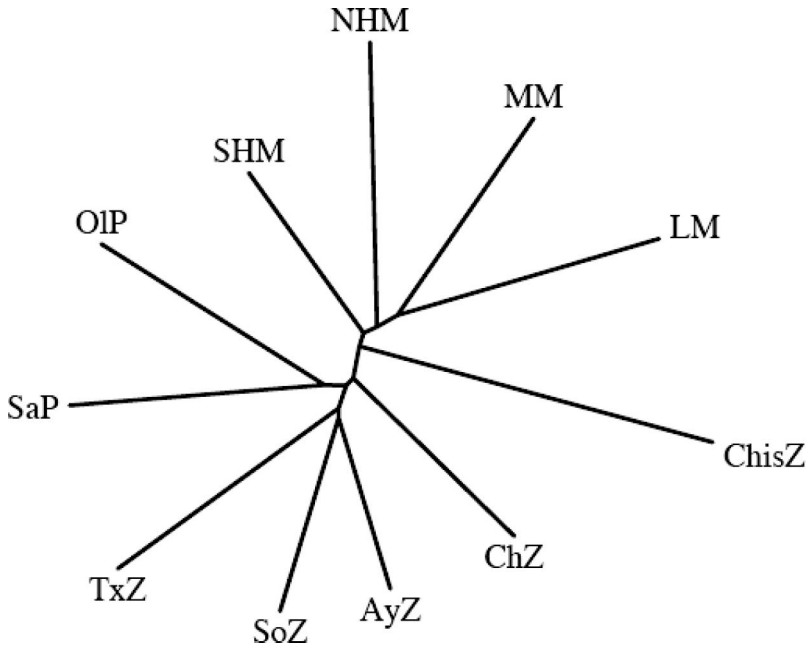


Fig. 6. Mixe-Zoquean tree, based on the *N* estimates of most recent shared ancestor, using the Fitch algorithm to infer the tree.

well as various minor errors. Some of the incorrect results of Holm's method in reconstructing the Mixe-Zoquean family tree may result from his manner of interpretation, but it is also clear that the *N* values themselves are not accurate. We turn now to the possible reasons for this.

4.2 The Use of Reconstructions as Data

Holm's method relies on shared retentions as data, and that is its first point of vulnerability. Two languages can be grouped together under one node when they show a similar enough pattern of shared retention, even if shared innovations – almost unanimously considered more substantial evidence – are lacking. In our case study, this error arises with the languages OlP and SaP. Wichmann (1995, p. 11) explains that "the Mixean languages of Veracruz, Sayula Popoluca [SaP] and Oluta Popoluca [OlP], are characterized by common retention rather than by shared innovation. It has not been possible to confirm that they form a true subgroup as is assumed in Kaufman's classification. The situation is probably that OlP and SaP have been different dialects ever since

Common Mixean times. Common retentions and probably some conversion, however, have made them so much alike that they share a host of features not found anywhere else.'' As mentioned in Section 2, Kaufman has now changed his view on SaP and OlP, so it can now be considered uncontroversial not to assign SaP and OlP to their own subgroup. But Holm's method conflates history of contact and mutual influence with close genetic relationship, and the assumption of a random sample (see Section 3.1) is thereby violated.

A less obvious but much more serious source of error lies in the process by which roots are established as proto-roots. The first step in reconstruction, once it is clear that a group of languages forms a family, is to establish lists of cognates between the languages. Only a portion of these can justifiably be considered reflexes of proto-roots, and it is a long process of sifting and weighting the evidence before such decisions can be made. The easiest cognates to discard are those that are clearly borrowings from another family, such as Spanish loanwords in Mixe-Zoquean. More problematic are internal loans, i.e., borrowings within a family, which can easily be mistaken for proto-roots (cf. Brainard, 1970, p. 71; see also Embleton, 1986, p. 141, where she attempts to correct for this in the analysis of Romance wordlists). However, filtering out borrowing is not everything. Even if we assume that this has been done effectively, a basis is still needed to decide which of the remaining potential cognate sets are common to the family under investigation (shared retentions) and which are innovations within a particular sub-branch (shared innovations). Only shared retentions should be reconstructed for the proto-language (and only they belong in the etymological dictionary). Usually, some notion of the family's sub-grouping is employed to help make this decision. In the case of Mixe-Zoquean, Wichmann uses the general criterion that a root must be attested in both Mixean and Zoquean in order for there to be sufficient evidence to consider it a proto-root.[15] In fact, he goes further, ruling out certain cases where borrowing is likely, but not provable: ''there are over sixty items which are recorded for one or more Zoquean languages, but only one of the Mixean languages: Sayula Popoluca [SaP]. . . . In [another] twenty-five cases evidence for a pMZ [Proto-Mixe-Zoquean] etymon is found only in Zoquean, SaP, and OlP. In these cases, the reconstructed form is labelled pMZ, whereas in the cases where

[15]That this poses a problem to Holm's approach was already recognized by Kendall (1950, p. 42f).

only SaP is involved the label pZ [proto-Zoquean] is used. ... In dealing with the items that are restricted to Zoquean and SaP I follow the principle that all these items ... are considered loans until this be disproved'' (Wichmann, 1995, pp. 213–214).

There is a serious problem of circularity here. Sub-grouping is precisely what Holm's method is supposed to uncover, and yet the data employed to do so are typically established on the basis of what is known about the family's sub-grouping. The problem is particularly acute in cases like Mixe-Zoquean where the proto-language is assumed to have undergone an initial binary split. In such cases one typically reconstructs a proto-root only if it is attested in both branches of the family; this is proper historical linguistic methodology. However, recall that Holm's method requires a unique ''signature'' of retentions in order to recover the unity of a branch. This means that in the case of Mixe-Zoquean it is actually impossible for Holm's method to determine that Mixean and Zoquean are distinct branches, because *every single reconstructed proto-root* survives into both by definition! In reality there must have been some proto-roots that only survived in one of the two branches, but there is no way to identify them. At first it might seem reasonable to include roots that have been reconstructed only for Proto-Mixean or Proto-Zoquean in the analysis, as this will catch many such ''covert'' Proto-Mixe-Zoquean roots. An analysis we performed with such data included indeed caused the division between Mixean and Zoquean to fall out clearly. However, this is really not very noteworthy, because including Proto-Mixean and Proto-Zoquean data essentially forces such an outcome. Agreements within each of the two sub-branches are artificially increased due to the guaranteed inclusion of many items that are not proto-roots, but shared innovations in one of the branches. And Holm's method cannot properly take shared innovations into account.

4.3 Influence of the Amount of Available Knowledge

A further problem to consider is that the number of retentions that are identified for a language in an etymological dictionary depends not only on the historical development of the language, but also on the amount of available data. The chance of finding a reflex of a given proto-etymon becomes greater as the amount of available data increases.[16]

[16]Embleton (1986, pp. 22–24) cites various examples in which selectively available knowledge influences the reconstruction.

In order to test the effect of data availability, we have estimated the number of entries in all dictionaries that were used in the preparation of Wichmann (1995) (see Table 5). These dictionaries have various entries for derivatives and compounds, but Wichmann's comparative dictionary only has a few, focusing mainly on roots. However, we judge that the proportion of derivatives and compounds is approximately the same in all of the individual language dictionaries, so the number of entries should be proportional to the amount of data used in constructing the comparative dictionary. When there are sources from different dialects for a given language, which is the case for SHM, MM, LM, and ChisZ, the counts are based on the largest source, but we have added 10% for each additional source, based on the assumption (from experience) that this is roughly the number of new forms one finds in a smaller new source that is relatively homogeneous, i.e., focuses on the same semantic fields and is compiled for similar purposes by either one and the same fieldworker or by fieldworkers with similar backgrounds.

As expected, there is a significant correlation between the amount of available data for a particular language and the number of attested retentions, as can be seen in Figure 7. This is not necessarily a problem, as long as the "missing" retentions (i.e., those proto-roots that have reflexes in a language, but are not attested in available sources) are distributed randomly throughout the proto-lexicon. If this is the case,

Table 5. Estimates of the number of entries in the dictionaries used in the preparation of the Mixe-Zoquean etymological dictionary, ordered by number of entries.

Language	Number of Entries
LM	7000
ChisZ	6000
NHM	5600
MM	4100
OlP	4000
TxZ	4000
SaP	3600
AyZ	2000
ChZ	1600
SoZ	800
SHM	700

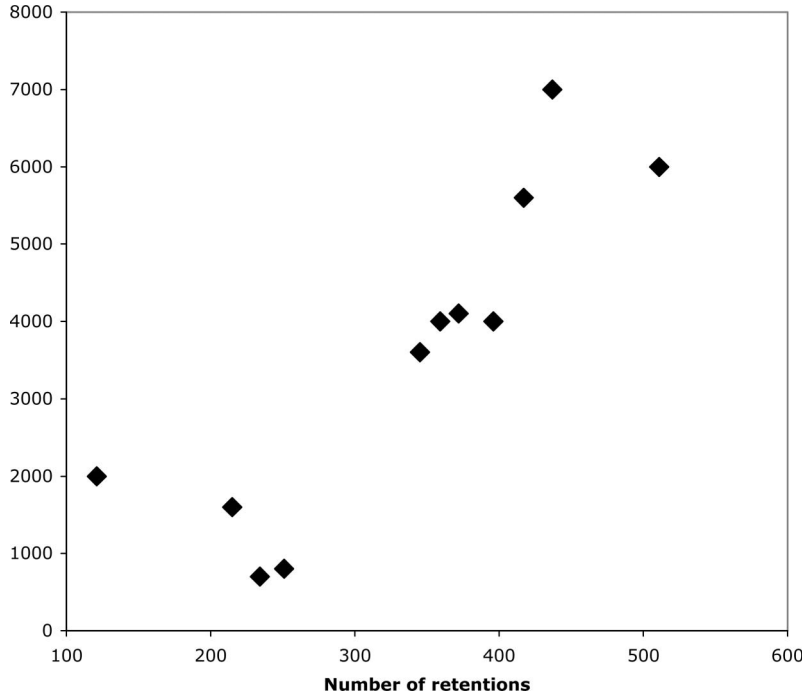


Fig. 7. Correlation between the amount of data available and the number of retentions r for each Mixe-Zoquean language (Pearson's $R^2 = 0.748$).

languages with missing retentions will simply appear to have undergone greater amounts of change than is actually the case, but the method should still produce accurate N estimates. The central question thus is whether the missing retentions are distributed randomly. It would be difficult to answer this question directly. However, the fact that two significant errors in the results of Holm's method are the placement of ChisZ and SHM might be taken as circumstantial evidence that the amount of available data does distort the results: ChisZ is one of the best documented Mixe-Zoquean languages, and SHM one of the poorest.

It is also important to realize that this problem is not restricted to families like Mixe-Zoquean where there is comparatively little information available. The same problem would arise whenever there are varying levels of documentation among languages, even if all the languages in question were amply documented.

5. COMPARISON WITH TRADITIONAL LEXICOSTATISTICS

Holm explicitly created the separation base method as an improvement over lexicostatistics. Addressing his criticisms from a theoretical point of view, we agree that closely related languages may be incorrectly classified if they differ significantly in their respective degrees of retention (the ''proportionality trap''). One should not exaggerate the importance of this problem, since the input data for lexicostatistics also include shared innovations, which at least partially counterbalances its possible effects. Even so, the problem does exist.

In order to judge the impact of this and other potential problems empirically, we have performed lexicostatistical analysis on a set of Mixe-Zoquean data, and our results may be compared to those arrived at under Holm's approach. We had word lists available for ten of the twelve extant Mixe-Zoquean languages, including two variants of Chimalapa Zoque (see Appendix). The languages for which we lack lists are AyZ and MM. The Mixe-Zoquean tree, however, remains largely the same without them. That is, there should be a Gulf Zoquean node whether or not AyZ is represented, and there should be an Oaxaca Mixean node whether or not MM is represented.

We used meaning lists containing translations of 110 different concepts. It was often the case that there was more than one possible translation into a Mixe-Zoquean language. However, in order to increase the relevance of the data to the problem at hand, we have only included translations that have cognates in one of the other languages in the list. Possible variants that do not have any cognates were removed, except in the cases where none of the translations of a particular meaning had cognates in any of the languages. In a few cases, a language had more than one variant with cognates in the other languages. We split these into two entries, resulting in a total of 115 entries. Our judgments of cognacy may be seen in the numerical list following the word list in the Appendix. Cognate forms are coded with identical numbers. The magnitude of the number is of no significance; the only feature coded is cognacy versus non-cognacy. These judgments are based on known phonological changes.

Several entries in the word list are cognate in all ten languages considered. Although such cognate sets might be employed to argue for the unity Mixe-Zoquean as a whole, they do not help in establishing its sub-grouping. As we are here only interested in the sub-grouping, we

have coded more fine-grained distinctions in these entries when possible. In some of them there is a difference in lexical derivation, i.e., some languages have a shared affixed or compound form that is not found in the other languages. For example, the word for 'ashes' clearly contains the root **ham* in all ten languages. However, some languages show further agreement in reflecting **kuy-ham*. We have coded two different cognate sets for this entry, although they are clearly related. Only in such cases with a *complete* cognate set with *lexical* differentiation did we decide to separate the cognates into two groups in order to increase the amount of evidence to be considered for sub-grouping. We did not use phonological criteria to establish such divisions.

To measure similarity between languages, we calculated the ratio of shared cognates to the total of possible shared cognates for each pair of languages. The number of possible shared cognates was not always 115, because not all entries were available for every language. Instead, we counted the number of entries with data available for both of the languages in question. These similarity measurements are summarized in Table 6.

In order to produce a tree, we applied the Fitch algorithm[17] to the matrix of cognate percentages (using one minus the percentages in order to produce a dissimilarity measure). The resulting tree is shown in Figure 8. Although it is an unrooted tree, the divisions strongly agree with Wichmann's (1995) classification. It is clearly a much better representation of the history of Mixe-Zoquean than any of the results produced by Holm's method: the division between the Mixean and Zoquean branches is strongly suggested, and furthermore, the Gulf-Zoquean (SoZ and TxZ) and Oaxaca Mixean languages (LM, SHM and NHM) are correctly placed in subgroups. Two subgroups disagree with Wichmann's (1995) classification, namely ChZ and ChisZ within Zoquean and SaP and OlP within Mixean. However, they split a very short distance after the previous splits, indicating that there is little evidence for these subgroups. Additionally, the Fitch algorithm only produces binary splits, so this result is about the nearest one can get to showing a three-way split, which is what Wichmann (1995) argued for. Recall also that a subgroup for ChZ and ChisZ was proposed by Kaufman.

[17]We used the `fitch` program from the `phylip` package, which implements this algorithm.

Table 6. Percentages of cognates based on a 110-word list for 10 Mixe-Zoquean languages.

	ChZ (Mar)	ChZ (Mig)	LM	NHM	OlP	SaP	SHM	SoZ	TxZ
ChisZ	65/95 = 0.684	72/107 = 0.673	41/108 = 0.380	43/109 = 0.394	43/103 = 0.417	45/106 = 0.425	43/109 = 0.394	59/102 = 0.578	69/106 = 0.651
ChZ (Mar)		82/98 = 0.837	38/95 = 0.400	36/96 = 0.375	33/92 = 0.359	41/92 = 0.446	39/96 = 0.406	55/92 = 0.598	64/93 = 0.688
ChZ (Mig)			37/107 = 0.346	37/108 = 0.343	36/102 = 0.353	40/104 = 0.385	39/108 = 0.361	59/102 = 0.578	63/104 = 0.606
LM				80/110 = 0.727	60/102 = 0.588	65/105 = 0.619	89/110 = 0.809	46/102 = 0.451	43/105 = 0.410
NHM					65/103 = 0.631	68/106 = 0.642	87/110 = 0.791	36/103 = 0.350	38/106 = 0.358
OlP						69/102 = 0.676	65/103 = 0.631	39/98 = 0.398	38/99 = 0.384
SaP							67/106 = 0.632	41/99 = 0.414	44/102 = 0.431
SHM								41/103 = 0.398	41/106 = 0.387
SoZ									81/100 = 0.810

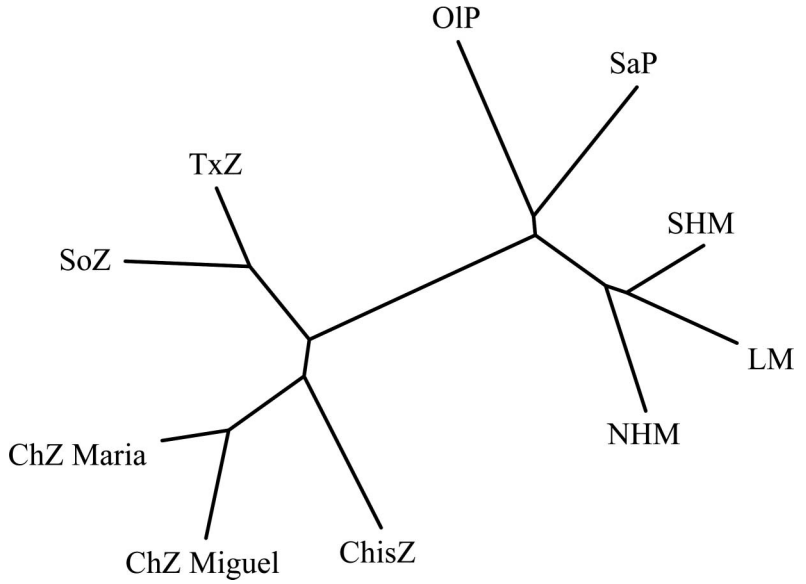


Fig. 8. Mixe-Zoquean tree based on the word list, using the Fitch algorithm to infer the tree.

The format of our word list data is, it so happens, also the format needed to apply an algorithm that has become widely used to infer phylogenies in biology: maximum parsimony. Maximum parsimony methods operate on data of the form given in Appendix B, normally referred to as "character" data, not on secondary data such as shared cognate percentages or other similarity measures. Using character data has the advantage of taking into account all available information, which is not the case with similarity methods (although the loss of information by transforming the data into a similarity matrix is apparently not that large, cf. Felsenstein, 2004, p. 147). Maximum parsimony methods search for trees that minimize the number of changes (innovations) that must have occurred in the development from the original proto-state to the state represented by the character data.

There are various maximum parsimony methods. We have used the Wagner parsimony method (Kluge, 1969, pp. 6–8), in which every change of state is taken to be equally likely. This means that for every entry in our word list, the change from any one cognate set into any other cognate set is equally likely. This is a reasonable assumption, because our word lists contain no claims about which forms represent reflexes of

proto-forms. We have to assume that every attested lexeme could reflect the original form, and thus that every possible change between the various lexemes in our list could have occurred.[18] The result of applying Wagner's Maximum Parsimony to our word list is shown in Figure 9.[19] Again, the split between Mixean and Zoquean can clearly be seen in the unrooted tree. The Gulf Zoquean and Oaxaca Mixean groups are also distinct. As before, ChZ and ChisZ form a subgroup, agreeing with Kaufman's but not Wichmann's tree. Interestingly, SaP and OlP do not form a group at all in this tree. Instead, OlP is incorrectly grouped with the Oaxaca Mixean languages. Neither Wichmann's nor Kaufman's classification calls for such a group. The result of the Maximum Parsimony analysis is not quite as good as the Fitch analysis of cognate percentages (Fig. 8). This could indicate the unsuitability of Maximum Parsimony methods for linguistic data (cf. Wichmann & Saunders, 2005). However, the tree is still a clear improvement over the results of Holm's method.

Finally, in Figure 10, we present a different way of depicting the structure of the dissimilarity matrix. This figure shows a split decomposition tree (Bandelt & Dress, 1992).[20] In such a tree, conflicting evidence is shown in the form of criss-crossing lines, so-called reticulations. Various possible trees are presented simultaneously, with the length of the branches indicating how much evidence there is for a particular group. For example, looking at the Mixean subgroup in the lower right, it can be seen that there is good evidence for Oaxaca Mixean (in the form of the length of the box separating LM, SHM, and NHM from the rest), but also a little bit of conflicting evidence grouping SaP together with NHM (in the form of the depth of box that separates LM, SHM, and NHM from the rest). The evidence for grouping OlP with the Oaxaca Mixean languages (as was done by the Wagner method above) can be discerned in the form of the little boxes at the division between OlP, SaP, and Oaxaca Mixean. The small size of the boxes shows that the amount of evidence for this

[18]In this case, detailed comparative work has been done, and we have good evidence for proto-forms in many cases. However, we are trying here to simulate a situation like that in which traditional lexicostatistics is usually applied, where such work has not yet been carried out.

[19]For this analysis, we used the `pars` program from the `phylip` package.

[20]This particular graph was made with the `SplitsTree` program, using the NeighborNet algorithm developed by Bryant and Moulton (2002).

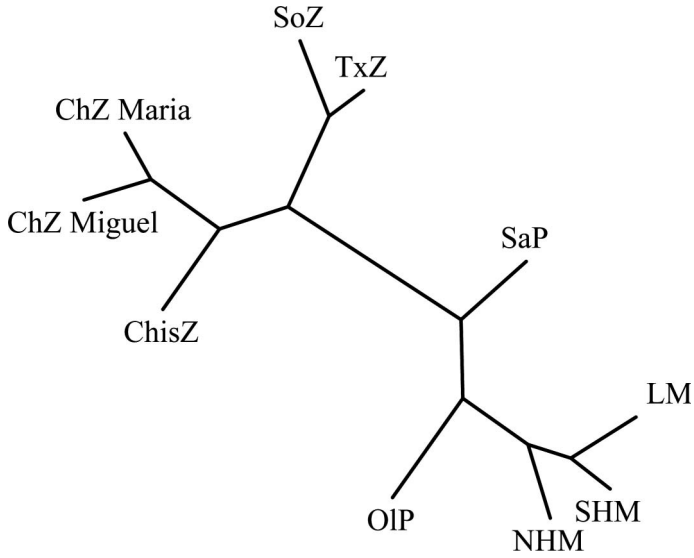


Fig. 9. Mixe-Zoquean tree based on the word list data, using the Wagner Maximum Parsimony algorithm to infer the tree.

subgroup is extremely small. Overall, the split decomposition tree in Figure 10 is an almost exact representation of the tree proposed by Wichmann (1995).

6. CONCLUSION

Holm's sub-grouping method did not give adequate results for our test case of Mixe-Zoquean. We have identified various reasons why the method failed, all of which focused on its empirical basis, namely etymological dictionaries, which are never complete or ideal. The method might work if we had perfect knowledge about the ancient heritage of a family, but this is clearly never the case. The establishment of proto-forms for a group of languages is a complex issue, and various circumstantial factors influence this process, such as preconceptions about the structure of the family and the amount of available data. All such factors have an influence on the results of Holm's method.

In conclusion, we believe that although Holm has a novel and interesting approach to the subgrouping problem, in practice it does not yield

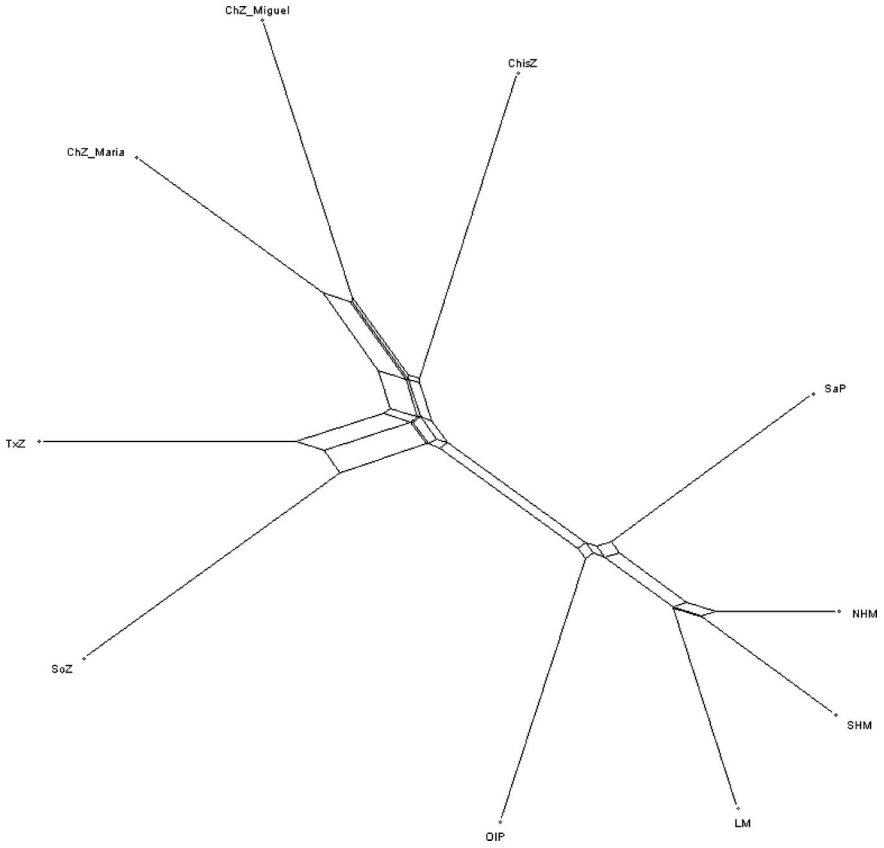


Fig. 10. Split decomposition tree of the Mixe-Zoquean family on the basis of the word list data, using character-wise coding and the NeighborNet algorithm to infer the tree.

good results. Furthermore, although his criticisms of lexicostatistics are valid in principle, the results we obtained from lexicostatistic analysis of a short word list were surprisingly good.

REFERENCES

Bandelt, H.-J., & Dress, A. W. M. (1992). Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, *1–3*, 242–252.

Bird, N. (1982). *The Distribution of Indo-European Root Morphemes*. Wiesbaden: Harrassowitz.

Brainard, B. (1970). A stochastic process related to language change. *Journal of Applied Probability*, *7*(1), 69–78.

Bryant, D., & Moulton, V. (2002). NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. *Proceedings of the Workshop in Algorithms for Bioinformatics* (http://www.mcb.mcgill.ca/~bryant/NeighborNet/preprint.pdf).

Clark, L. E. (1981). *Diccionario Popoluca de Oluta*. México, D.F.: Instituto Lingüístico de Verano.

Clark, L., & de Clark, N. D. (1960). *Vocabulario Popoluca de Sayula*. México, D.F.: Instituto Lingüístico de Verano.

Davies, P., & Ross, A. S. C. (1975). "Close relationship" in the Uralian languages. *Finnisch-Ugrische Forschungen*, *14*(1–3), 25–48.

Elson, B. F., & Gutiérrez, D. G. (1999). *Diccionario popoluca de la Sierra Veracruz*. Coyocán, D.F.: Instituto Lingüístico de Verano.

Embleton, Sh. (1986). *Statistics in Historical Linguistics. Quantitative Linguistics 30*. Bochum: Brockmeyer.

Embleton, Sh. (1991). Mathematical methods of genetic classification. In S. M. Lamb & E. D. Mitchell (Eds), *Sprung from a Common Source: Investigations into the Prehistory of Languages* (pp. 365–88). Stanford, CA: Stanford University Press.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, *155*(3760), 279–284.

Foster, G. M. (1943). The geographical, linguistic, and cultural position of the Popoluca of Veracruz. *American Anthropologist*, *45*, 532–546.

Harrison, W. R., Harrison, M. B., & Garcia, C. H. (1981). *Diccionario Zoque de Copainalá*. México, D.F.: Instituto Lingüístico de Verano.

Holm, H. (2000). Genealogy of the main Indo-European branches: Applying the separation base method. *Journal of Quantitative Linguistics*, *7.2*, 73–95.

Holm, H. (2003). The proportionality trap, or: what is wrong with lexicostatistical subgrouping. *Indogermanische Forschungen*, *108*, 38–46.

Holm, H. (2004). The new arboretum of Indo-European trees. Unpublished manuscript.

Holm, H. (2005). Genealogische Verwandtschaft. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds), *Quantitative Linguistics. Handbooks of Linguistics and Communication Science 27* (pp. 633–645). Berlin: Walter de Gruyter.

Hoogshagen, N. S., & de Hoogshagen, H. H. (1993). *Diccionario mixe de Coatlán*. México, D.F.: Instituto Lingüístico de Verano.

Johnson, H. A. (1998). *San Miguel Chimalapa Soke online dictionary*. Accessed December 2004 (http://www.albany.edu/anthro/maldp/mig.html).

Johnson, N. L., Kotz, S., & Kemp, A. W. (1993). *Univariate Discrete Distributions*. New York: John Wiley & Sons.

Kaufman, T. (1964a). Mixe-Zoque subgroups and the position of Tapachulteco. *Actas y memorias del XXXV Congreso Internacional de Americanistas*, Mexico City, 1962, Vol. II, 403–411.

Kaufman, T. (1964b). *Mixe-Zoque Diachronic Studies*. Unpublished manuscript.

Kaufman, T. (1974). Meso-American Indian Languages. *The New Encyclopædia Brittanica*. 15th ed. Vol. 22, pp. 767–774.

Kaufman, T., & Justeson, J. (2004). Epi-Olmec. In R. D. Woodard (Ed.), *The Cambridge Encyclopedia of the World's Ancient Languages* (pp. 1071–1111). Cambridge: Cambridge University Press.

Kendall, D. G. (1950). Reply to Professor Ross's paper. *Journal of the Royal Statistical Society, Series B (Methodological)*, *12*(1), 41–42.

Kluge, A. G., & Farris, J. S. (1969). Quantitative phyletics an the evolution of Anurans. *Systematic Zoology*, *18*(1), 1–32.

Knudson, L. M., Jr. (1980). *Zoque de Chimalapa Oaxaca*. México, D.F.: Centro de Investigación para la Integración Social.

Nordell, N. (1962). On the status of Popoluca in Zoque-Mixe. *International Journal of American Linguistics*, *28*, 146–149.

Pokorny, J. 1994 [1948–1959]. *Indogermanisches etymologisches Wörterbuch*. 3rd ed. Tübingen: Francke.

Saitou, N., & Nei, M. (1987). The neighbour-joining method: a new method for reconstructiong phylogenetic trees. *Moecular Biology and Evolution*, *4*(4), 406–425.

Sattah, Sh., & Tversky, A. (1977). Additive similarity trees. *Psychometrica*, *42*(3), 319–345.

Schoenhals, A., & Schoenhals, L. C. (1965). *Vocabulario Mixe de Totonpec. Mixe-Castellano. Castellano-Mixe*. México, D.F.: Instituto Lingüístico de Verano.

Wichmann, S. (1994). Mixe-Zoquean linguistics: A status report. In D. Bartholomew, Y. Lastra & L. Manrique (Eds), *Panorama de los estudios de las lenguas indígenas de México, vol. 1* (pp. 193–267). Biblioteca Abya-Yala, *16*. Quito: Abya-Yala.

Wichmann, S. (1995). *The Relationship among the Mixe-Zoquean languages of Mexico*. Salt Lake City: University of Utah Press.

Wichmann, S. (1999). On the relationship between UtoAztecan and MixeZoquean. *Kansas Working Papers in Linguistics*, *24*(2), 101–113.

Wichmann, S. (2002). *Diccionario analítico del popoluca de Texistepec*. México, D.F.: Universidad Nacional Autónoma de México.

Wichmann, S. (2003). Contextualizing proto-languages, homelands and distant genetic relationship: Some reflections on the comparative method from a Mesoamerican perspective. In P. Bellwood & C. Renfrew (Eds), *Examining the Farming/Language Dispersal Hypothesis* (pp. 321–29). McDonald Institute Monographs. Cambridge: McDonald Institute for Archaeological Research.

Wichmann, S., & Saunders, A. (2005). How to use typological databases in historical linguistic research. Manuscript under review. A preliminary version is available at: http://email.eva.mpg.de/~wichmann/wichmann_publ.html

SOFTWARE USED

Phylip by Joseph Felsenstein. Available online at: <http://evolution.gs.washington.edu/phylip.html>.

SplitsTree 4 (ß 14) by Daniel H. Huson and David Bryant. Available online at: <http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome.html>.

T-Rex by Vladimir Makarenkov. Available online at: <http://www.info.uqam.ca/~makarenv/trex.html>.

APPENDIX: MIXE-ZOQUEAN WORDLISTS

The word lists of this appendix were assembled from the following sources:

NHM (Totontepec):	Schoenhals and Schoenhals (1965)
SHM (Ayutla):	Personal communication from Yásnaya Elena Aguilar (native speaker)
LM (Coatlán):	Hoogshagen and Hoogshagen (1993)
SaP:	Clark and Clark (1960)
OlP:	Clark (1981)
TxZ:	Wichmann (2002)
ChisZ (Copainalá):	Harrison, Harrison and García (1981)
SoZ:	Elson and Gutiérrez (1999)
ChZ Maria:	Knudson (1980)
ChZ Miguel:	Johnson (1998)

Albert Davletshin was instrumental in the initial compilation of the lists from NHM, SaP, OlP, TxP, and ChisZ, for which we thank him heartily. The 110 item list, which consists of the 100-item Swadesh lists plus a 10-item addition by Yakhontov, is one currently used by Sergei Starostin and his collaborators. Most of the forms in the lists have subsequently been checked by us and we have added the rest of the lists.

We have standardized the orthographies using standard Americanist notation. However, vowel length is represented by doubled vowel symbols. The orthographic changes are straightforward and need not be indicated here, the only exceptions being the NHM vowels, for which we have made the following conversions with respect to the original source: e̲ → ɛ, u̲ → ï, a̲ → ə, o → ʊ, o̲ → o (the rest of the vowel symbols being unchanged), and the SHM vowels, where we have made the following conversions: ë → ə, a → æ, ä → a. These changes are intended to bring the vowel orthographies somewhat closer to IPA notation. Morpheme breaks that are clearly identifiable synchronically are indicated by a hyphen. ''Synchronically identifiable'' means that the morphemes to both sides of the hyphen are known by us to recur in the given language and, furthermore, that we understand their contribution to the semantics of the particular form. Morpheme breaks are not found in the original sources except in the case of TxZ. Verbs are cited as roots, stripped of inflectional

morphology, and phonological (ablaut) alternants are given, separated by a tilde.

In the complete versions of the lists there are regularly multiple possible translations of single concepts. When one of the translations, but not the other(s), was matched by a form in one or more of the other languages, we removed the alternative translation(s). Only when multiple cognate sets could be established from the alternatives of two or more languages or when there appeared to be no cognates did we leave all possible translations in the lists as given below.

Following the wordlist table there is a table where we have coded the cognates within each line of the word lists. Cognate forms are given identical numbers, but the ranks of these numbers are completely arbitrary and are not taken into account in the calculations. *NA* stands for "not attested". In a number of cases, all forms were cognate across the ten Mixe-Zoquean languages under consideration. If in such cases there were nevertheless lexical differences shared among a subset of the languages (e.g., the presence vs. absence of a derivational affix or a compounded element), these shared deviating forms were coded as belonging to a separate cognate set (see for example the translations of 'ashes'). This procedure served the purpose of extracting the maximum amount of information for subgrouping purposes. Otherwise, that is when compounded elements or derivational affixes were not shared among two or more languages, such elements were ignored when assessing relatedness. The decisions regarding the identification of cognates were informed by knowledge of regular sound changes, although in a few cases the lack of complete understanding of the etymology of a given form forced us to make somewhat more subjective decisions.

Appendix Table 1. 110-Wordlist for 10 Mixe-Zoqean Languages.

Meaning	NHM	SHM	LM	SaP	OlP	TxP	SoZ	ChZ Maria	ChZ Miguel	ChisZ
all	nəhʊm	tukəʔy	tuʔuk-ʔokəy	nu pəhən	niʔtiʔk	bumbe	ikuməh	aŋteneʔ	hemhiʔ	mumu
ashes	haahm	haam	kuy-haam	kuy-hahm	hamə	kuyam	kuy-ham	kuy-ham	kuy-ham	kuy-ham
bark	kĭp ak	ak	ak	ak	?	kuy-daak	?	naka	kuy naka	kuʔyu-naka
belly	tiints	mææts	moʔošak	tin-čay	puʔpu	tsek	?	tsek	tsek	tsek
big	məh	məh	məh	məh	taʔnək	wəəčeʔ	məh	wətiʔ	komiʔ	məha
bird	həəyva	hoon	hoon	hohn	muuši	hon	hon	hon	hon	paloma
bite	tsuʔuts	tsuʔuts	noots	kutoʔ	tsuʔts	keʔts	was	?	waš	keʔts
black	yək	yəhk	yik	yəgək	yəkəʔk	yək	yək	yək	yək-yək	yək
blood	nĭĭʔpun	nəʔəpy	nəʔpy	nipin	nəʔpiʔn	dəʔəpeñ	nəʔəpiñ	nəʔpin	nəpinʔ	nəʔpin
bone	pahk	pahk	pahk	pahk	pakə	pak	pak	pak	pak	pak
breast	hʊʔtaa	tsetsk	tsiʔits-ʔaniʔikš	moʔy	?	duun	nuunu	?	tsuʔts-iʔ	tsuʔtsi
burn (tr.)	tooy ~ toohy ~ toy	tsaʔay	noʔok	ak-toy	yak-toy	doʔ-tokeñ	noʔ	kəŋ	poŋ	poŋ
claw (nail)	šəəʔk	šooky	šooky	šoohk	šooki	kəʔ-tsuʔks	kətsəs	kətsus	kəʔtsuš	kəʔčus
cloud	vinĭ̈ts	nəmaʔa	hok	win-tuk	vin-hoyeʔk	ʔuʔksuʔ	uksə	?	ʔin-əʔ	oʔna
cold	šuš	šuš	təč	payik	tušəʔk	paak	pagak	wayaʔy	wayayʔ	pakak
come	mɛts ~ mɛhts	men	min	min ~ miʔn	min ~ miʔn	beñ	miñ	min	min	min
die	ooʔk	ook	oʔok	oʔk	ook	kaʔ	kaʔ	kaʔh	kaʔ	kaʔ
dog	ʊk	uk	uk	tak	šuʔni	čempaʔ	čimpa	nuʔu	nuʔ	tuwi
drink	ʊʊk	uuk	uk	uuk ~ uʔk	uuk	ʔuk	uk	uk	ʔuk	uhk
dry	təʔəhts	təʔəts	təʔəts	kuytəʔts	təʔəʔs	təts-kaʔ	ku-təts	yak-yəts	təts-iʔ	təʔts
ear	taatsk	tatsk	tətsʔak	taatsək	taatsəək	taatsək	taatsək	tatsək	tatsək	tatsək
earth	naaš	naš-wiʔñət	naaš	naaš	naašə	das	nas	nas	naš	nas
eat	kay	kæy	kay	kay	kay	kəʔəs	kəʔəs	kəʔs-oy	kəš	kəʔs
egg	tuʔut	tutsah	tuʔuty	kooy	tuʔtiʔk	kaʔnpuk	kanpu	poʔok	pohoʔk	poka
eye	viihn	ween	wiin	wihn	vinə	ʔeš-kuʔ	iš-kuy	witəm	wotəm	win
far	hekum	hækæm	həgem	yagats	heškeʔkeʔk	huʔum	huʔumə	yaʔhi	yah-əʔ	yaʔay
fat (n.)	on	on	wiin-ʔomb	oʔtsək	ona	?	pəhi	grasa	yew-eʔ	kiʔnə
feather	pahk	pəhk	tseyñiʔi-bəhk	?	pəkə	pək	pək	pək	pək	pək
fire	həəhn	həən	həən	həhn	hənə	hukut	huktə	hukutə	ʔaʔts-əʔ	hukətək
fish	ahkš	akš	akš	akš	koʔke	woʔŋ	təʔəpə	koke	koke	kokeh
fly (v.)	hɛy ~ hɛɛy	kæhkəʔək	keek	keek ~ keʔk	napap	kek	kek	kek	yuk-kek	sitiht
foot	tek	teky	teky	tan	kəʔšta	pak	puy	maŋ-kuʔy	maŋ-kuyʔ	neʔŋ-bak
full	uts	uhts	kuʔušy	patsik	utsi	kus-deʔ	kus-ne	yag-aŋ-tas	taš-əʔ	hoʔəyupə
give	məʔə	yak	moʔ	moy ~ moʔ	moy ~ moʔ	čeʔ	čiʔ	tsiʔ	tsiʔ	čiʔ
good	oy	ey	oy	oy	(ni)oya	wəə	wəə	?	wəhə	oye
green	tsʊšk	tsušk	tsuš	šušuk	tušuʔk	tsuus	tsuus	tsuhus	tsuš	tsuhtsu
hair (1)	vaahy	waay	waay			waʔy	way	waʔy	wayʔ	way
hair (2)				pəhk	pəkə					pək
hand	kəʔə	kəʔə	kəʔ	kəʔ	kəʔ	kəʔ	kə	kəʔ	tsaŋ-kuyʔ	kəʔ
head	kuvahk	kəpaahk	kowahk	kopak	koʔpaʔk	kopək	koobak	waʔy	kopak	kopak
hear	motu	mətoo	mədoogugəʔ	maraw	motov	batəŋ	aŋ-təən	matoŋ	matoŋ	matoŋ

(*continued*)

Appendix Table 1. (*Continued*).

Meaning	NHM	SHM	LM	SaP	OlP	TxP	SoZ	ChZ Maria	ChZ Miguel	ChisZ
heart	haʔvin	anmə-haʔan	huuky-hot	?	hutukotə	tsookoʔ	aanamah	tsokoʔy	tsokoyʔ	tsokoy
heavy	hemïts	heʔmšy	hemšy	uʔts	?	tsaʔks-naʔ	tsakš-ñi	hemets	hemeʔts	hemets
horn	vah	wah	muʔn	wah	vaha	waaŋaʔ	?	?	wekaʔ	waʔ
I	əts	əts	əəč	əə, əəts	əəs	ʔətsəʔ	əč	ətstsiʔ	də-š	əh
kill	yak-ʔook ~ yak-ʔooʔk	æk-ook	yah-ʔoʔok	ag-oʔk	yak-ook	yakaʔ	ik-kaʔ	yak-kaʔh	yak-ʔoy	yah-kaʔ
knee	koš	koš	koškohwahk	košk	košo	koosoʔ	kooso	koʔma	komaʔ	tungupyoʔk
know	nəhava	naw	nehwəəy	hawi	viniy	bus	hoodoʔŋ	mus	muš	muhs
leaf	aahy	aay	aay	ahy	ayə	ʔay	ay	ay	ʔayʔ	ay
lie	maʔva	koʔoky	koʔok	kəmna	šok	kas-deʔ	wooneʔe	moŋ-kətake	moŋ	əkgehk
liver	nəə hoot	hoht	hot	hooht	təʔshotə	tsookoʔ	tsoogoy	tsokoʔy	paʔt	paʔt
long	yən	yeny	yoony	yagats	yoneʔk	yaʔakaʔts	yagats	pəhi	pəh-iʔ	pəy
louse	əəʔts	aʔaht	aad	aawat	aavaʔt	awaʔt	aʔawat	awat	ʔawat	awat
man	yaaʔtyəhk	hətyəhk	yeʔey-dəhk	hayaw	yoʔohwa	pəəñ	pəəšiñ	pən	pən	pən
many	may	mæy	may	may	paʔko, taʔna	dač-ke, bəks-ke, səʔəs-daʔa	agi, tsaam, haʔyaŋ, waat^{y}i	səhəʔ	šəh-əʔ	wəwə, aŋ-baʔt-e, məʔčəki
meat	tsuʔuts	tsuʔučy	tsuʔuč	šiš	tsuʔči	šeš	maayi	sis	šiš	sis
moon	poʔo	poʔo	poʔ	poʔ	poʔa	pooy	pooya	sepeʔ	šepeʔ	poyah
mountain	kopk	kokp	koʔogop, kop	kopak	kopaʔk	kotsək	kootsək	?	kotsək	kotsək
mouth	aah	aa	aawak	ahw	avə	həp	həp	həp	həp	aŋnaka
name	šəə	šəə	šəə	nəʔhy	šəvə	dəy	nəəyi	nəhi	nəhi	nəy
near	təmhi	hənkon	wingon	tom	tomeʔk	taʔan	taʔañi	tomeʔ	tom-eʔ	tome
neck	yoʔkt	yoʔkt	yoʔok-pahk	yoʔk	?	kən-kəʔ	?	?	wintuʔ	kənə
new	nam	hempy	hemy	namay	namə	homə-naʔ	hoomi	homeʔ	homeʔ	home
night	tsʊʊ	koots	adzuuʔəʔ	tsuʔ	tsuhə	tsuʔ	tsuʔu	tsuʔhi	tsuʔ	tsuʔ
nose	həhp	həhp	həhp-put	həhp	həpə	keñ-kuy	kiiñi	kinə	kinək	kinə
not	kaʔa	kaʔt	kaʔ	kaah	əʔə	eñč	d^{y}a	wat kaha	ya	haʔne
one	tʊʔk	tuʔuk	tuʔuk	tuʔk	tuʔk	tum	tuum	tumə	tumə	tumə
person	hayu	haʔay	həy	hayaw	haykaʔk	pəəñ-čaʔ	?	?	?	pən-daʔm
rain	tʊʊ	tuh	tuu	tuu	tuuhə	tuh	tuh	tuh	tuh	tuh
red	tsapts	tsaps	tsapts	tsabats	tsapaʔs	tsaʔpaʔts	tsabats	tsapats	tsapats	tsapas
road	tʊʊʔ	tuʔu	tuʔ	toow	tuʔaʔw	tuŋ	tuŋ	tuŋ	tuŋ	tuŋ
root	aaʔts	aats	tikts	tikčik	čistsəʔw	čečək	t^{y}iičək	waʔts	waʔts	watsi
round	pigïk	pehk	pik	hoyoy	?	woy-woy-če	woyo	?	?	huyuh
salt	kaan	kaan	kaan	kaan	kaana	kaan	kaana	kana	kana	kana
sand	pʊʔʊ	puʔu	puʔ	poʔy	puʔə	poʔoy	poʔoy	haweʔ	poeʔ	waʔna
say	vaʔañ	nə-kapš	mənaʔn	nəm	nəm	dəm	nəm	nəm	nəm	nəm
see (1)	iš	eš	iš	iš		ʔeš	iʔiš	is	ʔiš	ʔis
see (2)				eʔp	ʔeep					
seed	təəmt	pahk	pahk	puuh	pakə	puh	pak	puh	puh	puh
short	kona	kon	kon	?	konoʔk	dokkoʔ	noko	konoʔ	kon(n)-oʔ	kono
sit	tsïïna	tsəən	inyəy, ištak	tsəəna	huunii	kon	koñ	tsənay	tsən	poks
skin	ak	nəʔak	ak	ak	akə	daak	naaka	naka	naka	naka

(*continued*)

Appendix Table 1. (*Continued*).

Meaning	NHM	SHM	LM	SaP	OlP	TxP	SoZ	ChZ Maria	ChZ Miguel	ChisZ
sleep	maʔa	maʔa	maʔ	maa ~ maʔ	maa ~ maah ~ maʔh	boŋ	moŋ	moŋ	kap-ney	əŋ
small	piʔk	mutsk	muts	čičic	čuuču	həsə-baa	šutʸu	?	wakšiŋʔ	čiks
smoke	hok	hok	həən-hok	wiʔšik	hoko	hook	hooko	oʔmaʔ	ʔomaʔ	hokoh
snake	tsaaʔn	tsaʔany	tsaʔny	tsanay	tsanaʔy	tsaañ	tsaañ	tsahin	tsahin	tsan
stand (1)	tena	tənæy	tenaay	tena	teniy	hoop-deʔ	teeñ		ten	tenay
stand (2)								maʔŋəy	maŋ-ney	
star	maatsa	matsaʔa	madzaʔ	maahts	maatsaʔaʔk	baatsaʔ	maatsa	matsaʔ	matsaʔ	matsa
stone	tsaa	tsah	tsaa	tsaʔ	tsaahə	tsaʔ	tsa	tsaʔ	tsaʔ	tsaʔ
sun	šəə	šəʔan	šəə	šəhw	šəvə	haam	haama	hama	hama	hama
swim	nə-vïïʔn	nə-yum	yaab	yun	yun	pəʔn	puʔn	yun	yuk	hemu
tail	tuʔšta	tutst	piʔišy	tuʔhts	tuʔsta	tuʔts	tuʔts	tuʔts	tuʔts	tuʔts
that	šiiʔ	hæʔæ	hehəmbə	yahyəʔn	hamah	?	heam	?	ga.də	əʔ
thin	pehi	pehy	pehy	pehay	hətəʔt, niipakə, pakə	waayeʔ, əʔks-kəʔdaʔa, kal-pak-daʔa	čehče, wayay	piʔiʔ	kitsa	tseya, kay, pahka-pahka
this	yəʔə	yaʔat	hadaʔ	ayəəh	heʔ	yəʔəhəʔ	heʔepək	yəʔ-pəʔ	yəʔ	yəʔ-wə
thou	mits	mehts	miič	mii(č)	miis	bič	mič	mits(tsíʔ)	mi-š	mih
tongue	toots	toots	toots	toohts	tootsə	?	tots	yen-kuʔy	tots	tots
tooth	təəts	təəts	təəts	təəhts	təətsə	təts	təts	təts	təts	təts
tree	kïp	šoh	kipy	kuhy	kuyə	kuy	kuy	kuy	kuy	kuy
two	mehtsk	mætsk	mets	mečk	mes-ko	wəs-naʔ	wəs-teen	mets-aʔaŋ	mets-aʔaŋ	metsa
walk/go (1)	yoʔoy	yoʔoy	yoʔoy	yoʔy	vit	weč	witʸ		wit	wiht
walk/go (2)								tuʔŋah	tuŋ-ʔah	
warm	an	an	an	powik, toyap	hokoʔš	?	?	nuts	yuk-nuts	pih, yoʔk-a
water	nəə	nəh	nəə	nəʔ	nəəhə	dəʔ	nə	nəʔ	nəʔ	nəʔ
we	əəts	əəts	əəč aht	əhtsat	ətsaateʔk	ʔəts-taʔ	?	newinynen	neywin	təh
what	ti	tii	ti	ti	ti	čeʔ	tʸi	ti	ti	tiyə
white	pooʔp	poop	poob	poʔp	poopoʔ	popoʔ	poʔobaap	popoʔ	popoʔ	popo
who	pən	pən	pən	pən	pən	ʔee	iapaʔap	i-wə	ʔi-wə	i-wə
wind	poh	poh	poh	šeʔm	haamu	saaw	saawa	?	šawa	sawa
woman	təʔəštəhk	tišytyəhk	toʔošyhəy	toʔšay	mahaw	yoomaʔ	yoomo	yomaʔ	yomaaʔ	yomo
worm (1)	naš-tsaaʔn	tənəʔək	təng	tsugut	kumu	tsuʔukeñ	tsuuʔkiñ	tsuʔkin	kumuʔ	
worm (2)						toʔt				toʔt
year	huuhnt	həməht	həməht	šiwit	šiviʔt	ʔamče	aamtʸəy	amintə	ʔamintəʔ	ame
yellow	puʔts	puts	puʔts	puʔts-puts	puʔs-puʔs	puʔuč-eʔ	puʔuč	puʔtsiʔ	puʔts-puʔts	puʔtsə

Appendix Table 2. Coding of cognate sets.

Meaning	NHM	SHM	LM	SaP	OlP	TxP	SoZ	ChZ Maria	ChZ Miguel	ChisZ
all	1	2	2	3	4	5	6	7	8	5
ashes	2	2	1	1	2	1	1	1	1	1
bark	1	1	1	1	NA	2	NA	2	2	2
belly	1	5	2	1	3	4	NA	4	4	4
big	1	1	1	1	2	3	1	3	5	1
bird	1	2	2	2	3	2	2	2	2	4
bite	1	1	2	3	4	5	6	NA	6	5
black	1	1	1	2	2	1	1	1	1	1
blood	1	1	1	1	1	1	1	1	1	1
bone	1	1	1	1	1	1	1	1	1	1
breast	1	2	2	3	NA	4	4	NA	5	5
burn (tr.)	1	8	2	3	3	5	6	7	4	4
claw/nail	1	1	1	1	1	2	2	2	2	2
cloud	1	5	2	1	1	4	4	NA	1	3
cold	1	1	2	3	4	5	5	6	6	5
come	1	2	2	3	3	2	2	2	2	2
die	1	1	1	1	1	2	2	2	2	2
dog	1	1	1	2	3	5	5	6	6	4
drink	1	1	1	2	2	1	1	1	1	1
dry	1	1	1	1	1	1	1	1	1	1
ear	1	1	1	1	1	1	1	1	1	1
earth	1	1	1	1	1	1	1	1	1	1
eat	1	2	2	2	2	3	3	3	3	3
egg	1	1	1	2	1	4	4	3	3	3
eye	1	1	1	1	1	2	2	3	3	1
far	1	1	1	2	3	5	5	4	4	4
fat (n.)	1	1	2	3	1	NA	5	6	7	4
feather	1	1	2	NA	1	1	1	1	1	1
fire	1	1	NA	1	1	2	2	2	3	2
fish	1	1	1	1	2	3	4	2	2	2
fly (v.)	1	2	2	2	3	2	2	2	2	4
foot	1	1	1	2	3	5	6	7	7	4
full	1	1	2	3	1	6	6	7	7	5
give	1	3	1	1	1	2	2	2	2	2
good	1	1	1	1	1	2	2	NA	2	1
green	1	1	1	1	1	1	1	1	1	1
hair (1)	1	1	1	NA	NA	1	1	1	1	1
hair (2)	NA	NA	NA	1	1	NA	NA	NA	NA	1
hand	1	1	1	1	1	1	1	1	2	1
head	1	1	1	1	1	1	1	2	1	1
hear	1	1	1	1	1	1	2	1	1	1

(*continued*)

Appendix Table 2. (*Continued*).

Meaning	NHM	SHM	LM	SaP	OlP	TxP	SoZ	ChZ Maria	ChZ Miguel	ChisZ
heart	1	4	2	NA	2	3	4	3	3	3
heavy	1	2	2	3	NA	2	2	1	1	1
horn	1	1	2	1	1	1	NA	NA	3	1
I	1	1	1	1	1	1	1	1	2	1
kill	1	1	1	1	1	2	2	2	2	2
knee	1	1	1	1	1	1	1	2	2	3
know	1	1	2	2	2	3	4	3	3	3
leaf	1	1	1	1	1	1	1	1	1	1
lie	1	2	2	3	4	5	6	7	7	8
liver	1	1	1	1	1	1	1	1	2	2
long	1	1	2	3	4	3	3	5	5	5
louse	1	2	1	2	2	2	2	2	2	2
man	1	1	1	2	3	4	4	4	4	4
many	1	1	1	1	2	4	5	6	6	3
meat	1	1	1	2	1	2	3	2	2	2
moon	1	1	1	1	1	1	1	2	2	1
mountain	1	1	1	1	1	2	2	NA	2	2
mouth	1	1	1	1	1	3	3	3	3	2
name	1	1	1	2	1	2	2	2	2	2
near	2	1	1	2	2	3	3	2	2	2
neck	1	1	1	1	NA	2	NA	NA	3	2
new	1	2	2	1	1	2	2	2	2	2
night	2	2	1	1	1	1	1	1	1	1
nose	1	1	1	1	1	2	2	2	2	2
not	1	1	1	1	2	3	4	1	4	4
one	1	1	1	1	1	2	2	2	2	2
person	1	1	1	1	1	2	NA	NA	NA	2
rain	1	1	1	1	1	1	1	1	1	1
red	1	1	1	1	1	1	1	1	1	1
road	1	1	1	1	1	1	1	1	1	1
root	1	1	2	2	2	2	2	3	3	3
round	1	1	1	2	NA	3	3	NA	NA	2
salt	1	1	1	1	1	1	1	1	1	1
sand	1	1	1	1	1	1	1	3	1	2
say	1	5	2	3	3	4	4	4	4	4
see (1)	1	1	1	1	NA	1	1	1	1	3
see (2)	NA	NA	NA	1	1	NA	NA	NA	NA	NA
seed	1	2	2	3	2	3	2	3	3	3
short	1	1	1	NA	1	2	2	1	1	1
sit	1	1	2	1	3	5	5	1	1	4
skin	1	1	1	1	1	2	2	2	2	2

(*continued*)

Appendix Table 2. (*Continued*).

Meaning	NHM	SHM	LM	SaP	OlP	TxP	SoZ	ChZ Maria	ChZ Miguel	ChisZ
sleep	1	1	1	1	1	3	3	3	4	2
small	1	2	2	3	3	4	5	NA	6	3
smoke	1	1	1	2	1	1	1	3	3	1
snake	1	1	1	1	1	1	1	1	1	1
stand (1)	1	1	1	1	1	2	1	NA	1	1
stand (2)	NA	NA	NA	NA	NA	NA	NA	1	1	NA
star	1	1	1	1	1	1	1	1	1	1
stone	1	1	1	1	1	1	1	1	1	1
sun	1	1	1	1	1	2	2	2	2	2
swim	1	3	2	3	3	5	5	3	6	4
tail	1	1	2	1	1	1	1	1	1	1
that	1	6	2	3	2	NA	2	NA	5	4
thin	1	1	1	1	2	4	5	1	6	3
this	1	1	2	1	3	1	4	1	1	1
thou	1	1	1	1	1	1	1	1	1	1
tongue	1	1	1	1	1	NA	1	2	1	1
tooth	1	1	1	1	1	1	1	1	1	1
tree	1	2	1	1	1	1	1	1	1	1
two	1	1	1	1	1	2	2	3	3	1
walk/go (1)	1	1	1	1	2	2	2	NA	2	2
walk/go (2)	NA	NA	NA	NA	NA	NA	NA	1	1	NA
warm	1	1	1	2	3	NA	NA	5	5	4
water	1	1	1	1	1	1	1	1	1	1
we	1	1	1	1	1	1	NA	2	2	3
what	1	1	1	1	1	1	1	1	1	1
white	1	1	1	1	1	1	1	1	1	1
who	1	1	1	1	1	3	4	2	2	2
wind	1	1	1	2	3	4	4	NA	4	4
woman	1	2	2	2	3	4	4	4	4	4
worm (1)	1	2	2	3	4	5	5	5	4	NA
worm (2)	NA	NA	NA	NA	NA	1	NA	NA	NA	1
year	1	1	1	2	2	3	3	3	3	3
yellow	1	1	1	1	1	1	1	1	1	1