WORKING PAPERS

IN

LINGUISTICS

The notes and articles in this series are progress reports on work being carried on by students and faculty in the Department. Because these papers are not finished products, readers are asked not to cite from them without noting their preliminary nature. The authors welcome any comments and suggestions that readers might offer.

Volume 46(5)

December

2015

DEPARTMENT OF LINGUISTICS

UNIVERSITY OF HAWAIʻI AT MĀNOA

HONOLULU 96822

DEPARTMENT OF LINGUISTICS FACULTY

2015

Victoria B. Anderson
Andrea Berez-Kroeker
Derek Bickerton (Emeritus)
Robert A. Blust
Lyle Campbell
Kenneth W. Cook (Adjunct)
Kamil Deen
Patricia J. Donegan (Chair)
Katie K. Drager
Emanuel J. Drechsel (Adjunct)
Michael L. Forman (Emeritus)
Gary Holton
Roderick A. Jacobs (Emeritus)
William O'Grady
Yuko Otsuka
Ann Marie Peters (Emeritus)
Kenneth L. Rehg (Adjunct)
Lawrence A. Reid (Emeritus)
Amy J. Schafer (Acting Graduate Chair)
Albert J. Schütz, (Emeritus, Editor)
Jacob Terrell
James Woodward Jr. (Adjunct)

LANGUAGE VS. DIALECT IN LANGUAGE CATALOGUING: THE VEXED CASE OF OTOMANGUEAN DIALECT CONTINUA[1]

EVE OKURA

This paper reviews the literature on the language-vs.-dialect question and mutual intelligibility testing. It discusses conflicting internal classification schemes from different sources involving Mixtec and Zapotec dialect continua (members of the Otomanguean language family). The paper shows how various catalogues differ in their purposes and thus in their methods of dealing with these dialect continua, comparing and contrasting various internal classification schemes. It also demonstrates how the *Catalogue of Endangered languages* has solved the issue of the Mixtec and Zapotec dialect continua.

KEYWORDS: dialect continuum; dialect continua; dialect chain; mutual intelligibility; catalogue; Otomanguean; Mixtec; Zapotec; Catalogue of Endangered Languages; ELCat

1. INTRODUCTION. Linguistic studies from the well-known earlier ones, such as Hockett 1958 and Haugen 1966, and recent ones, such as Blust 2013, have discussed the linguistic problem of determining whether varieties in dialect continua should be considered separate languages or dialects of a single language. Varying degrees of intelligibility between varieties make it difficult to find a satisfactory solution. There have been disagreements regarding what kind of criteria should be used to determine what makes an entity a distinct language as opposed to a dialect, e.g., linguistic features, genetic relationship, sociocultural attributes, and mutual intelligibility. A number of methods have been developed in attempting to measure mutual intelligibility alone (e.g., Voegelin and Harris 1951; Wolff 1959; Casad 1974; Bradley 1976; Agard 1984). For many linguistic purposes this dilemma can be ignored. However, for languages in a language catalogue, it is necessary determine how to represent dialect continua. Are all varieties to be entered as dialects of a single language? How should one respond to conflicts regarding the number of entities and the names of varieties in a dialect continuum? The *catalogue of endangered languages* (*ELCat*) faced these questions when entering data for Otomanguean dialect continua.

This paper discusses how the question of language versus dialect has been treated in the linguistic literature, how various catalogues have dealt with this issue, and what was done with Otomanguean dialect continua in *ELCat*. It proposes a solution the problem of representing the dialect continua involving Otomanguean languages in language catalogues, especially in *ELCat*.

One might think that defining "a language" would be a straightforward task. It would seem fundamental to the entire field of linguistics, and—to borrow terminology from *The structure of scientific revolutions* (Kuhn 1962)—a primary idea upon which all other secondary, tertiary, etc. ideas in the field would be built. However, how to determine what constitutes a language as opposed to a dialect can be surprisingly difficult, especially in cases of dialect continua. As Robert Blust stated, "It is one of the easiest problems to talk about in linguistics and one of the most difficult to answer" (personal communication, 2014). The difficulty is illustrated by the reaction of Albert J. Schütz, who has extensive experience classifying language varieties of Fiji: "If someone asks you how many languages or dialects are in an area, the best thing to do is turn around and walk in the other direction" (personal communication, 2014), because the asker typically expects a simple answer, but no simple answer exists.

[1] I would like to thank Lyle Campbell for his guidance on this project. I would also like to thank William O'Grady and Robert Blust for their helpful feedback. Gabriela Pérez Báez and Terrence Kaufman provided invaluable and essential information on the Zapotec language complex. Thank you also to *ELCat* team's Russell Barlow for editing the Mixtec and Zapotec data and to Albert J. Schütz for corrections. I would also like to express appreciation to Stephen Cicirelli for his editorial suggestions that helped make this paper more readable.

Thus, while defining language boundaries may be fundamental in linguistics, the proposed answers to the question of language vs. dialect are far from unified.

In many areas of inquiry in linguistics one can respond in just the way suggested by Schütz: when confronted with the question of language versus dialect—turn around and walk in the other direction. However, when it comes to data for "macrolanguages" (defined below) in a catalogue of languages, one is not afforded the luxury of sweeping this question quietly under the rug. This is especially relevant in the case of *ELCat*, as it seeks to have language entries based solely on linguistic criteria (as opposed to having separate entries for dialects, or for defining languages by social, political, or cultural criteria).

The structure of ELCat database requires that its data be entered for a specific language (not for dialects or varieties of languages). In some cases, it is unclear whether a specific linguistic variety is an actual language, or if it is one of several dialects of a language. This uncertainty is a problem even in some of the most current scholarly literature. This kind of uncertainty as to language classification is not as rare as one might think, occurring in a variety of geographic regions and in a number of unrelated language families. These kinds of uncertainties occur, for example, in the cases of the West Romance Dialect Continuum in Europe (Indo-European language family) (Downes 1989:19), Cree and Ojibwe dialect continua in northeastern North America (Algic language family) (Mithun 2001), Austronesian dialect chains in the Pacific (Blust 2013), and many of the 60 linguistic entities *Ethnologue* has labeled "macrolanguages" (Lewis et al. 2013).

Ethnologue defines a "macrolanguage" as "multiple, closely related individual languages that are deemed in some usage contexts to be a single language." The term was coined to deal with language groups that had been given a single ISO 639-2 code (i.e., labeled as a single language in the former ISO code system), but were found to have mutually unintelligible languages within them [2] Thus, "macrolanguage" is—to some degree—a reaction to the ISO 639–2 system, so which language groups happen to be called a "macrolanguage" has an element of randomness to it in that it is not defined only by linguistic or intelligibility criteria; only those language groups that happened to be mislabeled in ISO 639-2 are included. Nor does it respond to the reverse situation—cases in which multiple mutually intelligible varieties have been given different ISO 639–2 or 639–3 codes – that is, where dialects of one language have been considered separate languages. This is because *Ethnologue* does not classify language varieties strictly on linguistic criteria or on mutual intelligibility. Rather, *Ethnologue* classifications are based on ISO 639–3, whose classifications are based on three criteria:

1. mutual intelligibility
2. shared literature or "ethnologuistic" identity, even when levels of mutual intelligibility alone would not be enough to classify varieties as the same language
3. where mutual intelligibility is strong enough to be considered the same language, "well-established distinct ethnolinguistic identities" are considered reason enough to enter the varieties as separate languages.[3]

For example, each variety of Mixtec is labeled with the specific variety name together with "Mixtec," for example "Mixtec, Apasco-Apoala." "Mixtec" then becomes a type of surname for all of those related varieties. However, no attempt is made to define what type of an entity "Mixtec" is as a whole – a single language or a collection of separated languages and dialects.

Ethnologue includes Cree and Ojibwe among those 60 macrolanguages. Zapotec is also listed as one of *Ethnologue*'s macrolanguages. Interestingly, Mixtec is not categorized as a macrolanguage in *Ethnologue*. Rather, each Mixtec variety has a separate entry, with no mention of being part of a macrolanguage.

This paper discusses how the issue of language versus dialect has been treated how various catalogues, with focus on the solution adopted for the Otomanguean dialect continua in *ELCat*.

[2] https://www.ethnologue.com/about/problem-language-identification#MacroLgsID
[3] Ibid.

The Instituto Nacional de Lenguas Indígenas (INALI) recognizes some of the challenges faced in correctly identifying languages based on linguistic criteria, as opposed to listing only by the colloquial names for language varieties, which do not necessarily treat linguistic varieties in a systematic way.

> Es posible identificar algunos parámetros que generan confusión o malas interpretaciones de los datos. Esto es más notorio en los casos en los que más de una lengua se conoce con el mismo nombre, como el chontal (chontal de Oaxaca y chontal de Tabasco). Aun cuando hay una parte del nombre que sirve para distinguir estas dos lenguas, en algunos censos han aparecido referidas estas lenguas tanto con el nombre distintivo como con el nombre que no hace distinción (chontal). Un caso similar es el de los nombres usados como genéricos. Los casos más problemáticos pueden ser, para el INEGI, el chinanteco, el mixteco y el zapoteco. En estos casos aparecen el nombre genérico y los nombres de las variantes. (Ósorio and Alarcón 2009:88)

Thus, according to INALI, "los casos más problemáticos" (the most problematic cases) for distinguishing between separate languages that are all locally referred to by a shared general, generic name, are "el chinanteco, el mixteco y el zapoteco." This paper focuses on two of these three language groups—Mixtec and Zapotec. The sections for each of these language groups will:

1. Report various competing internal classifications from different sources for each language group;
2. Address how these language groups are represented by comprehensive language sources:
   a. *Ethnologue*
   b. *Glottolog*
   c. Insituto Nacional de Lenguas Indígenas (INALI)
   d. The Catalogue of Endangered Languages (ELCat)
3. Search primary sources to determine which internal classification schemes are the most accurate in current literature (i.e., based on linguistic criteria with data obtained by rigorous methods);
4. Acknowledge the uncertainties, challenges, and limitations that remain.

2. LANGUAGE VS. DIALECT. The problem of determining what is a language and what is a dialect is not new. Hockett (1958:8) presented the issue in the introductory pages of his *A course in modern linguistics*: "We cannot always judge whether the speech of two groups should be counted as separate languages or only as divergent dialects of a single language." Hockett addressed dialect continua in a manner quite different from *Ethnologue*'s method. First, although he discussed the role of ethnolinguistic identity, he used terminology to talk about the issue based on linguistic criteria and mutual intelligibility. Hockett's "L-simplex" refers to a clearly defined language (whether by linguistic criteria or by tradition). His L-complex "consists of any idiolect plus all other idiolects which are linked both to the first and (consequently) to each other" (Hockett 1958:324). He discussed various combinations of how these concepts are exhibited in real-world examples. He stated that a language could be both an L-complex and an L-simplex if it has been traditionally called by a single language name, but has been found to consist of some unintelligible varieties (e.g.,Menomini and Chocktaw) (Hockett 1958:324). German, however, is different:

> In other instances the correlation is not so neat. If by "German" we mean what is usually meant by the term—all the idiolects of Germany, German-speaking Austria, and German-speaking Switzerland—then this "language" is more than a single L-simplex but less than an L-complex. There are pairs of idiolects in "German" which are not mutually intelligible: a speaker from Switzerland and one from the boundary region near Holland, neither of whom has learned standard German, cannot understand each other. All the idiolects of German, as defined, belong to a single L-complex, but the latter is more inclusive, since it includes also all the idiolects of Dutch and Flemish (Hockett 1958:324).

Hockett also included Romance languages, varieties of English, and varieties of Mandarin as types of dialect continua that have varieties that are mutually intelligible and some varieties that are mutually

unintelligible. Where varieties located at the beginning and end of the chain are unintelligible, but each of the varieties in between is intelligible to the varieties nearest to it, to varying degrees.

Haugen (1966) continued the discussion. He recounted the origins of the English terms “language” and “dialect,” and the history of their use in the English languages (along with similar terms in French):

> The taxonomy of linguistic description—that is, the identification and enumeration of languages—is greatly hampered by the ambiguities and obscurities attaching to the terms “language” and “dialect.” Laymen naturally assume that these terms, which are both popular and scientific in their use, refer to actual entities that are clearly distinguishable and therefore enumerable. A typical question asked of the linguist is: “How many languages are there in the world?” Or: “How many dialects are there in this country?” (Haugen 1966:922).

Haugen (1966) observed that this issue is relevant for languages in Europe and Africa. While the Romance languages have often been used as an example of a dialect continuum, Blust (2013:34) acknowledges that the question is relevant even for varieties of English (e.g., American English and Australian English). However, the language-dialect problem extends beyond the well-known cases involving Indo-European languages. Blust (2013:34) highlights the situation in the Austronesian language family as well, and points out the problematic nature of some of the methods previously used to make language boundary distinctions.

Various methods have been attempted to define linguistic boundaries in these ambiguous cases. Typically, “mutual intelligibility” is used as the criterion to determine whether two linguistic entities are dialects of the same language or are separate and independent languages; i.e., if speakers of the two varieties in question can understand each other (if mutual intelligibility exists), then the varieties can be classified as dialects of the same language; if speakers of the varieties cannot understand each other, the varieties can be classified as two different languages.

Voegelin and Harris (1951:323) described four methods for determining the difference between a language and a dialect: (1) the “ask the informant” method; (2) the “count sameness” method; (3) the “structural status” method; and (4) the “test the informant” method. Method (1), “ask the informant,” involves simply asking speakers whether or not another person/group/town speaks their same language. There are a number of problems with this approach. First, sometimes cultural differences result in groups reporting that another group’s language is different even if it is actually mutually intelligible. Second, the speaker may be basing the “differentness” of another variety on criteria that is not sufficient from a linguistic perspective to distinguish a dialect from a language (e.g., the speaker may say it is a different variety because it has some differences, but it is not different enough to make it mutually unintelligible). Third, speakers from different language groups could be bilingual or multilingual. This adds ambiguity in answers to the question, “Do people in X town speak the same language as people in this town?” Methods (2) and (3) are more rigorous than method (1) (method 3 being preferable to method 2). Schütz’s study of Fijian dialects exemplifies method (3), in which he determined an estimate of the number of Fijian dialects based on phonological analyses (Schütz 1963). Method (2) does not account for possible borrowings in attempting to determine genetic relationships. Method (3) would be more useful than (2) in reflecting genetic relatedness. Methods (2) and (3) are based on linguistic similarity as opposed to mutual intelligibility (although method 2—lexical “sameness”--could perhaps translate into at least some minor degree of mutual intelligibility). Method (4) addresses mutual intelligibility as opposed to linguistic criteria and genetic relationships. A combination of all four of these methods would be preferable. However, method (4) “testing the speaker” could involve a variety of methods, some more useful than others. Some of these are discussed in the following sections.

Wolff (1959) outlined some of the problems of testing for distance between or among dialects. The first of these is that the Voegelin and Harris test involved translating. Translations can risk altering the original intended meanings and/or risk producing unnatural sounding language. The second problem is that there may be cultural reactions to different ways of speaking and to people who usually speak that way. These cultural reactions could also affect results of intelligibility tests. A third problem is that testers were using audio recordings instead of testing speakers listening to a live person, and the quality of the

audio recording then comes into question. In reaction to these problems, Wolff (1959:36) proposed alternate methods for more effective testing:

> Dialect distance testing is effective under two conditions: (1) that the informant has not, prior to the test, learned the non-native dialect; (2) that the informant is free from any temperamental resistance to translating between his native dialect and one or more dialects with which he is unfamiliar.

Wolff (1959:36) also suggested that mutual intelligibility may not coincide with internal classification schemes.

> Linguistic (phonemic, morphemic, lexical) similarity between two dialects does not seem to guarantee the possibility of interlingual communication; similarly, the existence of interlingual communication is not necessarily an indication of the linguistic similarity between two such dialects."

Wolff also highlighted some of the problems with intelligibility testing. One of these problems was that tests were not as helpful if asymmetrical intelligibility existed. Another problem is high levels of bilingualism. Bilingualism renders the test useless—a researcher would not be able to determine the distance between two varieties if the speaker-consultant knows both varieties. The speaker's understanding would reflect bilingualism rather than intelligibility. In other words, "While ability to translate obviously presupposes some type of intelligibility, the reverse is not necessarily true" (Wolff 1959:34).

Haugen (1966:926) also addressed the inherent difficulties and limitations with attempts to "measure" comprehensibility between speakers of different varieties: "Between total incomprehension and total comprehension there is a large twilight zone of partial comprehension in which something occurs that we may call "semicommunication":

> The very notion of an area divided into a given number dialects, one neatly distinct from the next, had to be abandoned. The idea that languages split like branches on a tree gave way to an entirely different and even incompatible idea, namely, that individual linguistic traits diffused through social space and formed isoglosses that rarely coincided. Instead of a dialect, one had a "Kernlandschaft" with ragged edges, where bundles of isoglosses testified that some slight barrier had been interposed to free communication. Linguistics is still saddled with these irreconcilable "particle" and wave" theories.

Haugen described dialects as a type of "pre-language," a spoken variety of language that has not been standardized and/or put into writing. He concluded that:

> The four aspects of language development … as crucial features in taking the step from "dialect" to "language," from vernacular to standard, are as follows: (1) selection of norm, (2) codification of form, (3) elaboration of function, and (4) acceptance by the community (Haugen 1966:933).

Casad (1974) cited Voegelin and Harris (1951); Olmsted (1954); Hickerson, Turner, and Hickerson (1952); Pierce (1952); Biggs (1957); and Wolff (1964). He discussed their various intelligibility tests, and provided a step-by-step instruction manual explaining how to conduct dialect intelligibility tests. He also gave a brief overview of Bradley's (1976) methods used in testing Mixtec dialect intelligibility, a topic of special interest in this paper. The method for which Casad gave instructions and which Bradley used was the traditional SIL intelligibility test that involves using tape recordings and testing speakers on their comprehension of the recordings. Bradley made some alterations to the method, used for later tests. Previous tests used both sentence and narratives; Bradley used only narratives. He had all subjects listen to the same narratives, eliminating unnecessary variables present in past studies that used different narratives for different groups of participants. In addition, participants only had to answer questions rather than reciting the narrative (Casad 1974:61–62). While cautions were taken to eliminate potentially faulty data (e.g., questions that more than 50% of participants got wrong, and responses from participants who got more than 50% wrong), there were other problems with the test. For example, some of the participants had no formal education, so they had to learn how to take a test for the first time as they responded to questions. This could have resulted in respondents appearing to understand less than they actually did due

to unfamiliarity with the test-taking method. This, in turn, could result in lower intelligibility scores, making language varieties seem more divergent than they actually were.

Agard (1984) proposed that the solution to this question began with returning to the semantics of the terms "language" and "dialect," and more rigorously redefining what these actually mean in the world of linguistics. He said:

> Crucial to our descriptive method are, first and foremost, acceptable working definitions of both language and dialect. And by "acceptable" in this frame of reference is meant, of course, something much more precise and rigorous than the traditional definitions based on extrasystemic, socio-political, or merely behavioral criteria. We need definitions based, rather, on intrasystemic (in a word, structural) criteria. There are those who pronounce it impossible to redefine such age-old notions on any new basis, and would favor coining new terms instead. But one may disagree with this oblique approach and claim that linguists can, and should, continue to avail themselves of the consecrated terms "language" and "dialect," though distinguishing the one from the other in a different way from the layman (Agard 1984:41).

While Agard was concerned with defining "language" and "dialect," Trudgill (1986) was concerned with dialectology's relation to sociolinguistics. He placed dialectology under sociolinguistics. He also gave an overview of macro (geolinguistics) versus micro sociolinguistics (individual face-to-face interactions that, multiplied, result in language movement and change). Trudgill warned of the risk of dialect boundaries based on a single sound change (which problem Josserand avoids by using multiple phonological, lexical, syntactic, and other criteria in creating Mixtec dialect maps; see below), and on using dialect maps to determine language movement.

Thomason and Kaufman (1988) coined the terms "emerging pidgin" (1988:171, 192–93), "emergent creole" (1988:148, 153, 158, 163), and "emerging contact language" (1988:150–51) to describe creoles and pidgins that were separating from each other diachronically, i.e., on the trajectory toward becoming distinct languages. The term "emerging language" has extended beyond the scope of contact languages to include dialects of a language on the edge of diverging into independent languages (Golla 2007:20–22).

Blust (2013:33) also discussed various approaches to determining distinct languages, as well as limitations and questions that arise with each approach:

> Do we group speech communities into languages on the basis of intelligibility, cognate percentages, structural similarity, some combination of these, or in some totally different way? Second, are languages discrete entities with clearly demarcated boundaries, or like many features of nature are they, at least in some cases, natural continua upon which boundaries can be imposed only with an inescapable measure of arbitrariness?

Like Hockett (1958), Blust pointed out that some varieties take several days of exposure to them to become intelligible e.g., American English with Australian English, and with some languages of the Austronesian language family. Blust also stressed how the *kind* of test chosen can affect mutual intelligibility results:

> Quantitative measures of lexical similarity provide another means for determining language boundaries. In practice this has generally been carried out through use of a standard 100-word or 200-word lexicostatistical test-list. The composition of such lists varies slightly from user to user, with special adaptations made for individual language families (thus 'sun' and 'day', while perfectly good for Indo-European languages, are repetitive in AN, since many of these languages express the former by 'eye of the day', and both 'eye' and 'day' appear independently on the list). One of the difficulties that has emerged in using lexicostatistical percentages for locating language boundaries is that different researchers use different cut-off points, with obviously different results. Dyen (1965a) defines speech communities as dialects of the same language if at least 70% of their basic vocabulary is cognate, while Tryon (1976) uses 81% for the same purpose. Many AN languages share a highest cognate percentage with another language that lies somewhere between these figures, and the choice of which one to adopt will therefore have significant consequences for the results of a language count (Blust 2013:34).

The uncertainty about what constitutes a language makes linguistics more and not less like the other sciences—physics, biology, astronomy, etc. These fuzzy boundaries are evident in classification of

objects in astronomy. Until recently, Pluto was considered a planet. However, a team of scientists decided to redefine the term "planet," so that Pluto is now considered a "dwarf planet" or "Plutoid" (NASA 2015)—bearing uncanny similarity to the Max Planck Institute's term "languoid" in *Glottolog*, to be explained in section 3.2.

3. DIALECT CONTINUA IN LANGUAGE CATALOGUES. *ELCat*), *Ethnologue*, *Glottolog* (Hammarström et al. 2015), and *INALI* each take different approaches to the treatment of linguistic varieties in these catalogues. Each also cites different definitions of the types of linguistic varieties included in their respective systems. As mentioned previously, *ELCat* seeks to include independent languages as defined by linguistic criteria.

3.1 *ETHNOLOGUE*. *Ethnologue*'s definition of "language" used for the purposes of its catalogue includes social and cultural criteria:

> Where there is enough intelligibility between varieties to enable communication, the existence of well-established distinct ethnolinguistic identities can be a strong indicator that they should nevertheless be considered to be different languages.
>
> These criteria make it clear that the identification of "a language" is not based on linguistic criteria alone.
>
> The language entries in *Ethnologue* include a listing of dialect names. In most cases, those listings are not based on rigorous dialectology. Rather, these lists include all names reported to us which may, at one time or another, have been used in reference to a local variety of a language. Names listed may be alternate names for the same linguistic variety (Lewis et al. 2013).

This corresponds to the common colloquial definition of what "a language" is. While there is no single correct definition of language, each source uses the definition that suits its purposes. *Ethnologue* also uses the three-letter International Standards Organization (ISO) codes that SIL has assigned to each language. These, too, are based on the definition of a language based on ethnolinguistic criteria, and not solely on linguistic criteria. This definition serves a useful social purpose. In addition, *Ethnologue* often presents a calculation of the percent of mutual intelligibility between language varieties. However, sometimes varieties that are over 90% mutually intelligible to each other are given entries under separate language names. *Ethnologue* keeps thorough track of each of these degrees of mutual intelligibility (based on SIL studies) where the information is available. This directory of languages also clearly refers to some entities as "dialects" as opposed to languages, although the two varieties may be given separate entries. Given its purposes and the criteria it uses (just outlined), it is important to note that the number of entries for a language group in *Ethnologue* does not reflect the number of actual languages in the group as determined by linguistic criteria, but reflects the number of known/documented dialects.

3.2 *GLOTTOLOG*. The Max Planck Institute's *Glottolog* is perhaps the most comprehensive bibliography of each of the world's languages, and resolves the "language vs. dialect" problem with an innovative solution of its own, which basically just ducks the issue. It says:

> Glottolog provides a comprehensive catalogue of the world's languages, language families and dialects. It assigns a unique and stable identifier (the Glottocode) to (in principle) all languoids, i.e. all families, languages, and dialects. Any variety that a linguist works on should eventually get its own entry. The languoids are organized via a genealogical classification (the Glottolog tree) that is based on available historical-comparative research (see also the Languoids information section). (Glottolog 2014)

Thus, *Glottolog* just includes "any variety," and makes no distinction between dialects and separate languages. *Glottolog* uses SIL's ISO language codes for languages that have corresponding entries.

3.3 *INALI*. The *Instituto Nacional de Lenguas Indígenas* (*INALI*) addresses the issues of language classification and nomenclature by using two different kinds of classification: (1) *autodenominaciones* (self-designations) and (2) *agrupaciones lingüísticas* (linguistic groupings). The *autodenominaciones* are names that speakers of a language variety use to refer to themselves. These distinctions sometimes coincide with linguistic distinctions, but often they do not, as they are based on social, cultural, and

political criteria. The *agrupaciones lingüísticas* are based on linguistic criteria. However, in a tactic similar to *Glottolog's* "languoids," even within the *agrupaciones lingüísticas*, INALI uses sociolinguistic criteria, referring to entities as "*variantes*" (variants, varieties), thereby circumventing the issue of having to classify linguistic entities as either languages or dialects:

> La **VARIANTE LINGÜÍSTICA** es la categoría que alcanza el mayor grado de detalle de los niveles de catalogación aplicados en este trabajo y se define como una forma de habla que: a) presenta diferencias estructurales y léxicas en comparación con otras variantes de la misma agrupación lingüística; y b) implica para sus usuarios una identidad sociolingüística que contrasta con la identidad sociolingüística de los usuarios de otras variantes (INALI 2008:37).

INALI defines its "variante lingüística" as partially determined by sociolinguistic identity. The term "variante" (variant) works well for its purposes. However, it becomes difficult to determine internal classifications using only the distinction of "variante." Whereas SIL uses town names to for dialect names in *Ethnologue*, INALI avoids using town names, and instead coined original names for these linguistic varieties based on wider geographic areas (e.g., "chinanteco del norte" [Northern Chinantec]). They did this intentionally for cases in which a linguistic variety is spoken in more than one town, with the idea that speakers who live in other towns may feel as though they are not being represented if their language is called by the name of another town.

3.4 ORGANIZING CLASSIFICATIONS FOR *ELCAT*. In order to determine how many and which Mixtec and Zapotec languages to include in *ELCat*, the challenge was to figure out which *Ethnologue* dialects, *Glottolog* "languoids," and *INALI* "variantes" correspond to each other, or how they correspond, if at all. In addition to each source having a different number of divisions for both of the language groups in question (Mixtec and Zapotec), they each also often use different names to refer to what appear to be the same linguistic varieties. *INALI* uses unique names (not used by any other source), and does not use ISO language codes (which sometimes to do not correlate to languages, but rather to dialects, or even more to sociocultural/political entities, instead of keeping with strictly linguistic criteria). This makes it necessary to determine which languages in other sources correspond to *INALI* entities. However, the assumption that there are unambiguous correspondences in other sources, just called by some other name, is wrong, as it is possible that what *INALI* refers to as "chinanteco del norte" does not have any clear correspondent in *Ethnologue*, but rather is composed of parts of several *Ethnologue* varieties. Glottolog and *Ethnologue* overlap somewhat in nomenclature, and Glottolog does use SIL's ISO codes where applicable, but Glottolog also includes many languoid names not used by *Ethnologue* and that do not have ISO codes. In other cases, where *Ethnologue* lists a greater number of entities than Glottolog, the reverse is true.

The primary method used here to determine which linguistic entities from each source correspond to those listed in other sources was to look at the lists of towns where each *INALI variante* is spoken, and to compare those to the towns associated with that a particular language named for a certain name in *Ethnologue*. A clear one-to-one correlation would be ideal, but this is rarely the case. Where *INALI* lists more than one town for a *variante*, the single variety corresponds to those several dialects in *Ethnologue* called by those town names. Where several different *INALI* varieties are spoken in the same town, these several variantes are listed as a single *Ethnologue* dialect. Admittedly, this is an imperfect technique with obvious limitations. For example, it is possible that what *Ethnologue* calls by a town name may actually be referring to only one linguistic dialect, and would not be relevant to other INALI variantes that happen also to be spoken in that same town. For greater accuracy, it will eventually be necessary to determine what exactly *INALI* and *SIL* mean by each of the linguistic varieties they name, and how exactly these distinctions were determined. Until then, this present study addresses only information available in published sources supplemented by consultations through personal communication (most particularly with Terrence Kaufman and Lyle Campbell).

For a classification of the full Otomanguean language family, see Campbell 1997:158 and Kaufman 1990. See figure 1 for the distribution of the Otomanguean languages.

4. OTOMANGUEAN DIALECT CONTINUA IN *ELCAT*.

FIGURE 1. Otomanguean languages. (*Reproduced from INALI 2009*)

4.1 MIXTEC. The disparities between sources are much more drastic in the Mixtec situation than with some other dialect continua. Whereas *Glottolog*, *Ethnologue*, and *INALI* differ only by one to two in their total number of varieties within some language complexes, (e.g., Chinantec—14, 13, and 11, respectively), for Mixtec the range of difference is much greater: *Ethnologue* has 52 dialects, *Glottolog* cites 71 languoids, and *INALI* lists 81 varientes. Terrence Kaufman (personal communication, March 29, 2014) recommended Josserand's 1983 dissertation, *Mixtec Dialect History* as the most accurate linguistic sources for Mixtec internal classifications, and *ELCat* has only 12 Mixtec languages, based largely on Josseerand 1983, who has c.12–22 Mixtec entities. These dramatic differences suggest the need for more research here.

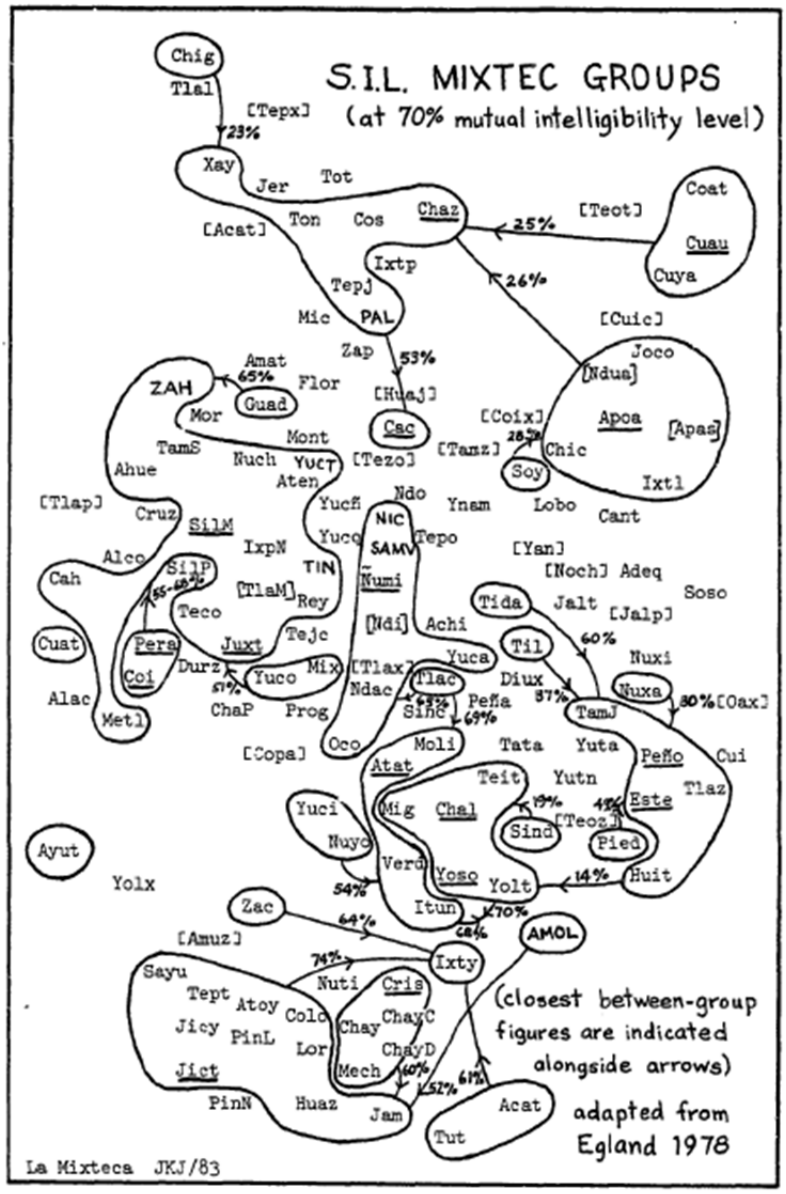
Glottolog lists 71 languoids in 12 subgroups. It should be recalled that *Glottolog* differs; it does not pretend to present a classification of languages, but rather includes every language, dialect, or variety ever cited in the literature. It even includes forms that were mentioned, but have been proven incorrect. It labels them as "spurious," but still includes them. Therefore, the inclusion of 71 Mixtec languoids in *Glottolog* is not a claim that there are 71 Mixtec languages. Its number of 71 comes from totaling the different independently listed linguistic entities listed in Josserand 1983.

While Josserand did extensive, rigorous linguistic work on Mixtec varieties, with a thorough analysis of sound correspondences, she did not declare an exact number of Mixtec languages with precise language names. Instead, she and most scholars who work on Mixtec, avoid drawing clear lines between dialects and languages in the Mixtec complex—hence why INALI referred to Mixtec as one of the most problematic nomenclature/classification issues in Mexico. Her total number of circled areas was 29. However, Josserand did adapt a 1978 Egland map in her dissertation, in which she circled together dialect groups that share 70% mutual intelligibility or higher (figure 2).

I determined the language names represented by the abbreviations on the map from the legend in Josserand's (1983) appendix, and converted the map data in the following list of Bradley and Josserand Mixtec varieties, adapted from Egland 1978, based on 70% mutual intelligibility (SIL mutual intelligibility tests).

I then compared the above list derived from Bradley's map (adapted by Josserand) and from Josserand's dissertation appendix with SIL's Mixtec dialect names in *Ethnologue* and its ISO codes, to try to determine how Josserand's 29 varieties and resultant internal classification related to existing names and divisions for Mixtec varieties. It was not clearly evident how much the map had been "adapted," and what exactly that adaptation entailed.

FIGURE 2. Josserand's map of the Mixtec dialect area (reproduced from Josserand 1983:148)



In addition to the 29 Mixtec "languages" derived from the 1978 map, Josserand also referred to a study by Spores (as reported by Josserand) of the Mixtec area, and subgroupings he created based on data from multiple disciplines:

> Spores did, however, discuss with me in 1978 a number of towns which might fall into distinct dialect areas, mostly within the Mixteca Alta and central Mixteca Baja, which I present here as an indication of the level of specificity possible with a more precise knowledge of Mixtec communities and social organization. These groupings reflect Spores' integration of archaeological, ethnohistorical and modern information in his delimitation of interaction spheres within the Mixteca. Comparison of these groups with those defined

> by linguistic criteria shows a marked coincidence between Spores' groupings and my own: (Josserand 1983:128–29)

The 18 subgroups Spores came up with, which Josserand said showed "a marked coincidence" with her own subgroupings based purely on linguistic data, are:

Spores's 18 subgroups of the Mixtec area (based on archaeological, ethnohistorical, etc. data)

1. Apoala, Apasco, Sosola and the eastern frontier with Chinantec, Cuicatec and Zapotec;
2. Coixtlahuaca, Huautla and Tequixixtepec;
3. Tonalá, Chila, Petlalcingo, Mariscala, Acatlán and the towns on the northern frontier with Nahuatl and Tlapanec;
4. Huajuapan;
5. Silacayoapan and its ranchos on the Guerrero border;
6. Tecomaxtlahuaca and Juxtlahuaca;
7. Tlaxiaco and its ranchos, Cuquíla, Ñumí, and Mixtepec;
8. Teposcolula, its ranchos, and another ten communities which use the Teposcolula market, including Tayata and Achíutla;
9. Tilantongo and its ranchos, and Mitlatongo;
10. Chalcatongo and San Miguel el Grande;
11. Yucuañe and nine or ten surrounding communities;
12. Teozacoalco and Peñoles;
13. Putla;
14. Zacatepec;
15. Tututepec and Jamiltepec;
16. Yolotepec;
17. Yanhuitlán, Chicahua, Soyaltepec, Cántaros, Coyotepec, Nochixtlán and Tonaltepec;
18. Tamazulapan, Tejutla, Teotongo, and Chilapa de Díaz.

The same method was used to correlate the 29 languages to *Ethnologue*'s Mixtec languages and ISO codes with Josserand's, I checked each town mentioned in Spores's list of 18 languages to see if it showed up anywhere in *Ethnologue*'s 52 varieties. Where any common town names were listed both in Josserand and *Ethnologue*, those *Ethnologue* varieties are listed as corresponding entries.

For several of Spores's varieties I was unable to find corresponding *Ethnologue* entries. They may exist, but if they do, the correspondences were unclear, perhaps due to the use of different names for the varieties. There are 11 of Spores's varieties, corresponding to 17 *Ethnologue* varieties, that do not fit. This leaves seven of Spores' languages and 35 of *Ethnologue* varieties as yet unaccounted for. Josserand stated that:

> A preliminary grouping of Mixtec communities into dialect areas, using Bradley and Josserand's phonological criteria (1978, 1982) appears in Josserand, Jansen and Romero, 1978 and In press //see if you can find a reference to the published version; if you do, add it here in parentheses, e.g., "and In press [1987]" and add the reference to your list of references//), but **since that study was written, much new material has been analyzed, which must be taken into account in any discussion of dialect areas**. (Josserand 1983:462) (Bold added for emphasis)

The new material that Josserand mentioned included lexical, phonological, and syntactic analyses of Mixtec varieties. Based on these data, she developed a much sounder, linguistically based internal classification of Mixtec varieties. However, because of to the nature of dialect continua, she—like other scholars—did not attempt to distinguish precise names of distinct Mixtec languages, or to give an exact number. However, she did clarify that:

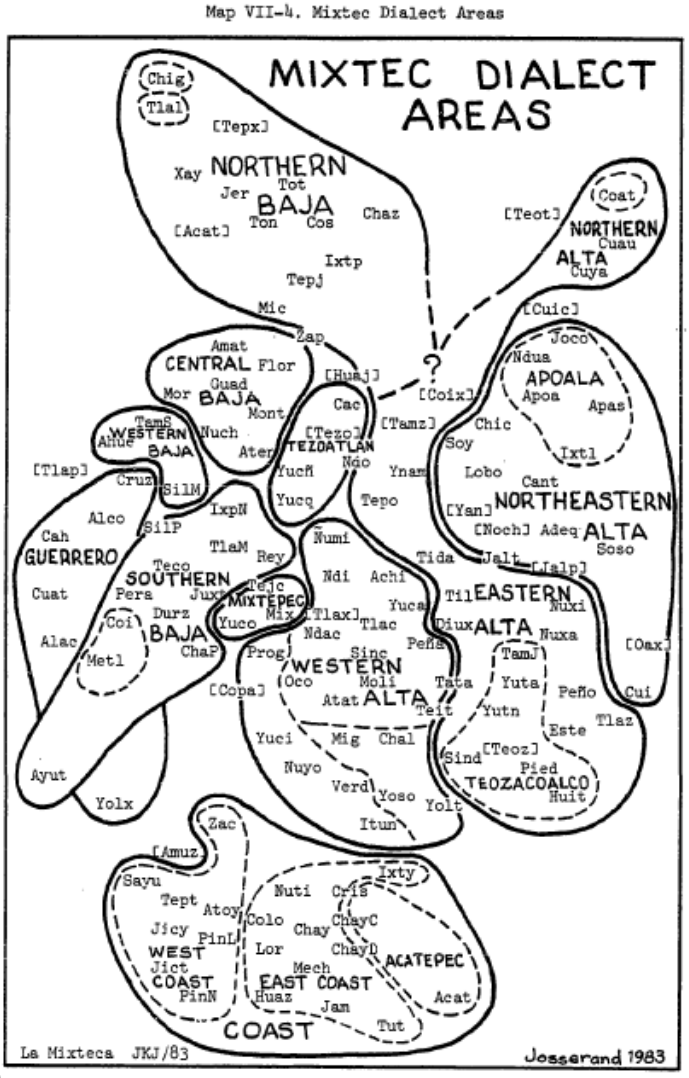
> By any measure, modern varieties of Mixtec must include **over a dozen mutually unintelligible varieties, or languages, each with many local dialects**. In my own sample of 120-Mixtec speaking

villages, *the speech of any one village never coincided in all its* details *with that of any othe*r. (Josserand 1983:457–58; bold and italics added for emphasis).

Based on this statement, there are approximately “over a dozen” actual Mixtec languages. There are a greater number of local dialects, and each of the 120 towns she surveyed has slight variations differing from other towns’ varieties. A challenge has been to determine how many Mixtec languages should be included in ELCat, and which ones they are.

Josserand’s 1983 map of the Mixtec dialect area was used as the basis for the language divisions. This map (see figure 2) differs from previous work in that Josserand applied the new phonological, lexical, etc. linguistic findings to determine subgroupings, as opposed to the Egland map (figure 2), which determines

FIGURE 3. Josserand’s 1983 Mixtec Dialect Area map. (Josserand 1983:470)



subgroupings purely on SIL mutual intelligibility tests (some of which consisted of 100-word Swadesh lists). Although the boundaries in the 1983 map (figure 3) seem less concise than would be ideal for language classifications, this was followed for Mixtec classifications in ELCat, because the dissertation contained much detailed linguistic analysis, though it did not have a list or chart of what it considered the distinct Mixtec languages to be.

This map contains 12 distinct language area boundaries. It also contains dotted-line boundaries within some of those 12 language areas, whose meaning is not clear. It is presumed that they refer to varieties

distinct from each other, but which share certain linguistic features which place them in a subgroup together (the solid line within which they are contained). Since Josserand said there were "over a dozen" distinct Mixtec languages, I think that these dotted-line divisions may indicate boundaries of distinct languages within subgroups, as the solid lines only indicate the 12 divisions (which is not "over a dozen," but exactly a dozen).[4] The 12 clearly defined, solid-line linguistic boundaries in the map are:

1. Northern Baja
2. Central Baja
3. Western Baja
4. Southern Baja
5. Guerrero
6. Tezoatlan
7. Mixtepec
8. Northern Alta
9. Northeastern Alta
10. Eastern Alta
11. Western Alta
12. Coast

If the dotted-line divisions also turned out to be distinct, mutually unintelligible languages within the 12 areas, and those 12 were subgroups of a language family rather than specific languages, the number of Mixtec languages would increase to 22:

1. Northern Baja
2. Chigmecatitlán (Santa María Chigmecatitlán, N Baja 22)
3. Santa Catarina Tlaltempan (Santa Catarina, N Baja 21)
4. Central Baja
5. Western Baja
6. Southern Baja
7. Coicoyán-Metlatonóc
8. Tezoatlan
9. Mixtepec
10. Guerrero
11. Northern Alta
12. Coatzospan (San Juan Coatzospan, NE Alta 27)
13. Northeastern Alta
14. Apoala
15. Eastern Alta
16. Teozacoalco
17. Western Alta Group I
18. Western Alta Group II
19. Western Alta Group III
20. West Coast
21. East Coast
22. Acatepec

[4] Credit goes to Russell Barlow for editing and correcting the Mixtec data.

The dotted-line divisions in Western Alta (17, 18, and 19) were unnamed in the map, so I refer to them as **Western Alta Group I** (northernmost subgroup within Western Alta, consisting of: *San Juan Ñumí, C Alta 1; Santiago Nundichi, C Alta 10; San Miguel Achiutla, C Alta 2; San Bartolomé Yucuañe, C Alta 3; San Mateo Peñasco, C Alta 4; San Agustín Tlacotepec, C Alta 9; Santa Cruz Nundaco, C Alta 11; San José Sinicahua, C Alta 8; Santa María Tataltepec, C Alta 5; San Juan Teita, C Alta 6; San Pedro Molinos, C Alta 7; Santo Tomás Ocotepec, C Alta 12; San Esteban Atatlahuca, C Alta 16*), **Western Alta Group II** (the southeastern group within Western Alta, consisting of: *San Miguel el Grande, C Alta 17; Chalcatongo de Hidalgo, C Alta 18; Santiago Yosondua, C Alta 20; Santa María Yolotepec, C Alta 22*), and **Western Alta Group III** (the westernmost group, consisting of: *San Miguel Progreso, C Alta 13; Santa María Yucuhiti, C Alta 14; Santiago Nuyoo, C Alta 15; Santa Lucía Monteverde, C Alta 19; Santa Cruz Itundujia, C Alta 21*). The map also contains a dotted line between Northern Baja and Northern Alta, possibly suggesting that these two groups share more linguistic features with each other than they do with other Mixtec varieties.

Ethnologue's 52 Mixtec varieties and INALI's 81 Mixtec "variantes" were matched to Josserand's 12 language areas based on town names when listed in those catalogues, and town locations, where town names did not match to any in Josserand's data. Because Josserand's original intended meaning of the dotted lines is uncertain, *ELCat* team decided to enter the 12 larger Mixtec varieties.

4.2 ZAPOTEC. Campbell (1997:158) has described the Zapotec complex as "a number of mutually unintelligible languages estimated to number between 6 and 55 distinct languages." Zapotec has had some of the most recent scholarly attention, and yet its internal classification also remains to be resolved. *Ethnologue* cites 57 Zapotec dialects (Lewis et al. 2013). *Glottolog* lists 65 Zapotec "languoids" (Hammarström et al. 2014), whereas *INALI* catalogues 62 "variantes." *Ethnologue*, *Glottolog*, and *INALI* further subdivide Zapotec into numerous dialect varieties, even exceeding Campbell's range of 6 to 55 "distinct *languages*." *Ethnologue*, *Glottolog*, and *INALI* each acknowledge that the entities the catalogue lists may include cases of mutually intelligible dialects, not only distinct languages.

Within *Glottolog's* list of 65 Zapotec languoids, some are considered varieties of the same dialect, while others are independent languages. This number (65) is more than three times the actual number of distinct Zapotec languages according to the most accurate estimates based on strictly linguistic criteria (Kaufman 2014). *INALI* also numbers Zapotec varieties into the sixties.

Gabriela Pérez-Báez (2011:945) outlines the Zapotec situation in an article about the Juchiteco variety:

> Juchiteco belongs to the Zapotecan branch of the Otomanguean stock of Mesoamerican languages. The dialectal complexity of Zapotec languages has been observed dating back to Fray Juan de Córdova's Arte en lengva zapoteca (Cordoua, [1578] 1987) and a number of proposals have been made for the classification of Zapotec languages throughout the history of their documentation (a detailed overview of this is in Smith-Stark, 2003). The Summer Institute of Linguistics proposes a flatter classification, with Zapotec as a macrolanguage that includes 57 member languages (Lewis, 2009). Similarly, Mexico's Instituto Nacional de Lenguas Indígenas identifies 62 language variants in the Zapotec language group. Kaufman (personal communication) considers Zapotec languages to constitute a language complex defined as a set of closely related languages. He proposes five language areas or virtual languages within the family of Zapotec languages: northern, central, southern, eastern and western. Following Kaufman's classification, Juchiteco is considered to be a variety of central Zapotec. The interaction between speakers of Zapotec languages who participated in a dialectal survey conducted by the Project for the Documentation of the Languages of Mesoamerica (PDLMA) under Terrence Kaufman's supervision and coordinated by Mark Sicoli between 2007 and 2009 supports at least some twenty distinct languages as sub-branches of the aforementioned five virtual languages (Sicoli, personal communication).

As Pérez-Báez stated, there are five Zapotec language areas. Kaufman's notion of 'language area' includes dialects of a particular language and emerging languages, varieties that may or may not be mutually unintelligible and therefore may be distinct languages or may be dialects of a particular language. According to Terrence Kaufman's most recent research, within those five language areas, there are approximately 18 independent (mutually unintelligibly) Zapotec languages (personal communication,

2012). Kaufman gave the following names to the five language areas, and the distribution of the 18 languages among them:

Zapotec language areas and number of distinct languages in each area (Kaufman 2012)

1. Northern (approximately five languages)
2. Central (approximately seven languages)
3. Southern (approximately four languages)
4. Western (one language)
5. Papabuco (one language)

Total: approximately 18 languages

In review, Kaufman stated that the approximately 18 Zapotec languages would fall into the five language areas in the following manner: Northern – 5; Central – 7; Southern – 4; Western – 1; Papabuco – 1. In the data, only four entities are listed at the emergent language tier under the Northern area: (1) Ixtlán; (2) Villalta; (3) Choapan; and (4) Rincón. The question, then, is, "What would the fifth Northern Zapotec language be?" One possibility would be to split Ixtlán into Southeast Ixtlán and West Ixtlán, as those two varieties have the next lowest mutual intelligibility of all of the lower tier Northern Zapotec varieties. Splitting Ixtlán would account for the estimated five Northern Zapotec languages. However, since this was uncertain, *ELCat* lists only the four most certain of the approximately five Northern Zapotec languages.

The Central area appears at first to be the least problematic, as the number of Central Zapotec languages given by Kaufman (2014) should match the number of Central Zapotec languages in the actual 1989 data. These total seven; they are: (1) North Central Zimatlán: Asunción Mixtepec; (2) West Eujtla: Ayoquesco, Nexila (Santa Cruz), Zabache; (3) Ocotlán; (4) Tlacolula; (5) Albarradas; (6) Oaxaca; and (7) Etla: Mazaltepec (Santo Tomás), Tejalapan.

The total number of Zapotec languages in *ELCat* database (19) is based primarily on Kaufman's estimate of approximately 18 Zapotec languages. Some of the specific languages in *ELCat* also conform to specific entities Kaufman identifies as Zapotec languages. Attempts were made to include multiple ISO codes for some language entries, to account for the discrepancies between ISO code dialect varieties and actual independent languages. However, because there were automatic additions to ELCat database initially, and because the process of entering data into *ELCat* requires an SIL ISO code (which may or may not correlate to an independent language), some of the languages entered in *ELCat* may need to be revised.

Zapotec Languages suggested for ELCat[5]

1. Chopan Zapotec [zpc]
2. Ixtlán Zapotec [zaa, zpd, zae]
3. Rincón Zapotec [zar, zsr]
4. Villalta Zapotec [zad, zav, zpu, zpq, ztc, zat]
5. Albarradas Zapotec [zas]
6. Asunción Zapotec [zoo]
7. Ayoquesco Zapotec [zaf]
8. Etla Zapotec [zpy]
9. Ocotlán Zapotec [zac, zpv, zpn]
10. Tlacolula Zapotec [zab, ztt, ztj, zaw, zaq, zpf]
11. Zaachila Zapotec [ztx]
12. Isthmus Zapotec [zai]
13. Petapa Zapotec [zpe]
14. Tlacolulita Zapotec [zpk]

[5] This final list was a result of collaboration between the author and Russell Barlow.

15. Yautepec Zapotec [zpb]
16. Coatlán Zapotec [zps, zpt, ztp, zao, zam, zpr, zap, ztg, ztl, zpo, zpb]
17. Mixtepec Zapotec [zpm]
18. Zimatlán Zapotec [zph, zpp, zpl]
19. Papabuco Zapotec [zte, zpw, zpz]

Zaachila Zapotec (11) is spoken in Oaxaca, so it corresponds to Oaxaca Zapotec. Some varieties have been deleted from ELCat database, and others have been added, but it is likely more work is needed to resolve to some of the Zapotec dilemmas outlined above. Notwithstanding the challenges and the approximations, ELCat is the catalogue closest to representing the actual number of Zapotec languages, as opposed to including all dialect varieties.

5. CONCLUSION. A rigorous process was used to determine (1) which languages in Otomanguean dialect continua exist to begin with (what the language divisions are); (2) how to enter the decided-upon language entities into *ELCat*; (3) how to make them correspond to conflicting classification schemes; and (4) how to enter that data to make it as user-friendly and searchable as possible. In sections IV, V, and VI, the overall framework of the process may have been obscured somewhat by the volume of specific language variety details. The following is a list of the general steps taken, with most of the language-specific details were cropped out. This process could be applied to other geographic and linguistic areas with ambiguous dialect-language boundaries.

1. Determine on which criteria the catalogue will base its entries (Dialects? Languages only? Sociocultural criteria of languages—e.g., tribe names? Strictly linguistic criteria? Intelligibility?)
2. Research and determine the major sources with conflicting internal classifications (e.g., in number and names of languages) of the dialect continuum in question.
3. Determine what the goal was for each of the conflicting sources (e.g., was it to list all dialects, like *Ethnologue*? Was it to list all varieties based on cultural and social identity, like *INALI*?). If intelligibility tests were used, try to determine what those tests entailed, and how valid they appear to be (e.g., a 100-word Swadesh list would not be a very reliable mutual intelligibility test upon which to base an internal classification scheme).
4. Select the classification scheme that is the best for the catalogue's criteria—in the case of *ELCat*, preferred sources were determined by those classifications based on the most accurate linguistic criteria as opposed to cultural/political nomenclature.
5. Create a spreadsheet or a table that shows which dialects (by ISO code) align with which divisions in the chosen classification scheme (i.e., which dialects/varieties group together—have enough mutual intelligibility—to roughly form a "language").
6. Use geographic clues to assist in determining which dialects/varieties (i.e., ISO codes) cluster together for one language entry. For example, in each of the Otomanguean cases in this paper—Mixtec and Zapotec—dialect or variety names are often based on the town where the variety is spoken. *INALI's* published data lists all towns and locations in which each linguistic *variedad* is spoken. I sorted through each of these locations, and when I found the name of a town that matched a variety in *Ethnologue* or some other source, I aligned that variety in that other source with this variety in *INALI*. It is not a one-to-one correspondence. Often, a single variety may be spoken in many areas. Likewise, speakers of differing varieties may reside in a single town.

(*INALI* also points out that when people from several different towns speak a particular variety, it may be insulting to dwellers in other towns when their linguistic variety is given only the name of one of the towns in which the language/dialect is spoken. This eclipses the speakers of that variety in other areas, and it may unintentionally give arbitrary preeminence to the town chosen as the language's name).

7. Enter the name of that language into the catalogue. List all of the ISO codes associated with that language in its entry. This allows catalogue-users to search for languages by other known names and varieties.
8. In the comments section of the language entry, make a note of the range of opinions of the number of varieties of that language complex (e.g., Mixtec: *Ethnologue* has 52 Mixtec dialects; *INALI* lists 81 "variedades"). This clarifies to users that the discrepancy in the number of languages was not an oversight, but a well thought-out, thoroughly researched, deliberate decision.
9. List the varieties associated with that language. Also include a brief explanation of the classification scheme chosen for the current catalogue, and the nomenclature used.
10. While this is not ideal, if no language names exist for the classification scheme chosen (e.g., some of Josserand's subgroups did not have names, but were divided into subgroups). In such case, one might have to invent a useful, accurate, and descriptive name—preferably one based on prior classification schemes. Keep the nomenclature consistent across groupings. If this is done, be sure to make a note of the coined name, and which known names correspond to it (if there are several dialects that combine to form this previously unnamed linguistic entity). If there is a single variety that corresponds to it, it would be best to use the name already in use in the literature.

Ideally, what needs to be done in the future is to develop a new, scientifically validated mutual intelligibility test reliable in areas with dialect continua, and to create new maps of language areas with explanations of the specific linguistic criteria upon which intelligibility percentages/cut off points are determined (for language vs. dialect distinctions).

In the future, yet another possible step toward refining the accuracy of linguistic data of an endangered languages catalogue could be to conduct mutual intelligibility tests to measure the differences between revitalized versions of a language and natively spoken versions of the language (e.g., Hawaiian, Māori, Ojibwa, etc.). Even in this case (as in step 1 in the process outlined previously), the question remains whether to use mutual intelligibility or other linguistic criteria, as the two are not equivalent. To return to Wolff's (1959:35) insight: "Since rather more seems involved than comparability of linguistic units, such as phonemes and morphemes, the question naturally arises what the true significance of intelligibility is." Josserand's Mixtec study was founded initially on Bradley's intelligibility tests, and upgraded based on Josserand's linguistic data—a combination of the two.

In the cases of Mixtec, Zapotec, and even "Pluto," the scientific categories we impose on the continua of the natural world may be mere reflections of human attempts to imagine order in chaos. Perhaps the more accurately we describe what is, the more we will end up with Plutoid and languoid-like categories.

REFERENCES

AGARD, FREDERICK. 1984. *A course in Romance linguistics*. Georgetown: Georgetown University Press.

BIGGS, BRUCE. 1957. Testing intelligibility among Yuman languages. *International Journal of American Linguistics* 23:57–62.

BLUST, ROBERT. 2013. *The Austronesian languages*. Canberra: Australian National University, Asia Pacific Linguistics.

BRADLEY, HENRY. 1976. *Dialect intelligibility testing: The Mixtec case*. (manuscrito inédito).

CAMPBELL, LYLE. 1997. *American Indian languages: The historical linguistics of Native America*. New York: Oxford University Press.

CASAD, EUGENE. 1974. *Dialect intelligibility testing*. *Publications in linguistics and related fields 38*. Oklahoma: Summer Institute of Linguistics of the University of Oklahoma at Norman.

Catalogue of endangered languages (*ELCat*). 2014. The University of Hawai‘i at Mānoa and Eastern Michigan University. http://www.endangeredlanguages.com

DOWNES, WILLIAM. 1998. *Language and society*, 2nd ed. Cambridge, UK: Cambridge University Press.

GOLLA, VICTOR. 2007. North America. In *Encyclopedia of the world's endangered languages*, ed. by Christopher Moseley, 20–22. New York: Routledge.

HAMMARSTRÖM, HARALD; ROBERT FORKEL; MARTIN HASPELMATH; and SEBASTIAN BANK. 2015. *Glottolog* 2.6. Jena: Max Planck Institute for the Science of Human History. (Available online at http://gottolog.org, Accessed on 2014-03-11.)

HAUGEN, EINAR. 1966. Dialect, language, nation. *American Anthropologist*, New Series, 68(4):922–35. Stable URL: http://www.jstor.org/stable/670407 Accessed: 29/01/2015.

HICKERSON, HAROLD; GLEN D. TURNER; and NANCY P. HICKERSON. 1952. Testing procedures for estimating transfer of information among Iroquois dialects and languages. *International Journal of American Linguistics* 18:1–8.

HOCKETT, CHARLES F. 1958. *A course in modern linguistics*. New York: The Macmillan Company.

JOSSERAND, JUDY KATHRYN. 1983. *Mixtec dialect history*. Tulane University PhD dissertation.

KAUFMAN, TERRENCE. 1990. Tlapaneco-Subtiaba, OtoMangue, and Hoka: Where Greenberg went wrong. In *Language and prehistory in the Americas,* ed. by Allan Taylor. Stanford: Stanford University Press.

KUHN, THOMAS. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

LEWIS, M. PAUL; GARY F. SIMONS; and CHARLES D. FENNIG (eds.). 2013. *Ethnologue: Languages of the world*, eighteenth edition. Dallas, Texas: SIL International. Online version: www.Ethnologue.com. Accessed: 3/19/14.

MITHUN, MARIANNE. 2001. *The languages of native North America*. Cambridge, UK: Cambridge University Press.

NASA. 2015. What is Pluto? (Aug. 4, 2015) http://www.nasa.gov/audience/forstudents/k-4/stories/nasa-knows/what-is-pluto-k4.html accessed 11/3/15

OSORIO, ARNULFO EMBRIZ, and ÓSCAR ZAMORA ALARCÓN (Coordinadores). 2009. *Catálogo de las lenguas indígenas nacionales: variantes linguísticas de México con sus autodenominaciones y referencias geoestadísticas,* ed. by LÓPEZ SÁNCHEZ, JAVIERÑ ALEJANDRA ARELLANO MARTÍNEZ, ALMANDINA CÁRDENAS DEMAY, and MARISOL FLORES CASTRO. Instituto Nacional de Lenguas Indígenas (INALI). Lunes 14 de enero de 2008. Diario Oficial (Primera Sección). 3.1.3. Catálogo. 37, 69–79, 84–96. (Tercera sección). 1–3.

PÉREZ BÁEZ, GABRIELA. 2011. Spatial frames of reference preferences in Juchitán Zapotec. *Language Sciences* 33:943–60.

PIERCE, JOE. 1952. Dialect distance testing in Algonquian. *International Journal of American Linguistics* 18:203–10.

SCHÜTZ, ALBERT J. 1963. A phonemic typology of Fijian dialects. *Oceanic Linguistics* 2(2):62–79.

THOMASON, SARAH GREY, and TERRENCE KAUFMAN. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

TRUDGILL, PETER. 1986. *Dialects in contact*. New York: Basil Blackwell.

VOEGELIN, CHARLES F., and ZELLIG S. HARRIS. 1951. Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society*, (Jun. 12, 1951), 95(3):322–29.

WOLFF, HANS. 1959. Intelligibility and inter-ethnic attitudes. *Anthropological Linguistics*. Urbanization and Standard Language: A Symposium Presented at the 1958 Meetings of the American Anthropological Association (Mar., 1959), 1(3):34-41. Published by: The Trustees of Indiana University of behalf of Anthropological Linguistics Stable URL: http://www.jstor.org/stable/30022192 accessed: 20/01/2015

WOLFF, HANS. 1964. Intelligibility and inter-ethnic attitudes. In *Language and culture in society: A reader in linguistics and anthropology,* ed. by Dell Hymes, 440–45. New York: Harper & Row.

eveokura@hawaii.edu