LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 7 (2013), pp. 331-359
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/4606

Languoid, Doculect and Glossonym: Formalizing the Notion 'Language'

Michael Cysouw
Philipps Universität Marburg

Jeff Good
University at Buffalo

It is perfectly reasonable for laypeople and non-linguistic scholars to use names for languages without reflecting on the proper definition of the objects referred to by these names. Simply using a name like *English* or *Witotoan* suffices as an informal communicative designation for a particular language or a language group. However, for the linguistics community, which is by definition occupied with the details of languages and language variation, it is somewhat bizarre that there does not exist a proper technical apparatus to talk about intricate differences in opinion about the precise sense of a name like *English* or *Witotoan* when used in academic discussion. We propose three interrelated concepts—LANGUOID, DOCULECT, and GLOSSONYM—which provide a principled basis for discussion of different points of view about key issues, such as whether two varieties should be associated with the same language, and allow for a precise description of what exactly is being claimed by the use of a given genealogical or areal group name. The framework these concepts provide should be especially useful to researchers who work on underdescribed languages where basic issues of classification remain unresolved.

1. THE PROBLEM: NO CONSENSUS ON WHAT IS A LANGUAGE. [1] The underlying problem that has led to the proposals in the present paper is the well-known and widely discussed issue of how to define the notion of 'language' as opposed to 'dialect' (cf. Anderson 2010, Hammarström 2008), summarized tongue-in-cheek by Weinreich's famous dictum that "a language is a dialect with an army and a navy."[2] There is simply no easy cover-all definition of the term 'language' that would satisfy all users and at the same time be scientifically rigorous.[3] This underlying problem leads to the practical issue of understanding

[1] The concepts presented in this paper arose after lengthy discussions in 2006 while we were both at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. We kindly acknowledge Bernard Comrie for making it possible for us to take part in the fruitful atmosphere for scholarly discussion there. We would also like to acknowledge the role that input from Martin Haspelmath, Sebastian Nordhoff, Gary Simons, and Bernhard Wälchli, as well as a number of anonymous reviewers, played in shaping many of our thoughts on the material presented here. The current paper was written while Michael Cysouw was funded by ERC Starting Grant 240816.

[2] The discussion about the origin of this statement is succinctly summarized at http://en.wikipedia.org/wiki/A_language_is_a_dialect_with_an_army_and_navy.

[3] By 'scientifically rigorous' here, we refer to a system which allows the same standards to be applied in all cases. The notion of 'rigorous' can also be associated with the notion of 'formalized.' However, we present here only a partly formalized version of our model since our primary goal is its justifica-

E-ISSN 1934-5275

exactly what a given scholar is referring to when employing a given language name. Because a term like *English* is not rigorously defined, different individuals might (and will) use it with different intended meanings. Likewise, any higher-order entities, like genealogical families or areal groups, and lower-order entities, like dialects or sociolects, suffer from the same problem of not being strictly defined, possibly leading to misunderstanding and miscommunication.

The fact that people often use the same names with rather different intentions is a fact of life in informal discourse. However, one of the distinguishing features of scholarly communication is attending to this problem when it may cause crucial misunderstandings. The need for the linguistic community to address the terminological issues surrounding 'language' is clearly long overdue. It is further receiving new impetus from the rapidly increasing efforts to document the world's present linguistic diversity and to digitize legacy scholarly resources, since adequately accomplishing these tasks requires a consistent means to identify the entities described by these resources. Moreover, the demand for a rigorous means to refer to entities like 'languages' goes well beyond linguistics (see Dobrin & Good 2009:626), and, even within linguistics, we are seeing renewed attention to the creation of systematic catalogs of information about the world's languages (see, e.g., Hammarström & Nordhoff 2011, Nordhoff & Hammarström 2011, Heaton et al. 2013).

The purpose of this paper is, therefore, to propose a conceptual and terminological framework which we believe can provide the foundation through which the problem of referring to 'languages' can be addressed. While we do not claim our framework is the only way to approach this problem in a rigorous way, to the best of our knowledge, no one has proposed a competing system that can also achieve the same goals. We begin by explaining the role of the framework we are proposing within the wider context of documenting the world's languages (§2) and then situate the discussion with respect to work done on standardization for referring to languages via systems of language codes (§3). We introduce the three key terms of our framework in §4 (i.e., *glossonym*, *doculect*, and *languoid*) and discuss them in detail in §5, 6, and 7 respectively, with §8 offering a summarizing overview of our model. §9 is a discussion of selected issues of implementation, and §10 offers a brief conclusion.

The model to be developed here has already been introduced, unsystematically, in previous work by various authors. Good & Hendryx-Parker (2006:5) is the first published usage of the term *languoid*. The first public presentation outlining the entire terminological framework seen here was Cysouw & Good (2007). This terminology has since been used in various places in the literature, e.g. Bowern (2008:8) and Wälchli (2009:78), both using the concept *doculect*, and Haspelmath (2009:45, fn.14), which mentions the concept of *languoid*. The term *languoid* is also used in various online catalogs like WALS online,[4] Freebase,[5] and Glottolog/Langdoc[6] (see also Nordhoff & Hammarström 2011). Indeed, the

tion on linguistic grounds. If it becomes widely adopted, the development of more formal definitions may be in order, but we believe they would be out of place in this context given the nature of our intended audience.

[4] http://wals.info/

[5] http://freebase.com/

[6] http://www.glottolog.org/

primary motivation for writing this paper is to develop an explicit model in which these terms are embedded to allow references to them in the literature and in online sources to be more adequately contextualized.

2. OVERARCHING GOAL.

2.1 METAMODELS FOR LANGUAGE DOCUMENTATION. The subject matter of this paper is somewhat unusual in a linguistic context, perhaps best characterized along the lines of 'metatheory.' That is, we are not concerned with directly theorizing about the nature of a key linguistic concept, but, rather, are interested in what sort of data encoding model for information about the world's languages would allow us to rigorously discuss what we mean when we refer to a given language in the first place. Thus, for instance, while we could imagine the points made in this paper informing a redesign of a resource such as the Ethnologue, they are not directly relevant to the current construction of the Ethnologue itself (or to its associated language codes—see §3).[7] The presentation of information in the Ethnologue is designed around the assumption that we can enumerate the languages of the world. As linguists, we know this is a fiction, but, for some purposes, it is a very convenient one, which is presumably why the editors of the Ethnologue have adopted it in a publicly-oriented work. Our concern here, by contrast, is not to develop some sort of new and better Ethnologue, but to understand what kind of metamodel is required to allow resources like the Ethnologue to be built on a more solid scholarly foundation.

While we believe our development here of an explicit metamodel around the concept of language is novel (though not completely without precedent, as briefly discussed in §4), there are recent parallels in the development of metamodels to facilitate technical aspects of language documentation more generally. The most prominent current instance of this is probably found in Lexical Markup Framework (LMF) (Francopoulo et al. 2009, Francopoulo & George 2013), which provides recommendations for the construction of electronic lexicons. However, unlike published lexicon encoding standards such as those found, for example, in the Text Encoding Initiative guidelines (TEI Consortium 2013) or in Lexicon Interchange Format (Hosken 2006), LMF does not provide a concrete standard for encoding lexical data but, rather, a standardized way of describing the structure of arbitrary kinds of lexical entries—it is thus a standard for creating standards, or a kind of 'meta-standard.' The resources created by the LEXUS tool, designed to facilitate the construction of endangered language lexicons, are, in fact, encoded in LMF since the framework provides a good balance between the advantages of standardization and the need to ensure that documentary linguists have the flexibility to encode lexical data in ways that they believe to be responsive to a language's lexical patterns and user needs (see Ringersma & Kemp-Snijders 2007).[8] This is a clear case, therefore, where a metamodel has been put to work for documentary linguists.

Another proposed metamodel of relevance to documentary linguistics is connected to the notion of a data category registry, where it is possible to register and 'publish' the

[7] See http://ethnologue.com/. We refer to this resource as the 'Ethnologue' here without making reference to specific editions or publications since our comments apply to it quite generally and focus on the online version of the resource which is frequently updated. Where relevant, we will indicate when our comments about its structure are relevant to specific versions only.

[8] See also http://tla.mpi.nl/tools/tla-tools/lexus/.

linguistic data categories employed in a given resource or project. Wright et al. (2010) (see also Windhouwer et al. 2012) discuss aspects of the use of this metamodel to create a specific data category registry known as ISOcat.[9] While ISOcat is intended to be of general service to the documentary linguistics community, the metamodel on which it is built could be used as the basis of the formation of other data category registries if a given subcommunity were to decide its interests were better served outside of ISOcat. As a further example, within the ISO 639 family of standards for language identification—of which the ISO 639-3 standard is best known in documentary linguistics (see §3)—one also finds the ISO 639-4 standard (see Gillam et al. 2007), which, among other things, can be understood as laying out a metamodel for language coding systems in general.

Metamodels such as these—as well as the one we develop below—are, by their nature, abstract objects, and their relevance to the more usual activities of linguists is not always immediately apparent. It is important, therefore, to bear in mind that, whether it is explicitly laid out or not, the way we talk about the world's linguistic diversity is intimately tied to some metamodel of 'a language,' such as the idea that the referent of a language name can be effectively clarified via the kind of information that is found in an Ethnologue entry or that languages belong to families whose documentary status can be determined by 'summing up' the state of documentation for the languages that belong to them (see, e.g., Hammarström 2010). Moreover, when it comes to arriving at the most accurate possible answers to key questions for documentary linguistics, it seems unlikely that there is any metamodel more important than the one through which we define 'languages' themselves. It is only through the application of some metamodel that we can try to answer such basic questions as, "How many languages are spoken in the world today?," "What percentage of the world's languages are still mostly undocumented?," or "What is the world's most endangered living language family?"

Of course, questions like these have been explored in the existing literature in the absence of an explicit metamodel of the sort we are proposing here (see, e.g., Whalen & Simons 2012). However, we believe that the convergence of new technologies, in particular web-based means of data dissemination, with the increasing emphasis on documenting the world's underdescribed languages, makes this an important time to explicitly consider our metamodel for languages in order to ensure it is aligned as closely as possible to our research practices and needs. We expect that the number of linguists actively engaged in developing such a metamodel will always be relatively small in proportion to the field as a whole. At the same time, we also believe that particular subcommunities of linguists are especially strongly impacted by the version of this metamodel that the field may make use of, with those involved in language documentation and description, as well as typological investigation, being especially reliant on it. After all, work on English will thrive whether or not it is based on a solid conceptual foundation of what we mean by the 'English language.' However, we may not even be able to recognize whether or not some underdocumented speech variety should be allocated the resources we usually devote to documenting a 'language' if we are not clear about which varieties are taken to belong to which languages in the first place.

[9] See http://www.isocat.org/.

2.2 THE GOALS OF THE PRESENT METAMODEL. As already indicated, the leading motivation behind this paper is our belief that the field of linguistics cannot adequately document the world's linguistic diversity without devising a more rigorous way to talk about languages. We should first make clear, however, that we are not arguing that the problems with strictly defining the notion 'language'—or even deciding just what is a 'language'—implies that the term 'language' should be abandoned. In laypeople's discussions, it will of course still be used as it always has been, and rightly so. Likewise, the informal usage of the term 'language' is perfectly reasonable in most cases of both linguistic and non-linguistic scientific discussion, and even we will use the word informally here in cases where we do not believe its potential ambiguity will cause any problems in understanding. However, there are clearly situations in which rigorous application of terminology is needed to allow for the precise specification of one's intentions, to discuss and resolve disagreements, and to ensure that the same kinds of things are being compared when we engage in activities like trying to make numerical statements about the world's 'languages.'

We should also make clear that, in using the term 'language' here, we are almost solely concerned with its use as a label for specific human languages rather than as an abstract cover term for human language in general. We are moreover concerned with 'languages' as attested (or, at least, attestable) entities rather than as abstract mental entities. This orientation arises from the fact that we are primarily motivated in this paper by the practical problem of facilitating more explicit scientific communication among linguists and other scholars with a stake in coming to a better understanding of the nature and distribution of the languages of the world.

A rigorous system for discussing languages is most obviously important for the purposes of cataloging language data and building cross-linguistic databases. But, there are broader considerations as well: Constituencies outside of linguistics often have a large stake in the issue of which varieties are or are not considered to constitute 'languages.' As a field, linguistics is well aware of the impossibility to answer such questions definitively. However, we are at present incapable of systematically codifying our knowledge of the world's languages in terms which would allow those outside of the field to make informed decisions on issues such as language policy based on the information we have collected. Moreover, since online databases, in linguistics and elsewhere, are increasingly coming to be viewed as the authoritative means through which reference information is disseminated, it is clear that we cannot view activities such as developing catalogs of the world's languages as mere bookkeeping efforts. Rather, they actively shape the understanding of our object of study in fundamental ways (see also Dobrin et al. 2009). Thus, the metamodel, on which such catalogs are based, though typically obscured from the view of the average user due to its function as 'infrastructure' rather than 'product,' should not be viewed as just a technical tool, but a more fundamental expression of our conception of the epistemology of the field.

We believe the proposals here can play a crucial role in such applications. However, at the same time, we should stress their foundational nature. That is, they present a guide to improve existing practice in ways which we expect will be beneficial in the long-term, but they cannot in and of themselves 'fix' problems with existing practice. This will have to await the concrete implementation of resources making use of the proposed framework. Indeed, one prominent initiative making use of a variant of this framework has already come online, namely the Glottolog/Langdoc project (Hammarström & Nordhoff 2011; Nordhoff

& Hammarström 2011).[10] Furthermore, we do not intend to claim that our model, in and of itself, represents a full replacement for existing approaches to referring to and cataloging language varieties. Rather, we view it as one piece of larger set of models which will undoubtedly need to be developed.

3. STANDARDIZATION VIA LANGUAGE CODES IS INSUFFICIENT. Superficially, the topic of this paper may appear to be similar to work done on creating standardized sets of codes for referring to languages, most prominently associated with the (largely Ethnologue-derived) ISO 639-3 codes, which are intended to provide three-letter language identifiers for all of the world's languages.[11] However, as discussed in §2, we are interested in a rather different problem, and the framework developed here is meant to complement such efforts, not to supersede them. We consider ISO 639-3 to be a very important and useful effort dealing with problems of language identification, particularly for resources collected by non-linguists or linguists who are not specialists of the languages contained in these resources.[12] We focus on ISO 639-3 here due to its general prominence at present and due to its significance in the domain of standardization for language resources, in particular for resource metadata, but the same general points apply to comparable standardizing efforts, such as the Linguasphere Register (Dalby 1999–2000).[13]

Efforts like ISO 639-3 can be usefully compared to the Linnaean system of binomial nomenclature in biology. This represents a good approximation of the world's biological diversity, and, although it might not be perfect, it is still highly practical and good enough for most situations. Simons (2009) usefully compares language codes to time zones, which only function by partly dissociating solar time from standard time to prevent each locality from being a time zone unto itself. ISO 639-3 codes similarly try to carve up the world of language variation in a way that represents a seemingly useful balance between capturing diversity and facilitating interoperation. Yet, any such fixed list or classification does not help to resolve disagreements about categorization, and it is not designed to do so.[14] The

[10] See http://www.glottolog.org/.

[11] Roughly speaking, the Ethnologue provides information on codes associated with living, or recently extinct, languages (http://ethnologue.com) and The LINGUIST List provides information on older extinct languages in the context of its MultiTree project (http://multitree.org/codes/). The current list of ISO 639-3 codes can be found at http://sil.org/iso639-3/.

[12] In referring to ISO 639-3, we primarily mean the codeset and associated documentation that is published by the standard's registration authority, SIL International. The main characteristics of this codeset and the rules that guide the registration authority are described in ISO (2007). For present purposes, the most significant distinction between our proposal and the system adopted in ISO 639-3 is that we treat explicit connections to an actual documentary resource as critical in the specification of a given language, while ISO 639-3 officially treats the 'denotation' of a given three-letter identifier merely as one or more glossonyms (see http://www.sil.org/iso639-3/scope.asp). §4 discusses glossonyms in more detail.

[13] See http://www.linguasphere.info/.

[14] The fact that ISO 639-3 has adopted a coding scheme making use of only three letters as language identifiers is significant here. This scheme permits the construction of approximately 15,000 identifiers, which is sufficient to code all languages given the current estimated number of the world's languages, roughly on the order of 5,000 to 10,000. However, this limit makes the system unsuitable for coding fine-grained distinctions among all the world's language varieties. The ISO 639-6 stan-

list of three-letter codes in ISO 639-3 is intended to represent a consensus regarding our present knowledge about the delimitation of the world's languages. It is true that it can be updated, but only to the extent that consensus allows.

However, consensus about the identification of languages is often hard to achieve and, moreover, often turns out to be incorrect as new facts becomes known. Therefore, we expect that language experts will never be fully satisfied with the range of decisions that are taken to develop a standard like ISO 639-3, especially with regards to the delineation of groups of closely related speech variants into specific languages. In some cases, it may be that a given expert simply disagrees with current consensus. In others, it may be that a lack of information has made that consensus inherently fragile, and everyone agrees that it could change quite abruptly if more was known about the linguistic situation of a specific group or area.

Nevertheless, there are cases in which there is a clear need for a standard like ISO 639-3 that aims to be comprehensive, and this requires that some kind of decision be taken regarding how to code varieties whose precise status is unclear or controversial. We, therefore, believe efforts like ISO 639-3 have an important place both within the field and beyond. What we believe is also needed, however, is a complementary system, which can provide, among other things, a systematic means to support description of a lack of consensus as well as the documentary basis for consensus decisions. This would allow the precise nature of disagreements to be made clearer, thereby facilitating the linguistic research needed to achieve or further verify consensus to the extent that this is possible or desirable in a specific instance. The need to represent consensus, lack of consensus, and the basis of consensus is, of course, a general problem for researchers, even though we limit the scope of our discussion to issues surrounding the specification of 'language' here.

4. DOCUMENTATION AS THE FOUNDATION. The leading idea behind the model being developed here is that it must be based on aspects of language classification that are largely uncontroversial. When we want to be able to engage in substantive discussions in order to resolve, or merely cleanly delineate, disagreements in what constitutes a 'language' or a 'family,' or even understand the documentary basis of the claim that some set of varieties constitutes a 'language,' we need a common vocabulary based on uncontroversial entities and concepts that will allow us to rigorously examine the basis for different claims. Notions like *English* or *Altaic*, for example, cannot be defined at a fine level of detail without controversy, but we still need a way to talk about them. The Ethnologue and The LINGUIST List together keep track of information providing basic documentation for each ISO 639-3 code, thereby clarifying the intended meaning of these codes. However, these descriptions are still complex combinations of different kinds of data (including location, alternative names and references), which could easily be disputed as defining a given language.

As a solution to these problems, we begin with the desideratum that (for the purposes of language identification in the context of scientific linguistic research) languages can only be identified by reference to existing documentation.[15] The motivation behind this

dard has been proposed to partly remedy this by using codes made of four alphanumeric characters. However, for the approach described in this paper that would still not provide a sufficient number of codes, as will be become clear below.

[15] By 'documentation' here, we mean the word in the general sense of 'material providing informa-

proposal is that, while what constitutes a specific language in some abstract sense will perhaps always be controversial, there is no controversy in simply saying that there exists a given book, sound file, manuscript, or article that contains data documenting some language variety, even if there is disagreement about how that variety should be classified. While, at first glance, this desideratum may seem quite obvious—after all it merely seems to be a more specific instantiation of the general scholarly practice of citing sources—we can merely note here that it is not the basis of the Ethnologue or ISO 639-3.[16] The Ethnologue does include sources as part of the information associated with a given language (understood as an entity with an ISO 639-3 code), but, at least in terms of presentation, these are treated as an additional source of information about the language, whereas we propose a reversal of the code–source relationship as usually understood: The resources are the primary data, with the codes taken to be a kind of metadata applying to the set of resources defining a 'language.'

Of course, even when one bases a language's definition in existing documentation, there might be disagreement about exactly which pages of a document belong to one particular manuscript, or whether a book should be treated as one entity or a collection of separate works. But, these problems are not linguistic issues and are unlikely to have a significant impact on research. Moreover, we should make it clear that, in claiming that it is uncontroversial to say that some object contains data on a language variety, we are not saying that the data itself is uncontroversial. Campbell (1997:13–15), for example, discusses a number of cases of 'fake' or 'mistaken' languages in his survey of American Indian languages. These fake languages may have no connection to any variety that was ever actually spoken. Yet, their mere existence, albeit as fanciful constructions, is not in itself controversial. We, furthermore, do not mean to suggest that there is any threshold for the necessary extent of documentation about a language for it to be 'sufficient' for use in linguistic research. Even a language only known by name would count as documented in the present context. Troyer et al. (1995:9–10), for example, discuss a language for which no data was available beyond a name and a claim of speakers. This constitutes documentation, if not of a linguistically very interesting kind.

The basic entities that we claim should form the foundation for a rigorous definition of languages are thus actual documentary records attesting to the existence of some language variety. We propose to refer to the linguistic varieties documented by these records as DOCULECTS ('documented lect,' see §6). All further language-like objects, like dialects, sociolects, languages, genealogical groups, or areal groups, can then be defined by referring to a collection of doculects on which they are based. The definition of all these language-like objects is thus essentially the same: They are all defined via sets of doculects. Although there are very many different kinds of such language-like objects, we propose to use the name LANGUOID to refer to the superset of all those language-like entities (see §7). A crucial aspect of this conceptualization is that the grouping traditionally called 'language' does

tion' rather than in the more specific sense adopted in the literature on language documentation (see Woodbury 2011).

[16] Similarly, the newer Catalogue of Endangered Languages (ELCat) project (Heaton et al. 2013), which is careful to include bibliographic references supporting the information in its entries, does not necessarily reference to primary documentary resources but, rather, in many cases relies on other general reference works.

not have an inherently preferred or basic status in this model for classifying the world's linguistic variation. This allows for rigorous debate about what constitutes a 'language' without interfering with the system's core classificatory apparatus. Finally, we must also pay attention to the ways that names are used to refer to languoids and doculects, since, ultimately, it is via names that linguists usually talk about different languoids. To preclude any conceptual confusion between the entities and their names, we propose the term GLOSSONYM here for the names that are used to refer to languoids (see §5).

There are, of course, numerous kinds of information about languages that are of interest to linguists and beyond (e.g., autonymic reference, geographic location, sociolinguistic context, number of speakers). However, as discussed above, our interest here is not to devise an all-encompassing reference work for the world's languages or anything along those lines. Rather, our goal is construct a conceptual and terminological foundation for talking about languages that can be rigorously applied at a global scale. Our basic claim is that the triad of concepts proposed here (i.e., languoid, doculect, and glossonym) is a minimal system that nevertheless allows for a rigorous definition of language-like entities. Other kinds of information can easily be added on top of this foundation, but are not needed for their definition. Precedent for this kind of system can be found in earlier work cataloging or classifying languages, such as Loukotka (1968), though, of course, the utility in laying out a generalized and extensible system like this has increased greatly in recent years as online systems of information collection and dissemination have become critical tools for expanding our research horizons (see §2).

5. GLOSSONYMS: NAMES TO REFER TO LANGUAGE-LIKE OBJECTS.

5.1 DEFINITION. We propose the word GLOSSONYM as a technical term for names for a language, for a lect, or for a genealogically or areally related group of languages.[17] Names for such entities are of course commonplace in the linguistic literature—and beyond—as references for languages, language families, etc. However, in and of itself, the glossonym as understood here does not have reference. A glossonym is just a signifier used in the naming of language-like entities. Thus, by saying that there is a glossonym *English*, we only intend to describe the fact that someone has used this string of characters to designate some language-like object.[18]

Glossonyms as defined here are thus not names in the usual sense and might better be called 'glossonym text strings.'[19] That is, a glossonym is not a name, but only the form used to convey a name of a particular type. However, because 'glossonym text string' is somewhat tedious in practical usage, we abbreviate this to simply glossonym. We use the term

[17] The term *glossonym* was suggested to us by Christian Lehmann. However, his usage of the term glossonym seems to be different from our, rather restricted, definition. Unlike *languoid* or *doculect*, the term glossonym has been in use for some time (see, for example, Matisoff 1986).

[18] In principle, glossonyms could take various forms, such as sounds, signs, or textual representations. For purposes of exposition, we will focus on their representation as text strings in this paper.

[19] In fact, it is possible to imagine that a glossonym, even when not conceptualized as a name, may consist of a more than a text string. For instance, the string itself could be annotated for information such as its script or the fact that it may be associated with a word in some language. These are, however, potentially general properties for text strings, rather than being specific to glossonyms, which is why we do not develop them here.

glossonym in this special technical sense partly in order to allow us to better understand the issue of glossonym homology, which we discuss in the next section.

5.2 HOMOLOGY. In principle, any collection of glossonyms could be represented as just an unstructured list of possible names for language-like objects. However, there are conventional associations among glossonyms in the sense that they can be related to each other independently of their specific reference at a given instance. That is, it is sometimes possible to, for example, say that two glossonyms are used to refer to the 'same' entity, even if the specification of that entity is not fixed. For example, the glossonyms 'Altaic' (in English), 'Altaisch' (in German) and 'ałtajskie' (in Polish) conventionally refer to the same language family and can always be used interchangeably in their respective languages, even though what is meant by 'Altaic' may differ across sources, making it a naming equivalence that holds independently of any claim about the extent and interpretation of the grouping. Indeed, this example is a specific illustration of an important organizational aspect of language catalogues: Translations and alternate spellings of language names may remain constant even when the meaning of the name changes. For the organization of knowledge about language varieties this implies that such relations need to be considered as relations among glossonyms and not as relations among their referents.

There are three important inherent relations among glossonyms that we are aware of: spelling variants, language-specific morphological variants, and etymologically related variants. These are all relations among the names themselves, not among their referents. So these relations are genuine relations among glossonyms proper. We propose the term HOMOLOGY as a cover term for these relations.[20] On the synchronic side, linguists would normally refer to homologous glossonyms via concepts like derivation, inflection, or compounding (depending on the grammatical details). On the diachronic side, linguists would normally refer to homology using notions like cognate or loanword (depending on the historical scenario). And for spelling variants and transliterations, most linguists would probably simply treat them as 'the same thing.' The term homology is proposed here to refer to the super-set of all these kinds of relations among forms in which the forms themselves are in some broad sense 'the same thing.'

To make these distinctions clearer, we exemplify different kinds of homologies in the following section. Our primary goal in this discussion is to further justify the inclusion of glossonyms as distinctive objects in our system, rather than to fully explore all of the complications involved in their modeling. Accordingly, we focus on a descriptive presentation of different kinds of homologies rather than offering a formalization of them.

5.3 EXEMPLIFYING GLOSSONYM HOMOLOGY. Consider the glossonyms *Ashéninka, Pichis; Pichis Ashéninka;* and *Pichis Ashéninca*, each associated with the variety given ISO 639-3 code [cup]. In our conceptualization, these are three different, but homologous, glossonyms in the sense that, from a research perspective, they can be interchangeably used to refer to the same thing. Of course, issues of style or choice of research metalanguage may affect which is chosen, but those concerns are not specific to linguistic research.

[20] The term 'homology' was originally suggested by Michael Cysouw in reference to the biological concept of homology (cf. Steiner et al. 2011:94). It has a strong similarity, if not outright identity, to the term 'allofamy' found in Matisoff (1978:17).

Likewise, spellings in different orthographic traditions (including language-specific morphology) can create sets of homologous glossonyms. For example, *Holländisch* (as it is called in German), *hollandaca* (as it is called in Turkish), *голландский* (in Russian), or even オランダ語 (in Japanese) are homologous glossonyms. Furthermore, inflectional or derivational variants like *holländisch, holländische, holländischer* (in German) are homologous glossonyms. In contrast, the three glossonyms *Nederlands* (in Dutch), *holländisch* (in German), and *Dutch* (in English) are not homologous, though they are normally used to refer to a very similar entity.[21] The reverse situation arises with *Dutch* (in English) and *Deutsch* (in German), which are also homologous glossonyms in this conceptualization (across the diachronic dimension) though they do not refer to the same entity when used on their own (though in the glossonym *Pennsylvania Dutch*, the orthographic sequence *Dutch* can refer to the same entity as *Deutsch*). A similar problem is that the English homologous derivations *Turkic* and *Turkish* do not refer to the same entity (with the former referring to a family and the latter a language), though the Dutch homologue *Turks* does have both possible references.

Finally, note that it is quite possible for the same string of characters to arise more than once as a glossonym by pure chance. In such cases the glossonyms are identical, though not homologous, in the sense developed here. With short names there are many such examples, like the string *Aho* as a glossonym for a dialect of Eloyi, a Niger-Congo language from Nigeria (ISO 639-3 [afo]), but also as a glossonym for Aheu, an Austro-Asiatic language of Thailand (ISO 639-3 [thm]). In our model, we treat these instances of *Aho* as employing the same glossonym because they use the same string of characters. The fact that this glossonym refers to two different entities is not captured by the glossonyms themselves, but by the doculects with which the names are associated (see §6). When the same glossonym has been used to refer to different language varieties, access to information on its homologous relationships can be useful in disambiguating a glossonym's intended referent. For example, the *Aho* dialect from Nigeria has homologous glossonyms like *Afo, Afu*, and *Afao*, while the *Aho* language from Thailand has a homologous glossonym Aheu.[22]

At least from the point of view of the typical linguist (rather than an expert in terminology), standardized codes can be understood as a special kind of glossonym. For example, an ISO 639-3 code like [peh] is a glossonym, as is the code [bao] in the World Atlas of Language Structures (henceforth WALS, Dryer & Haspelmath 2011). These two codes, [peh] and [bao], apparently refer to quite similar entities (an Altaic language from China, otherwise known as Baonan or Bonan), but that is unimportant on the level of glossonyms. Conversely, one and the same glossonym of this kind might have completely different meanings in different contexts. For example, there is both a WALS-code [bao] and an ISO 639-3 code [bao], but they are used to refer to the Altaic language Baonan from China and the Tucanoan language Waimaha from Colombia, respectively. One special aspect of standardized codes, like ISO 639-3 or WALS codes, however, is that they do not have any homologs by definition. This is probably the most important characteristic that sets them

[21] See http://en.wikipedia.org/wiki/Names_for_the_Dutch_language for a concise summary of the complications when referring to the Dutch language.

[22] Further note that the string *aho* is used as an ISO 639-3 code for Ahom, an extinct language of India, and as a code in the World Atlas of Language Structures (Dryer & Haspelmath 2011) for Arapaho, an Algonquian language spoken in the United States.

apart as a special kind of glossonym.[23]

These examples illustrate the value in giving special consideration to how glossonyms interrelate regardless of their specific referent when trying to formulate a metamodel for the specification of language varieties. In many cases (though far from all) it will be possible to separate sets of homologous glossonyms into subsets closed under (mathematical) transitivity, i.e. the 'ideal' situation in which there is a collection of homologous glossonyms that can all be used interchangeably or whose use can be determined by aspects of context not relevant to scientific questions of linguistics (e.g., serving as translational equivalents across different metalanguages). However, establishing such groups requires extensive documentation and collection of glossonyms and their usage, which, of course, requires a much more rigorous means of defining what we mean by particular 'languages' in the first place.

6. DOCULECTS: THE BASIS OF RIGOROUSLY DEFINED LANGUAGES (AND BEYOND)

6.1 DEFINITION. We propose the term DOCULECT ('documented lect') for a linguistic variety as it is documented in a given resource.[24] This term is deliberately agnostic as to whether or not that variety can straightforwardly be associated with a particular 'language' or 'dialect' and, instead, merely focuses on the fact that there is a document either about the relevant variety or directly recording that variety in some way (e.g. as a book written in that variety).

The motivation behind developing the notion of doculect is to offer a scholarly useful (though in practice slightly cumbersome) concept that offers a basis on which to compare differences in opinion about the identification of languages. There are two reasons for this. First, since it is impossible to debate the status of any language in a rigorous way when there is no information available about it, this implies that any variety that is to be subject to reasonable linguistic debate must be associated with documentation. If someone claims that there is a language called 'Gobbledygook' without providing any further information or reference to other sources, the lack of an evidentiary basis for its existence makes it essentially irrelevant for research purposes.[25] Second, although it is clearly difficult to reach consensus about what exactly is 'English' and what is not, it is trivial to agree on the fact that there is a particular document that is written in (or is about) a language which someone has named English. Indeed, this kind of agreement is so trivial as to essentially never be remarked upon.

A scholar may, of course, object to the claim embodied by the association of a specific glossonym with the content of a given resource by suggesting that the 'language' of the resource has been misidentified. Yet, even in contesting that identification, they nevertheless implicitly accept that the pairing itself exists.[26] Unlike languages, the existence of a

[23] Another way in which they differ from more usual glossonyms is that they are not elements of any natural language except by accident.

[24] The term *doculect* was suggested to us by Martin Haspelmath.

[25] Technically, in the model developed here, the pairing of the glossonym *Gobbledygook* with the resource where it is mentioned could constitute a doculect, but a useless one.

[26] Note that in the model we are developing here, there is no sensible way for a glossonym–resource pairing (i.e. a doculect, in our parlance) in and of itself to be 'incorrect' since the glossonym itself has no referent. Rather, disagreement arises only when the referent of a glossonym found in a given

specific doculect will only be contested in very unusual circumstances, allowing doculects to serve as an appropriate basis for rigorously defining what varieties are encompassed by reference to a given language.

We restrict our use of this term to linguistic varieties associated with concrete data, whether or not that data is 'correct' in some sense. This data need not be purely linguistic, but could instead comprise, for example, information such as speaker demographics, geographic location of speakers, cultural traits of speakers, or even just a statement that there is a community in some location that is claimed to speak a language with a certain name. Of course, for certain applications, some scholars may only want to examine doculects where the available information on them passes some pre-determined 'threshold'—for instance, that actual utterances or lexical items are given. We do not attempt to define any such a threshold here, or to make a division between 'real' linguistic data and 'other' data, since it is not at all obvious how to draw such lines. We view inclusion of such information, therefore, as something which should be built on top of our model, if users consider it desirable, rather than being part of its foundation.

Our definition of doculect implies that some language-like entities that are the subject of linguistic scholarship will never be associated with doculects, with the most conspicuous example being language families. A resource describing a language family may contain information on many doculects—including, of course, proto-languages—but the family itself would not be one. This approach further means that a resource containing data from a 'dialect' survey will be treated as not documenting any relevant 'language,' from the perspective of the doculect, but, instead, its data will be understood as comprising a number of doculects associated with the various dialects. Such a survey will also (implicitly or explicitly) claim that those doculects comprise varieties of a common 'language,' but that is a separate concern in our model (see §7).

We acknowledge that it may, at times, be difficult to clearly ascertain all the doculects found in a given resource, or to what doculect a given piece of data may belong. For example, dictionaries of a given 'language' may contain dialect variants for some entries, but not others (e.g., a dictionary of English giving British and American variants for certain words). Does this mean that entries not associated with dialect variants should be taken as belonging to multiple doculects? Or, should they be treated as belonging to a single 'standard' or 'unified' doculect? The framework developed here cannot answer such a question generally, and each problematic case will have to be resolved separately. Nevertheless, the fact that it reveals this as an issue needing consideration is an indication of the role of our framework as a foundation for a more principled discussion of issues such as what constitutes a 'language.'[27]

In formal terms, we define a doculect as a pairing [resource; glossonym].[28] The glos-

doculect is otherwise associated with a more general languoid (see Section 6).

[27] Of course, some of the problems we mention here are connected to the general concern of representing 'multilingual' data (which, of course, extends to 'multivarietal' data), which has already been given serious consideration for some time (see, e.g., Simons 1998:11–15).

[28] While the inclusion of a resource identifier in a formal definition of a doculect is essential (since this is what makes it a 'documented' variety), it could in principle be paired with some other identifier than a glossonym, for example a reference or series of references to all places in the relevant resource where documentation of that variety can be found. We propose to use a glossonym here because we

sonym in the ideal case refers to the actual string of characters used to refer to a linguistic variety in the resource itself (if it is about that variety) or in the metadata of a resource (if it is in that variety), though, in some cases, a glossonym may have to be constructed on the basis of information in a resource when a specific one has not been proposed. For example, a dialect survey of the varieties of a language as found across different villages may not give explicit names to each dialect but, rather, refer to them via the village names themselves. In such cases, one could simply construct a glossonym by using a qualified name along the lines of *variety of [glossonym] as spoken in [village]* in order to link each village's variety to an explicit doculect. The resource in the doculect pairing [resource; glossonym], in turn, must be uniquely determined by means of an appropriate identifier. A doculect can, therefore, be thought of as a string of characters used to refer to a documented language-like object in the context of a specific uniquely identified resource.[29]

Our claim, in effect, is that the minimum requirement for making a language variety 'real,' at least for the linguist, is the pairing of a resource with a glossonym, and it is only at the level of this pairing that we can build a rigorous means through which to structure debates about questions like what is or is not a language, whether two varieties represent the same language, whether or not two languages are part of the same language family—or even whether a 'language' is documented at all and to what extent. As is generally the case when developing a formal model based on real-world practice, there is a degree of simplification involved in this conceptualization. In this case, for instance, we do not explicitly characterize that a doculect, in informal terms, involves some sort of assertion on the part of an author that there is a coherent linguistic variety that can be named in the first place. Since our goal is to model knowledge of the world's languages, rather than the full set of logical statements required to make sense of scholarship in general, we believe such simplification is warranted. This follows from our general interest in developing the simplest reasonable framework instead of a logically 'complete' one.

6.2 THE PROBLEM OF RESOURCE GRANULARITY. There are, of course, practical problems in working with the notion of doculect. For example, it can be difficult to determine what exactly constitutes a single resource. Should each sound file from a fieldwork trip be treated as a single resource, all recordings from one session, or all recordings from an entire research period? We do not believe that such questions can be answered in any general way. Rather, they have to be decided upon in each particular situation. Moreover, problems of granularity like this can (and should) ideally be addressed without referring to the language(s) being documented in the resource. Indeed, in the model being developed here, this is especially appropriate since it can provide a means for preventing circularity in the definition of a doculect.

think it represents a reasonable balance between tractability and explicitness given that common practice in linguistics is to associate data in a given resource to a given variety by means of some name. Moreover, use of glossonyms would allow, in principle, for the exploitation of homologous glossonym relationships (see §5.2) to facilitate resource discovery.

[29] When multiple glossonyms are used to refer to the same variety within the same source, these can, in principle, be represented as a set of glossonyms belonging to one doculect, though by the formal definition given above they would technically be understood as different doculects associated with the same documentation. We do not see much hinging on the distinction.

An apparently more difficult issue arises in determining the integrity of a given collection of language data as presented in a single resource, given that it will inevitably contain internal variation making determination of its 'variety' problematic. For instance, a particular resource might be considered to consist of only one lect by one researcher, though another researcher might consider it to be two different lects mixed together into one resource. By way of example, consider a recording of a conversation in which the two participants each exhibit noticeable idiolectal variation: one researcher may say they represent the same dialect, while another would treat them as distinct dialects. However, in the model developed here, this problem, while quite significant to interested linguists, is only apparent in the context of language identification. Since there is disagreement on the nature of the 'language' in the resource, there will be subsequent disagreement in the doculectal assignment of the content resource. The 'lumping' linguist is likely to say the resource documents one doculect, while the 'splitting' linguist will likely say it documents two. In the latter case, there may not be an explicit glossonym in the resource for each of the two varieties, but one can be straightforwardly constructed from the resource metadata by using the speaker identifiers (e.g., *dialect of [glossonym] as spoken by [speaker]*). Nevertheless, there will still be an associated practical problem of ensuring that the presence of a 'lumping' doculectal assignment along with a 'splitting' one may hinder the ability of a non-expert to discover or make use of the material within that resource. We discuss how one might address problems like these in a general way in our treatment of languoids in §7 and merely re-emphasize here that the goal of our model is not to resolve all disagreements but, rather, make it possible for them to be debated more rigorously. In a hypothetical case like the one above, we view it as a positive feature that it would make the precise nature of the disagreement between a dialect 'lumper' and a dialect 'splitter' quite clear.

6.3 EXEMPLIFYING DOCULECTS. As an example of the conceptualization of doculects as a pair of source and glossonym, consider the following purely illustrative selection of sources dealing with the Huitoto subgroup of the Witotoan languages, a family spoken on the Peruvian-Colombian border in South America. In ISO 639-3 there is a separation of Huitoto into three different languages:

– Huitoto Minica (ISO 639-3: hto, alternate names 'Meneca' or 'Minica'),
– Huitoto Murui (ISO 639-3: huu, alternate names 'Bue' or 'Witoto', dialect 'Mica')
– Huitoto Nüpode (ISO 639-3: hux, alternate names 'Muinane Huitoto' or 'Nipode Witoto')

Consider, then, the *Diccionario Huitoto Murui* by Burtch (1983). The SIL Language and Culture Archives classifies this dictionary as belonging to Huitoto Murui.[30] However, in the introduction of Burtch 1983 it is stated that various entries in the dictionary are explicitly referenced as belonging to different dialects, namely:

– el dialecto murui del río Cara-Paraná
– el dialecto meneca (mɨnɨ́ca o mɨ́nɨca para los huitoto)

[30] See http://www.sil.org/resources/language-culture-archives.

– *el dialecto muinane (muìnánɨ para los huitoto muinane)*
– *el dialecto mɨca del río Cara-Paraná*

Accordingly, the data in Burtch (1983) represents five different doculects in the model developed here, with associated distinct glossonyms. We will formalize this using an ad-hoc book identification of *Burtch1983DiccionarioHuitotoMurui* separated from the glossonyms by a semicolon. Note that the glossonyms are the actual strings as used in the source, and multiple glossonyms for the same language variant used in this source are separated by commas for purposes of presentation.[31]

– [Burtch1983DiccionarioHuitotoMurui; *huitoto murui*]
– [Burtch1983DiccionarioHuitotoMurui; *murui del río cara-paraná*]
– [Burtch1983DiccionarioHuitotoMurui; *meneca, mɨnɨ́ca, mɨ́nɨca*]
– [Burtch1983DiccionarioHuitotoMurui; *muinane, muìnánɨ*]
– [Burtch1983DiccionarioHuitotoMurui; *mɨca del río cara-paraná*]

Next, the SIL Language and Culture Archives probably erroneously classifies the Vocabulario Huitoto Muinane by Minor & Minor (1971) as belonging to Huitoto Minica (ISO 639-3: hto). The introduction to the book clearly indicates that it should be classified as Huitoto Nüpode (ISO 639-3: hux), as it says that, "los esposos Minor comenzaron el estudio del idioma en 1952, viviendo con los huitoto muinane de Estirón" (Minor & Minor 1971:viii).[32] Treated as a doculect, this could simply be coded as follows, independent of any decision of the ISO classification of the data in this book.

– [Minor1971VocabularioHuitotoMuinane; *huitoto muinane de estirón*]

The two academic publications listed in the SIL Language and Culture Archives as dealing with Huitoto Nüpode (ISO 639-3: hux) are the article Witoto vowel clusters by Minor (1965) and the collection of wordlists from Nies (1976). The language name 'Witoto' in the title of Minor (1965) might suggest a very general paper dealing with Witotoan in general, but actually the paper is based on a very special variant, namely "the Muinánī dialect spoken by approximately 12 families situated on the Ampiyacu River in Peru, a little above the site of Pucaurquillo" (Minor 1965:131). This can be coded as a doculect as follows (with the glossonym shortened for purposes of presentation):

– [Minor1965WitotoVowelClusters; *witoto, muinánī*]

[31] In this example (and others below), we have removed all capitalization from the glossonyms. However, in any practical implementation it is possibly better to explicitly include differently capitalized glossonyms. The rules of capitalization differ strongly across the world's languages and should not be assumed in the organization of glossonyms.

[32] "The Minors began their study of this language in 1952 while living with the Huitoto Muinane of Estiron" (translation MC & JG).

Nies (1976) gives wordlists for various Peruvian languages, and it includes a number of doculects (only the last two of which belong to the Witotoan family). The fact that this source is listed as containing data on Huitoto Nüpode (ISO 639-3: hux) in the SIL Language and Culture Archives is probably an error. In Nies' collection of wordlists, there is also a list for Bora Muinane (ISO 639-3: bmr). This name Muinane was possibly mixed up with Huitoto Muinane.[33] However, in establishing the doculects, this difficulty can be completely ignored. The fact that a glossonym is mentioned suffices to establish the doculect, while connecting this doculect to a standardized identifier is a distinct concern, as we have discussed above. Referring to the actual glossonyms used in the source, the following doculects can be established in Nies (1976).[34]

- [Nies1976ListasComparativas; *campa nomatsiguenga*]
- [Nies1976ListasComparativas; *culina*]
- [Nies1976ListasComparativas; *amuesha*]
- [Nies1976ListasComparativas; *piro*]
- [Nies1976ListasComparativas; *chayahuita*]
- [Nies1976ListasComparativas; *achual*]
- [Nies1976ListasComparativas; *huitoto meneca, mɨnɨca*]
- [Nies1976ListasComparativas; *bora muinane*]

We see, then, that based on these four publications (Minor 1965, Minor & Minor 1971, Nies 1976, Burtch 1983) we can define fifteen different doculects. In defining these doculects, we have not taken any stance as to how they should be related to each other or classified (more about that will be said below in §7.3). In general, it should be rather uncontroversial to enumerate such doculects in almost all cases. Because of this, we propose to use doculect as the basic level of language identification and classification. Differences of opinion, or other kinds of scholarly discussion, can then be formulated in relation to these basic objects where precise reference is required, and we propose a model for accomplishing this in the next section.

7. LANGUOIDS: GENERALIZING THE NOTION 'LANGUAGE'

7.1 DEFINITION. We propose the term languoid ('language-like object') to refer to an entity used to designate any (possibly hierarchical) grouping of doculects, in principle ranging from a set of idiolects to a high-level language family.[35] We view it as a generalization

[33] The word *muinane* means 'downriver' in Witotoan, which explains why it is used for completely different languages.

[34] For ease of exposition in this paper, we only consider the problem of associating a given resource, however defined, with the doculects described within it. Clearly, for some applications, it would also be valuable to be able to explicitly associate subparts of a resource, e.g., different wordlists, with the doculect they are associated with. (This would also be useful in cases where data contains instances of code switching.) We assume that existing markup conventions, whether using inline or standoff techniques, would be sufficient to deal with this problem.

[35] The term *languoid* was originally suggested by Jeff Good in adaptation of a similar usage of the *-oid* suffix for language groups in African linguistics, like Bantoid or Nupoid (see also Good & Hendryx-Parker 2006).

of the notion embodied by the term 'language' insofar as both terms represent a grouping of varieties, with languoids simply representing any imaginable grouping without the constraints generally associated with terms like language, dialect, or family. As with doculects, we are concerned with explicitly specifying the sense of a languoid as proposed in a given resource, en route to developing an implementable metamodel. We, therefore, depart from the informal sense of languoid seen in previous work using the term, which retains some of the conceptual issues surrounding language (e.g., consideration of when a linguistic entity merits receiving a name), and formally characterize languoids using the following recursive definition: <resource; glossonym; *list*(languoids)>.[36] If this system were widely adopted, it is imaginable that formal statements of the composition of a languoid could be embedded directly into the resource that is describing them, in which case one would need to allow the reference to be the current resource.

The list of lower-level languoids in the definition of a higher-level languoid can be empty. Likewise, not all glossonyms might explicitly be stated, for example when not all subgroups in a tree are named. Obviously, this would fall short of ideal practice due to the lack of explicitness, but accepted practice in linguistics often falls short of the ideal, and it must still be modeled somehow. Such unspecified languoids could arise, for instance, when one is modeling the content of a standardized language code list in which the denotation of one of the codes is not made explicit, when a resource discusses a language family without clearly indicating which languages are assumed to belong to it, or in work discussing major languages where it is unusual, and often impractical, for authors to specifically cite the documentation that underpins our knowledge of them. The more interesting case, for the present context, however, is when a languoid resolves to a set of 'atomic' doculects that provide the basis for a rigorous definition of the languoid itself. In this sense, a doculect can be considered to be a special case of a languoid, with the important property that it does not allow for further embedding of languoids and, therefore, is the endpoint of any recursive formulation.[37]

Setting aside these special cases, a languoid can usefully be understood as a grouping mechanism for doculects. This grouping must itself be described in some resource, and it will (normally) be given a name or identifier in the resource (i.e., a glossonym). Thus, the languoid is defined by bringing together three pieces of information: a resource, a glossonym as used in the resource, and the set of doculects claimed to form a group.

Beyond this, there are two additional points to consider. First, since languoids are recursively defined, it is possible for them to contain internal hierarchical structure, e.g. three doculects A, B and C might form an intermediate group X, only to be joined with another doculect D on the next higher level to form languoid Y. In such a situation, both X and Y are languoids. Languoid X is 'simple' in the sense that all of its component languoids are

[36] We use the word 'list' in its general sense of a specification of a set of items, and do not mean to imply the order of elements is significant. The formal definition that we give does not attempt to give all possible restrictions on what constitutes a 'sensible' languoid (e.g., it does not rule out that a 'parent' languoid may have itself as a 'child') because our goal is to describe linguistic practice, which has its own independent restrictions, rather than to create a model which only allows formalization of what linguists deem to be 'correct.'

[37] We could then redefine a doculect formally as <resource; glossonym; list()>, though we will refrain from using this formulation due to our conceptual emphasis on the doculect/languoid distinction.

doculects. Languoid Y is a slightly different kind of entity because of its internal structural complexity. In the model developed here, all intermediate nodes in a hierarchical grouping of doculects are necessarily languoids.

Second, as already indicated, resources will often not mention all doculects on which a particular languoid is based, even if they mention some of them. Although they ideally should do so, this can clearly be impractical, or at least was impractical before recent advances in the digital representation of data. For instance, no existing resource describing the Niger-Congo family will reference all known doculects that form the evidentiary basis for the family. Similarly, ISO 639-3 three letter codes reference those languoids that the standard considers to be languages, but do not explicitly associate those codes with the associated doculects. Rather, the assumption is that the use of those codes as glossonyms in a documentary resource will facilitate discovering those doculects that comprise a languoid. Indeed, languoids associated with an external authority like ISO 639-3 can, for practical purposes, often be treated as doculects in the sense that they can serve as endpoints in a recursive languoid structure, assuming the authority is trusted. We leave open the general question of determining when a given data source may be more or less 'trusted'—an issue which applies equally well to doculects and languoids—since, while clearly important, it falls outside the scope of the more abstract model we are developing here.

The fact that languoids are a grouping mechanism whose structure is effectively that of a tree does not come with an expectation that all of the languoids that might be defined in order to capture scholarly knowledge of the world's languages will themselves neatly arrange into a single tree. While certain kinds of research activities (e.g., language classification) may work under an assumption that there is a single ideal 'languoid tree,' there are simply too many ways one might want to group languages (e.g., areally, genealogically, opportunistically, etc.) and too many opportunities for disagreement for a general model for documenting linguistic knowledge to be built around the idea that there will only be one way to group attested speech varieties. Of course, allowing for complex networks of languoids, rather than a single unified tree, creates problems for data processing, but we believe emerging techniques can allow the resulting difficulties to be overcome (see §7.3).

7.2 LANGUOIDS ARE AGNOSTIC ABOUT THEIR STATUS. In our definition of a languoid there is intentionally no mention of what kind of language-like entity a given languoid is supposed to be. Rather, languoids are simply any (hierarchical) grouping of doculects as proposed in some resource, ideally explicitly, but, of course, often only implicitly. In current practice, most frequently, such groupings are used to represent sources that are all understood to deal with the same dialect or language. But languoids can just as well be used for proposals of genealogical groupings of languages, areal clusters, macro-languages, or even typological samples. This agnosticism crucially allows for rigorous discussion about any language-like entity without the need to impose consensus about the status of the entity. For example, people might agree on the fact that a group of doculects is sensibly grouped together into a languoid, but they might differ on the assessment of whether this languoid should be considered a dialect or a language.

Likewise, the dividing line between a language (with various dialects) and a genealogical group (with various daughter languages) is often difficult to draw, so it seems fruitful to us to have the possibility to separate the issue of whether the composition of a languoid can be agreed upon from the issue of just what kind of language-like object a given languoid happens to be. The situation is similar with the question of areal convergence versus genealogical descent: Is a particular group of languages really a genealogical family, or are the languages included in the group similar due to long-term convergence? These are often difficult questions that can take decades to resolve (if, in fact, they are resolvable). Therefore, we propose not to use a limited set of classificatory terms in the core definition of language-like entities, but to consider notions like 'language,' 'dialect,' 'family' or 'Sprachbund' to represent an additional kind of information that can be independently associated with languoids (e.g., via ancillary metadata or other comparable devices).

As an example of this problem, consider Echeverri (2009), who requested a code change to ISO 639-3 to add the 'dialect' Mïca as a separate ISO 639-3 entity alongside the other three Huitoto languages (cf. §6.3). The request was rejected by the ISO 639-3 Registration Authority with the argument that, if there should be any code change, then it seemed more sensible to combine all Huitoto variants into one language, instead of splitting them further: "The requesters make a case for a distinct code element for Mïka in parallel with the three other varieties, while referring to all as 'dialects' of the Huitoto language […]. The evidence given might equally support a case for merging the existing three varieties of Huitoto into a single code element, rather than adding more varieties."[38] What is important here is that, although the registration authority rejects the specific request regarding 'language' classification, there does not appear to be any disagreement about the existence of the relevant languoids and their subgrouping. The example clearly suggests the need to distinguish between the existence of a given language-like entity from its categorization into a narrow set of languoid types like 'language' or 'dialect.'

7.3 LANGUOID COMPATIBILITY. The notion of a languoid is central to addressing the problem of rigorously examining differences of opinion in language classification. Consider a situation in which two different languoids appear to refer to more or less the same thing despite not having precisely the same composition (e.g., in the case of proposals for the structure and composition of language families). To bring order to such a situation, an intuitive first reaction might be to attempt to group languoids into sets of identical (or 'sufficiently' identical) languoids. However, while this might be possible in some simple cases, attempting to do this consistently across all languoids would quickly become intractable.

Therefore, instead of focusing on a simple, but rigid, notion like 'identity,' we believe that what is instead is needed is a more flexible notion that we term COMPATIBILITY, which we will develop here. That being said, we should make clear that our model would allow for the independent use of various notions of 'identity' or 'compatibility,' depending on the needs of the user, and this discussion can therefore be considered as much an illustration of how our model can allow for new kinds of comparison as it is a specific proposal for a metric of comparison.

In the sense to be developed, two languoids would be compatible to the extent that they do not contradict each other. Note that languoid-compatibility is not transitive in the mathematical sense, i.e. when A is compatible with B, and B is compatible with C, then it is not necessarily the case that A is compatible with C. Compatibility between groups

[38] See http://www-01.sil.org/iso639-3/cr_files/PastComments/CR_Comments_2009-011.pdf.

of languoids can be relatively straightforwardly investigated by using consensus trees or consensus networks from bioinformatics, as will be illustrated below.

For example, consider the doculects discussed in §6.3. Of those, the following doculects are given in the Ethnologue as being part of the Witotoan family.[39]

– [Burtch1983DiccionarioHuitotoMurui; *huitoto murui*]
– [Burtch1983DiccionarioHuitotoMurui; *murui del río cara-paraná*]
– [Burtch1983DiccionarioHuitotoMurui; *meneca, mɨnɨ́ca, mɨ́nɨca*]
– [Burtch1983DiccionarioHuitotoMurui; *muinane, muìnánɨ*]
– [Burtch1983DiccionarioHuitotoMurui; *mɨca del río cara-paraná*]
– [Minor1971VocabularioHuitotoMuinane; *huitoto muinane de estirón*]
– [Minor1965WitotoVowelClusters; *witoto, muinánī*]
– [Nies1976ListasComparativas; *huitoto meneca, mɨnɨca*]

Drawing on the Ethnologue family tree that was available online relatively recently (Lewis 2009), these doculects form the languoid in Figure 1 labeled Witoto proper. In the representation below, we use angle brackets to represent non-doculectal languoids, while maintaining the use of square brackets for doculectal languoids, but otherwise use the same source–glossonym pairing notation to refer to both types.[40]

```
<Ethnologue2009; witoto proper;
        <Ethnologue2009; minica-murui;
            <Ethnologue2009; huitoto minica;
                [Nies1976ListasComparativas; huitoto meneca, mɨnɨca],
                [Minor1971VocabularioHuitotoMuinane; huitoto muinane de estirón]
                >
            <Ethnologue2009; huitoto murui;
                [Burtch1983DiccionarioHuitotoMurui; huitoto murui],
                [Burtch1983DiccionarioHuitotoMurui; murui del río cara-paraná],
                [Burtch1983DiccionarioHuitotoMurui; meneca, mɨnɨ́ca, mɨ́nɨca],
                [Burtch1983DiccionarioHuitotoMurui; muinane, muìnánɨ],
                [Burtch1983DiccionarioHuitotoMurui; mɨca del río cara-paraná]
                >
            >
        <Ethnologue2009; nipode;
            <Ethnologue2009; huitoto nüpode;
                [Nies1976ListasComparativas; bora muinane],
                [Minor1965WitotoVowelClusters; witoto, muinánī]
                >
            >
        >
```

FIGURE 1. Languoid–doculect representation of Witoto proper in the Ethnologue in 2009

[39] In this list, the doculect [Nies1976ListasComparativas; bora muinane] is ignored due to its obvious misclassification.

[40] The SIL Language and Culture Archives includes many more resources about Witotoan languages, and this ad-hoc selection is only used for illustrative purposes. Furthermore, we recognize that a reference work like the Ethnologue is, in effect, a perpetual work in progress. We therefore distinguish between its proposal for the structure of this group of languages and the one we believe is more accurate in order provide useful examples for discussion, rather than intending this to be an academic criticism of the Ethnologue classification.

Taking a different perspective (based partly on points discussed in §6.3), we might propose a different division of this group into four dialects, all on the same level. Such a languoid could be represented as in Figure 2.

```
<ThisPaper2013Languoids; huitoto;
        <ThisPaper2013Languoids; murui;
                [Burtch1983DiccionarioHuitotoMurui; huitoto murui],
                [Burtch1983DiccionarioHuitotoMurui; murui del río cara-paraná]
                >
        <ThisPaper2013Languoids; minica;
                [Burtch1983DiccionarioHuitotoMurui; meneca, mɨnɨ́ca, mɨ́nɨca],
                [Nies1976ListasComparativas; huitoto meneca, mɨnɨca]
                >
        <ThisPaper2013Languoids; muinane;
                [Burtch1983DiccionarioHuitotoMurui; muinane, muìnánɨ],
                [Minor1971VocabularioHuitotoMuinane; huitoto muinane de estirón],
                [Minor1965WitotoVowelClusters; witoto, muinánī]
                >
        <ThisPaper2013Languoids; mika;
                [Burtch1983DiccionarioHuitotoMurui; mɨca del río cara-paraná]
                >
        >
```

Figure 2. A different languoid–doculect representation of Witoto proper

Because of their relatively simple tree-like structure, any number of such languoids based on the same doculects can be compared, for example by making a consensus network (Holland & Moulton 2003), allowing for the investigation of agreements and disagreements between the languoids. The consensus network of the two languoids presented above is shown in Figure 3, and it can be understood here to represent the extent to which the languoid structures in Figure 1 and Figure 2 group the doculects in the same way.[41] Boxes in Figure 3 represent cases of disagreement in the structure of the languoid; simple branching indicates agreement. At the bottom left, the various variants described in Burtch (1983) are shown. At the top, the various Muinane descriptions are grouped, and to the bottom right the Meneca descriptions are grouped. Note that there is actually a large amount of disagreement between the two languoids, as there is only one clear branch showing agreement—that is, the branch uniting the two dialects of Murui as described in Burtch (1983).

[41] This figure was made by using the software SplitsTree (Huson & Bryant 2006), available online at http://splitstree.org.

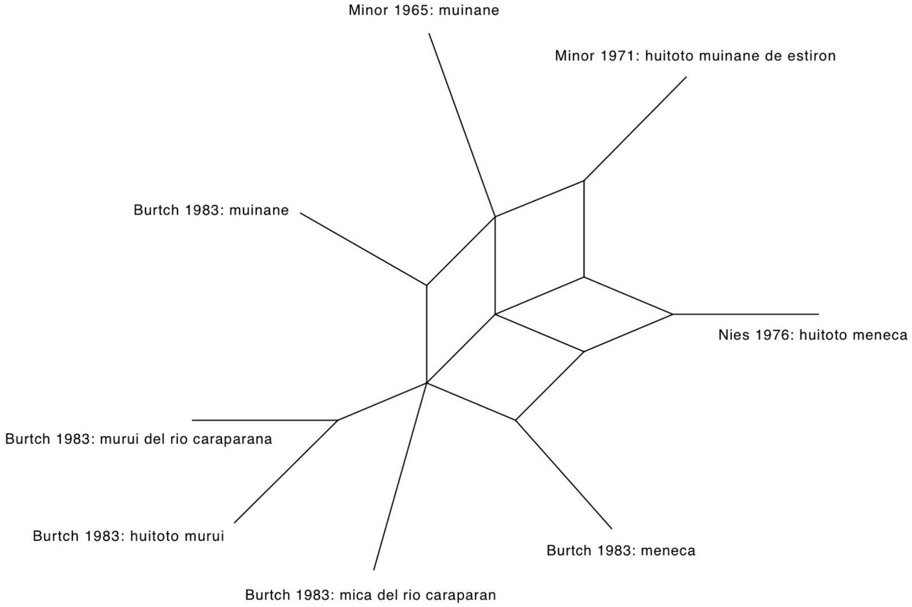
Minor 1965: muinane
Minor 1971: huitoto muinane de estiron
Burtch 1983: muinane
Nies 1976: huitoto meneca
Burtch 1983: murui del rio caraparana
Burtch 1983: huitoto murui
Burtch 1983: meneca
Burtch 1983: mica del rio caraparan

FIGURE 3. Consensus Network of two different languoids on Witotoan

There are various related methods available in the field of bioinformatics to investigate agreements among large sets of tree-like objects. One especially fruitful approach for the purpose of combining languoids seems to be the 'supernetwork' proposal of Huson et al. (2004), which is useful for comparing trees that do not share all their leaves. In general, the investigation of consensus between trees is a large field in bioinformatics, which has already seen significant applications in linguistics, especially for comparative linguistic studies (see, e.g., Nichols & Warnow 2008 for overview discussion of the application of phylogenetic methods from biology to questions of historical linguistics). Thus, linguistic data that can also be modeled in terms of trees with overlapping sets of leaves can readily make use of techniques that have been developed (or will be developed) by bioinformaticists. In practical terms, this means that if one were to develop a database of doculects and languoids on the basis of the model developed here, data processing technologies are already in place to facilitate exploration of the data.[42]

8. OVERVIEW OF THE MODEL. Having discussed the three elements of our model in some detail, we present a schematic overview in Figure 4. A set of resources, in some way documenting a linguistic variety, has been brought together as part of the specification of a languoid. Each resource is associated with a single glossonym to form a doculect. These doculects are in turn grouped together and paired with another glossonym to form

[42] When considering potentially useful technologies in this regard, the emergence of the linked data paradigm (see Chiarcos et al. 2012) as applied to data about languages is also likely to play a useful role in any implementation of the model developed here.

a languoid. Also depicted is the possibility that a given languoid may be associated with metadata, for instance an association with an ISO 639-3 code, an indication of number of speakers, geographic location, etc., and possibly even a characterization as to whether or not the languoid represents a 'language,' 'dialect,' etc. We depict this in dotted lines to schematize that such metadata, while of clear practical relevance, falls outside of our core conceptual scheme. Indeed, this is a necessity since it is in the specification of much metadata of this kind where we expect there to be the most disagreement, and this classificatory system is explicitly designed to help factor out consensus from disagreement. Finally, we have added a circular arrow from the languoid back into itself to schematize the recursive nature of the object.

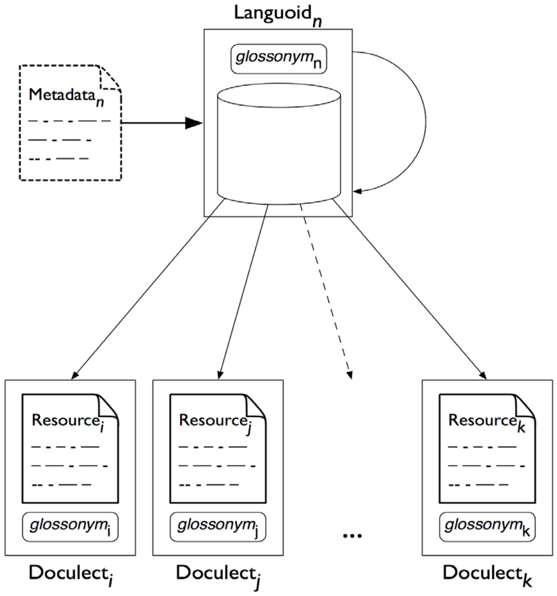

FIGURE 4. Overview of model

9. IMPLEMENTATION: TOWARDS CATALOGUING LANGUAGE VARIATION. The proposals in this paper are of a general nature. However, the concepts developed allow for a direct implementation. We will not explore this topic here in great detail, but see, for example, Hammarström & Nordhoff (2011) for discussion of some important details relevant to using them to catalog language resources. Nevertheless, there are a few specific points concerning practical implementation that we believe are worth remarking upon at this point.

First, in our conceptualization, any ISO 639-3 code (as used widely in current practice) is to be understood as referring to a languoid. This is because the existence of the code represents, in effect, a claim by the ISO 639-3 Registration Authority that there is some 'language' denoted by the code. From that perspective, it is clearly sensible to associate documentary sources to 639-3 codes, though our understanding of the meaning of such associations is slightly different from widespread conceptions. In our view, associating a source to an ISO 639-3 code should not be understood as stating that a given source is 'about' a certain language.[43] Quite to the contrary, it is the code itself that is further specified by virtue of being associated with that source. In other words, the collection of sources (here, doculects) associated with a given code in effect defines the meaning of the code (a languoid). In fact, there is nothing in the ISO 639-3 standard itself (as it consists merely of a listing of three-letter codes and associated language names) that would allow any other conception to be coherent for the purposes of linguistic scholarship.

Second, recall that, in the definition of doculects, the notion RESOURCE plays a central role.[44] A resource can take on many forms, though preferably it is a source that contains actual data about a specific language. Doculects are crucially defined by reference to such a resource. Any successful practical implementation of this organizational structure, therefore, implies that resources are uniquely identifiable. The need for such identifiers is a well-known problem for digital resources in and beyond linguistics, and there do exist proposals for how to manage them (e.g. Broeder et al. 2006). Unique identifiers exist, of course, for many traditionally published sources as well, for example in the form of DOIs or ISBNs.[45] However, these are not sufficient if the goal is to build a system accounting for any attested doculect. One problem is that not even all traditionally published sources have unique identifiers (e.g. older publications before modern identification systems were put in place, or individual articles in edited volumes). Especially when it comes to lesser-studied languages, the most important resources may be unpublished or be part of the gray literature, in which case often no clear identifiers exist. Moreover, in some cases multiple identifiers may exist for the 'same' source, at least from the linguist's perspective—for instance a paperback version of a book may have a different ISBN from a hardcover version—making it necessary to allow for a system in which the use of different identifiers does not result in problems of interpretation about the nature of the documentation underpinning a given languoid.

Finally, we should make clear that any practical implementation of the model we have developed here could usefully build on existing standards. For instance, both the IETF 'best current practice' description BCP 47 (RFC 5646) and ISO 639-4 allow for the description of aspects of doculects and languoids, though neither of these proposals was explicitly designed to do it in the way that we describe it here.[46] Both of these proposals assume that there is one optimal system for the designation of the world's languages, an assumption that is clearly invalid for those conducting research intended to expand our knowledge of the world's language varieties and the relationships that hold among them. In most practical use cases outside of the academic field of linguistics, the IETF and ISO proposals will suffice. However, it is the difficult cases that are of most interest to linguistics and these require a means for structured discussion and explicating dissent. This means has to be built on a different kind of conceptual and technical apparatus, as we have begun to develop in this paper.

[43] This is how the ISO 639-3 codes are used in the OLAC Metadata Standard, for example, where they can be used to indicate a language the resource is about (Simons & Bird 2008).

[44] See also the notion LECTODOC in Hammarström & Nordhoff (2011).

[45] See http://www.doi.org/ for more information on Digital Object Identifiers (DOIs).

[46] See http://tools.ietf.org/rfc/rfc5646.txt for a description of the current IETF BCP 47 recommendations for tags for identifying languages. For the description of ISO 639-4, see http://www.iso.org/iso/language_codes.html and Gillam et al. (2007).

10. CONCLUSION. We believe that the combination of the concepts glossonym, doculect, and languoid offers an appropriate foundation for the rigorous discussion of the otherwise rather vague notion of 'language.' Of course, the proposals here are not intended to argue that the term language should be dispensed with. Rather, in cases in which the term is not explicit enough to conduct research, the precise intention can be specified by a detailed explication of a languoid, whose definition crucially relies upon the notions of glossonym and doculect.

To summarize, a GLOSSONYM is a label (i.e. a string of characters) used as a name for a language (or language-like object). It is not a name in the strict sense of a label that has reference. A glossonym is just the 'form' of the name as used to refer to some kind of language-like entity. The foundational language-like object is a DOCULECT. A doculect is a named linguistic variety as attested in a specific resource. The documentation associated with a doculect is ideally a resource containing data on the language in question, but it can in principle even be a document claiming that a particular language exists without any actual linguistic data given (like in a census or a traveller's diary). Finally, a LANGUOID is a collection of doculects or other languoids, which are claimed to form a group. The most ubiquitous examples of languoids are those that embody a claim that some particular set of doculects all belong to the same language or dialect. Yet, languoids can also be associated with other groupings, like language families, or areal groups, reflecting the fact that, from a linguist's perspective, it is not possible to uniformly establish a basic concept of 'language' somewhere in the middle ground between individual utterances and the full range of worldwide linguistic variation, even if laypeople (and even sometimes linguists) have strong intuitions that this should be possible.

A possibly counterintuitive aspect of our approach is the extent to which we have not tried to ground our notions in the external reality of languages, e.g., by making use of a parameter like the locations in which a given language is spoken, or by considering the nature of its speakers. Of course, we do not deny the importance of such characteristics, but we believe they are not ideal to define languages in a rigorous scholarly way, which requires building, instead, on existing scholarly resources. Of course, to the extent that such information will be included in a resource associated with a given doculect or that it may be associated with ancillary metadata associated with a languoid, it would not all be lost. Such information, we believe, is simply not part of the foundation for more rigorously defining our objects of study.

References

Anderson, Stephen R. 2010. How many languages are there in the world? *LSA Frequently Asked Questions*. http://www.linguisticsociety.org/resource/faq-how-many-languages-are-there-world. (9 October, 2013.)

Bowern, Claire. 2008. *Linguistic fieldwork*. New York, NY: Palgrave Macmillan.

Broeder, Daan, Remco van Veenendaal, David Nathan & Sven Strömqvist. 2006. A grid of language resource repositories. *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*. Los Alamitos, CA: IEEE Computer Society.

Burtch, Shirley. 1983. *Diccionario Huitoto Murui*. Yarinacocha: Instituto Lingüístico de Verano.

Campbell, Lyle. 1997. *American Indian languages: The historical linguistics of Native America*. Oxford: Oxford University Press.

Chiarcos, Christian, Sebastian Nordhoff & Sebastian Hellmann (eds.). 2012. *Linked data in linguistics: Representing and connecting language data and language metadata*. Berlin: Springer.

Cysouw, Michael & Jeff Good. 2007. Towards a comprehensive languoid catalog. Paper presented at the *Language Catalogue Meeting*. Leipzig, Germany. 28 June, 2007. http://wwwstaff.eva.mpg.de/~haspelmt/Cysouw1.pdf. (9 October, 2013.)

Dalby, David. 1999–2000. *The linguasphere register of the world's languages and speech communities* (two volumes). Hebron: Linguasphere Press.

Dobrin, Lise M. & Jeff Good. 2009. Practical language development: Whose mission? *Language* 85(3). 619–629.

Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description*, volume 6, 37–52. London: Hans Rausing Endangered Languages Project.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The World Atlas of Language Structures online*. Munich: Max Planck Digital Library. http://wals.info/. (9 October, 2013.)

Echevarri, Juan Alvaro. 2009. Request for new language code element in ISO 639-3. Change request number 2009-011. http://www.sil.org/iso639-3/chg_detail.asp?id=2009-011. (9 October, 2013.)

Francopoulo, Gil, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet & Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation* 43(1). 57–70.

Francopoulo, Gil & Monte George. 2013. Model description. In Gil Francopoulo (ed.), *LMF: Lexical Markup Framework*, 19–40. London: ISTE Ltd.

Gillam, Lee, Debbie Garside & Chris Cox. 2007. Developments in language codes standards. In George Rehm, Andreas Witt & Lothar Lemnitzer (eds.), *Data structures for linguistic resources and applications*, 147–155. Tübingen: Narr.

Good, Jeff & Calvin Hendryx-Parker. 2006. Modeling contested categorization in linguistic databases. In *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and standards: The state of the art*. Lansing, MI. http://emeld.org/workshop/2006/papers/GoodHendryxParker-Modelling.pdf. (9 October, 2013.)

Hammarström, Harald. 2008. Counting languages in dialect continua using the criterion of mutual intelligibility. *Journal of Quantitative Linguistics* 15(1). 34–45.

Hammarström, Harald. 2010. The status of the least documented language families in the world. *Language Documentation and Conservation* 4. 177–212.

Hammarström, Harald & Sebastian Nordhoff. 2011. Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language* 3(2). 31–43.

Haspelmath, Martin. 2009. Lexical borrowing: Concepts and issues. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 35–54. Berlin: Mouton.

Heaton, Raina, Eve Okura & Lyle Campbell. 2013. The Catalogue of Endangered Languages in context. Paper presented at the 3rd International Conference on Language Documentation and Conservation (ICLDC 3). Hawai'i Imin International Conference

Center, Honolulu, HI. 28 February–3 March, 2013.

Holland, Barbara & Vincent Moulton. 2003. Consensus networks: A method for visualising incompatibilities in collections of trees. In Gary Benson & Roderic Page (eds.), *Algorithms in bioinformatics: Third international workshop*, WABI 2003, 165–176. New York, NY: Springer.

Hosken, Martin. 2006. Lexicon Interchange Format: A description. http://lift-standard.googlecode.com/files/lift_13.pdf. (9 October, 2013.)

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2). 254–267.

Huson, Daniel. H., Tobias Dezulian, Tobias Klopper & Mike A. Steel. 2004. Phylogenetic super-networks from partial trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1(4). 151–158.

ISO. 2007. *Codes for the representation of names of languages—Part 3: Alpha-3 code for comprehensive coverage of languages*. Geneva: ISO.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the world,* sixteenth edition. Dallas, TX: SIL International.

Loukotka, Čestmír. 1968. *Classification of South American Indian languages*. Los Angeles, CA: Latin American Center, UCLA.

Matisoff, James A. 1978. *Variational semantics in Tibeto-Burman*. Philadelphia, PA: Institute for the Study of Human Issues.

Matisoff, James A. 1986. The languages and dialects of Tibeto-Burman: An alphabetic/genetic listing, with some prefatory remarks on ethnonymic and glossonymic complications. In John McCoy & Timothy Light (eds.), *Contributions to Sino-Tibetan studies*, 3–75. Leiden: Brill.

Minor, Eugene E. 1965 Witoto vowel clusters. *International Journal of American Linguistics* 22(2). 131–137.

Minor, Eugene E. & Dorothy A. Minor. 1971. *Vocabulario bilingüe: Huitoto–español y español–huitoto (dialecto mɨnɨca)*. Lomalinda: Editorial Townsend.

Nichols, Johanna & Tandy Warnow. 2008 Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2(5). 760–820.

Nies, Joyce. 1976. *Suplemento a listas comparativas de palabras usuales en idiomas vernáculos de la selva*. Lima: Instituto Lingüístico de Verano.

Nordhoff, Sebastian & Harald Hammarström. 2011. Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In Tomi Kauppinen, Line C. Pouchard & Carsten Keßler (eds.), *Proceedings of the First International Workshop on Linked Science 2011 (LISC2011) in conjunction with the International Semantic Web Conference (ISWC2011)*, 1–6. Aachen: CEUR Workshop Proceedings. http://CEUR-WS.org/Vol-783/paper7.pdf. (9 October, 2013.)

Ringersma, Jacquelijn & Marc Kemps-Snijders. 2007. Creating multimedia dictionaries of endangered languages using LEXUS. In *INTERSPEECH-2007*, 1529–1532. Bonn: ISCA.

Simons, Gary F. 1998. The nature of linguistic data and the requirements of a computing environment for linguistic research. In John Lawler & Helen Aristar Dry (eds.), *Using computers in linguistics: A practical guide*, 10–25. London: Routledge.

Simons, Gary F. 2009. Linguistics as a community activity: The paradox of freedom through standards. In William D. Lewis, Simin Karimi, Heidi Harley & Scott O. Farrar (eds.), *Time and again: Theoretical perspectives on formal linguistics in honor of D. Terence Langendoen*, 235–250. Amsterdam: Benjamins.

Simons, Gary F. & Steven Bird (eds.). 2008. Recommended metadata extensions. Open Language Archives Community Recommendation. http://www.language-archives.org/REC/olac-extensions-20080222.html. (9 October, 2013.)

Steiner, Lydia, Peter F. Stadler & Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change* 1(1). 89–127.

TEI Consortium (eds.). 2013. Dictionaries. TEI P5: *Guidelines for electronic text encoding and interchange* (Version 2.3.0). TEI Consortium. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html. (9 October, 2013.)

Troyer, Duane, Paul Huey & Joseph Mbongue. 1995. A rapid-appraisal survey of Mmen (ALCAM 821) and Aghem dialects (ALCAM 810), Menchum Division, Northwest Province. Yaoundé: SIL Cameroon.

Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology*. 13(1). 77–94.

Whalen, D. H. & Gary F. Simons. 2012. Endangered language families. *Language* 88(1). 155–173.

Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.

Windhouwer, Menzo & Sue Ellen Wright. 2012. Linking to linguistic data categories in ISOcat. In Christian Chiarcos, Sebastian Nordhoff & Sebastian Hellmann (eds.), *Linked data in linguistics: Representing and connecting language data and language metadata,* 99–107. Berlin: Springer.

Wright, Sue Ellen, Marc Kemps-Snijders & Menzo Windhouwer. 2010. The OWL and the ISOcat: Modeling relations in and around the DCR. Paper presented at the *Language Resource and Language Technology Standards Workshop (LREC10-W4)—State of the art, emerging needs, and future developments*. http://www.windhouwer.nl/menzo/professional/papers/Wright_OWL_DCR.pdf. (9 October, 2013.)

Michael Cysouw
cysouw@uni-marburg.de

Jeff Good
jcgood@buffalo.edu