UNIVERSITY OF CALIFORNIA

Santa Barbara

Serialization and Hierarchy: From Data to Corpus in Linguistic Fieldwork

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Linguistics

by

Patrick James Hall

Committee in charge:

Professor Marianne Mithun, Chair

Professor Wallace Chafe

Professor Fermín Moscoso del Prado Martín

June 2014

The thesis of Patrick James Hall is approved.

_____
Fermín Moscoso del Prado Martín

_____
Wallace Chafe

_____
Marianne Mithun, Committee Chair

May 2014

Serialization and Hierarchy: From Data to Corpus in Linguistic Fieldwork

by

Patrick James Hall

ACKNOWLEDGEMENTS

I dedicate this thesis to my family, for their endless encouragement and support, and to my colleagues in linguistics, for bringing me home.

ABSTRACT

Serialization and Hierarchy: From Data to Corpus in Linguistic Fieldwork

by

Patrick James Hall

The structure of digital documentation should empower linguists to search the entirety of a documentary archive. This should be possible even when multiple tools were used to enter the data into the archive. Current practice tends to lead toward fragmenting of archives along the lines of the tools themselves: data entered in one tool cannot generally be searched outside of that tool, and thus data produced by distinct tools cannot be combined and aggregated. The solution to this problem requires an approach which is more general than critique of existing software. In this thesis I elaborate a general but simple way to digitize linguistic data which enables annotation of arbitrary levels of linguistic or annotative detail. A key benefit of this approach is that it allows for documentation which can be searched across levels of language structure, and across varied sorts of linguistic annotation. It is hoped that this simpler, more general design for archival documentation will contribute to the production of linguistic documentation, upon which in principle many sorts of software could operate, and which would be of use to linguists with many and varied interests.

Introduction

The capabilities of current software tools for managing linguistic fieldwork data do not meet the needs of working documentary linguists. Linguists have expectations about the minimum level of linguistic structure which is expected to be represented by a documentary archive. While various tools have impressive capabilities in handling specific subsets of the necessary data (time alignment, or morphological analysis, or annotation of phonetic detail), it is not currently as efficient as possible to investigate the many linguistic phenomena which involve phenomena exhibited across these layers. Searching across our current archives does not return manipulable groups, or does so only for a subset of the linguistic categories which are of interest. the instances of the categories which we maintain as theoretically defensible. Then, we could query across the whole array of levels of annotation, and search, summarize, and aggregate. We could also quantify, and point to result *sets* as uniquely identifiable *elements* of the archive.

I argue that the implementations in existing software do indeed store a subset of the language data and annotations that linguists care about, but that there is no thought of consistency in the design of that data, because our current tools were designed to solve problems other than fieldwork per se.

In language documentation, what is the nature of the content which is actually committed to the archival record? How should that data be structured? How should purely descriptive content — the elements of linguistic structures which constitute recorded data — be intertwined with the linguist's observations about those structures?

The answers to such questions are key to solving the data avalanche facing documentary linguists today. They would guide the design of software which could enable linguists to more efficiently encode their work for maximum utility and expressivity. The current thesis

is not itself a technical specification for software. However, I suggest that the next generation of software for aiding language documentation cannot be produced without resolution of the problem described here.

There is a single, key problem which is not resolved by current software tools for digital linguistic fieldwork. This is the problem of archiving not simply textual strings, but a representation of linguistic data relationships. The recorded linguistic data collected by linguists during fieldwork is hierarchical in nature, but the representations on which existing workflows and archival standards are based do not serialize that hierarchy in a manner which allows linguists to pose arbitrary queries.

The term documentation has acquired at least the following three senses in linguistics:

1. **Interactions between linguists and speakers which result in recordings of linguistic usage.** It is to this process that linguists refer when they speak of *doing* language documentation.
2. **All recorded evidence about how a language is used**. It is in this sense that a linguist might refer to "*the* documentation" of a language, e.g., "the documentation of Hiligaynon." Such references refer to all the usable linguistic evidence to which a theory of that data might refer, regardless of the format in which the data is stored (print, manuscript, recording, etc.).
3. **A persistent representation of linguistic data in a structured, digital form**. The content that is actually encoded into some persistent stored representation.

It is the third aspect, then, that is the focus of the current thesis. This is a timely topic, given the fact that practices in the creation of documentation have been drastically transformed by developments in technology (increased storage capabilities, inexpensive recording devices, and widespread access to powerful computers). The recent upsurge in practice of linguistic documentation has progressed rapidly in preserving and archiving documents which record the languages of the world. This increase is broadening the field's understanding of the variety of grammar, and of the utility of corpora based on a wide variety of data, including naturalistic and elicited content in a wide variety of genres.

Current tools and workflows do not assist linguists as fully as possible in efficiently recording the models described above. Linguistic archives tend to grow by accretion of incompatibly digitized materials, in such a way that it is very difficult to frame and execute queries over the fieldwork archive as a whole. Further, a considerable portion of the knowledge which linguists acquire during fieldwork is only implicitly recorded in the digital archive, and is effectively unretrievable through the sorts of intuitive searches framed in terms of familiar linguistic units. That information is only implicitly stored is problematic not only when querying the archive of a completed fieldwork project, but also for the during the project itself: without a way to search all previously recorded data completely, the linguist must rely exclusively on memory. In the next section, I propose a specific mechanism for resolving the root of these problems, one that elucidates how structured linguistic data is only recoverable from an archive if is truly serialized into that archive in the fist place. !

Serializing Linguistic Relationships

It is perhaps simplest to think of hierarchical structure visually, for instance, as Matryoshka dolls or nested physical boxes. Many fundamental categories in linguistics may of course be understood as being subdivided in this way. Obvious examples of this are the division of utterances or intonation units into words, words into morphemes, and morphemes into phonemes. It is not immediately obvious, however, how to represent such subdivisions in a file, which is to say, somehow capturing those subdivisions in a one-dimensional string. This procedure is not as trivial as it might first appear.

However, some definitions of ‘hierarchy’ are framed in one-dimensional or sequential terms. Pumain (2006) defines a common sense of ‘hierarchy’ as:

"…the organisation of a set into an ordered series of elements where each term is superior to the following according to some normative character. Ranking and classification methods demonstrate hierarchical orders of this type."

While Pumain (intentionally) leaves open the explicit definition of the "normative character," he is clearly implying that it is indeed possible to encode a hierarchy into a one-dimensional (string) representation. In computer science, the encoding of hierarchy into strings is known as *serialization*. The specifics of serialization turn out to be surprisingly relevant to the task of language documentation.

I should pause at this point in the discussion to acknowledge that while the model described here is partially built on a compositional or hierarchical representation of language, it would at best reductionist to suggest that language actually *is* hierarchical in that sense. Indeed, it is precisely the wholesale equation of hierarchical relationships with linguistic relationships which led to the heavy focus on tree structures in generative approaches to syntax — derived expressly from a disinterest in usage as such. It is true that a morpheme, for instance, is not simply made up of phonemes, words are not simply a string of morphemes, and so on. The properties of the constituents cannot be used to predict the properties of a "higher-level" element. It is true that what is interesting about these units is simply that they can be (in some sense) analyzed into decomposed sub-parts at a lower level. Rather, what makes linguistic units at all "levels" interesting is how they are used, what functions they carry out, their histories, and way they may relate to *any* other linguistic unit, whether or not that relationship is a constituitive one. While this explanation makes use of notational descriptions which may seem to be akin to those early presentations of linguistic data, I emphasize again that the model described here addresses only data structures as they are formatted in for the purposes of *documenting* language in a way which meets the needs of complete search.

This point will be developed further below, but in order to move from abstract to concrete, the serialization of a single word will here considered as a point of departure. The most frequent word in Hiligaynon is the article *ang*. The participants in the Hiligaynon field methods class concluded that this form could be understood to contain, at the phonological level, two phonemes: the vowel [a] and the velar nasal. However, for reasons of practicality, the orthographic representation chosen contained three symbols: «a», «n», and «g».1What would be represented in most phonetic notations as [ŋ], then, was symbolized by the orthographic sequence «ng».

This choice reflects in the simplest way the specific sense of hierarchy as a *documentary* practice (and not a statement about linguistic structure as such). There was plentiful evidence throughout our interactions with the speaker to conclude that this the structure of the word contained two phonemes. For instance, that the velar nasal is a unitary phoneme may be shown by normal distributional evidence: it appears initially in /ŋálan/ 'name', medially in /saŋá/ 'branch,' and finally in /damáŋ/ 'spider'. Near-minimal pairs opposing /n/ and /ŋ/ are also common.

But strictly speaking, and this is the crucial point, without some sort of programmatic method which can distinguish «n» from «ng», the occurrences of the velar nasal are not retrievable from the archive. Strictly speaking, one might even say that the fact that «ng» represents a unitary phoneme is not *in* the archive. That this is the case is exemplified by the following task: "Compare the frequencies of the phonemes /n/ and /ŋ/." Because the character n is a prefix of the sequence of characters ng, this task is quite difficult to carry out without resorting to advanced search techniques such as regular expressions. A better

1It was also relevant that the speaker was comfortable with this notation from general practice in Philippine writing systems, and it is of course also a familiar practice in English.

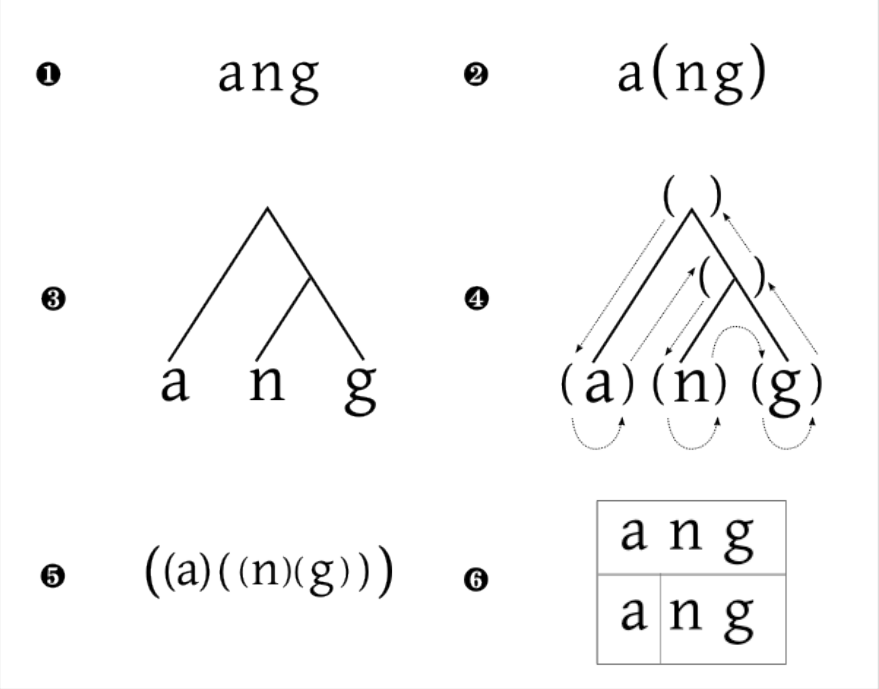
approach is to try to make the structure of the archives map as closely as possible onto the traditional linguistic terminology.

The hierarchy in question, of course, is simply the fact that orthographic «ng» corresponds to a single phoneme, namely [ŋ], so that we assume a correspondence of a(ng) to aŋ.

But in point of fact linguists do this almost by second nature: the sequence *ng* is simply recognized as a group, because that is how the velar nasal is most often spelled in real orthographies. And yet, in a very real sense, the fact that *ng* here represents a phoneme is nowhere serialized in the archive. That this is the case may be shown by the fact that counting phonemes in a textual archive cannot be automated without an abstraction of the set of what does and does not count as a phoneme.

If it is true that the correspondence of the two-letter orthography «n,g» to the unitary phoneme [ŋ] is not explicitly stored in the archive, then what different measures should be taken to ensure that it is explicitly stored, and is explicitly recoverable? Answering this question requires entering into the issue of serialization in further detail.

The Details of Serialization



Six representations of the phonemes of *ang*. (1) Orthographic transcription ; (2) Simple grouping of phonemes; (3) Hierarchical representation of phonemes; (4) Serialization ("Preorder Depth-first Traversal"); (5) Serialized form; (6) "Icicle" or nested chart

A containment hierarchy is a logical structure in which every level and sublevel may be subdivided completely into sublevels. On consideration of the form in (1), it is apparent that there is no explicit representation of hierarchy at all. Nonetheless, linguists are aware of the implied hierarchy, and they are able to analyze the structure of (1) into something akin to the structure in (2) without much thought. Computers, of course, have no such abilities, and thus such intuitions cannot be easily programmed. Rather, explicit serialization of the groupings that are intuitive to trained linguists must be indicated by the linguists themselves. In (3), a hierarchical view of the three characters is expressed visually. Note that in this case, the question of the relationship of «n» and «g» is not simply sidestepped by replacing the «ng» sequence with a single character such as «ŋ». While such a substitution might resolve the

problem in this case, it of course would not solve the more general problem of handling arbitrarily encoded groupings of data in an archive.

Fortunately, algorithms exists for carrying out such a serialization. (Cormen et al 2009:254) In (4), a rather complex visual representation of how such a serialization can be carried out is given. The premise here is that by traversing the hierarchy as mapped out by the sequence of arrows through the characters and abstract nodes in the hierarchy, and sequentially noting either the character in question or an opening or closing parenthesis (usually referred to as a "bracket" in the relevant literature), the serialization in (5) will be produced. The procedure which determines how the arrows traverse the tree is referred to as a "pre-order depth-first traversal." Note also that the serialization proceeds from left to right . That is to say, the left branch of the tree, which dominates only the character «a», is fully traversed before the algorithm proceeds to the right branch, a sub-node which dominates both «n» and «g». It is this sub-node which captures the grouping relationship which is so apparent to the linguist.

Image (6) is a fully equivalent chart of the relationships contained in (5), wherein each "floor" corresponds to an abstract node in the tree (4), or a level of bracket depth in (5). (Such diagrams are sometimes referred to as an "icicle" or "partition chart".)

The import of this rather complex analysis of such a short word is that hierarchy allows us to subdivide a sequence into a shorter set of symbols.

In short:

When storing hierarchical linguistic data, that data should be stored in a bracketed tree structure which represents the data in depth-first preorder.

The diagrammatic form in (6) above is a convenient way to summarize the same process at a higher level: that of the subdivision of a whole utterance, as in the following diagram.

1. *Baw, indi' ko guid malipatan ang mga memories sa CPU.*
2. baw indi' ko guid malipatan ang mga memories sa CPU
3. baw | indi' | ko | guid | malipatan | ang | mga | memories | sa | CPU
4. baw | indi' | ko | guid | ma | lipat | an | ang | mga | memories | sa | CPU
5. b a w | i n d i' | k o | g u i d | m a l i p a t a n | a n g | m a n g a | m i m u r e s | s a | s i p i y u

A diagram of the containment hierarchy of an utterance from the Hiligaynon archive: 1) transcription, 2) normalized transcription; 3) words; 4) morphemes; 5) phonemes. (The utterance means 'Wow, I really can't forget my memories at CPU.'

This diagram could have been equally accurately represented with a tree structure of the same sort as that presented in (4) above. To this point, the presentation is entirely monolingual, since the containment hierarchy in question is modeling only the content.

Translations, glosses, and other sorts of information are to some extent interpretive, and for that reason lie outside of the central "backbone" of the logical model, which is the containment hierarchy of source language transcription and analysis. In this regard, such annotations may be thought of as annotations on the containment hierarchy, rather than as being hierarchically ordered content with respect to each other. Said differently, annotations are related to each other only insofar as they are linked to elements in the containment hierarchy.

Language Data and Hierarchy

That such levels of analysis can be useful in linguistic description is hardly controversial, of course. That language can be described (in part) in terms of these compositional levels has been part and parcel of documentary linguists for at least a century (Lehmann 2004 cites Finck 1909, for instance, as the earliest instance of a modern style in glossing full texts, where full texts were given morpheme-level analyses throughout). When

working documentary linguists make use of the set of levels described above, they are also implicitly making use of the fact these that levels may be usefully modeled through the use of a containment hierarchy: linguistic elements at one level may be decomposed into lower levels, or assembled into higher levels.

And is this mismatch between the linguist's mental model and what gets committed to the archive which is where a lack of serialization causes damage to the final record. When a representation of usage, such as a transcription of a stream of speech, is converted into an archival form — most often a file — that the current digital array varieties models begin to cause fragmentation.

At this point it is worth questioning whether the amount of information encoded in the hierarchical representation is truly sufficient to capture the most important aspects of language structure. One simple argument that it is the simple existence of a long history of language documentation in print. While print sources do not of course serialize the hierarchy of language structure in a way which allows for complete search or search experiments (about which see below), the structure in which is ? implicit in linguistic annotation (interlinear glossing, parallel text, etc.) has proven to be a rich source for the comparative analysis of language structure. As shown above, the hierarchy of linguistic data may be serialized into a form from which that hierarchy can be reassembled.

Climbing up this stack from lower to higher levels, the difficulty of ordering or disentangling which level is at work on a given construction becomes more difficult: prosody, syntax, and discourse, for instance, interact strongly. The cognitive reality of these hierarchical structures may be questioned. Indeed, Bybee and McClelland (2005) strongly contest the assertion that a hierarchically designed corpus can capture the essential characteristics of the structure of language:

"…there is no analysis into units at any level or set of levels that will ever successfully and completely capture the realities of synchronic structure or provide a framework in which to capture language change." (Bybee and McClelland 2005)

Taken at face value, this claim would seem to negate the utility of fieldwork at all: if there is no concept of sublevels in language, then a database seems worthless (as would time-tested categories such as "word" and "morpheme.").

However, there is a grain of truth to the contention that analysis of linguistic structures is never beyond doubt, particularly in the validity of such analysis in diachronic terms. But a crucial aspect of fieldwork is that it records a synchronic moment in usage. Given this snapshot nature of recorded usage, the basic familiar linguistic units (phonemes, morphemes, words, and higher-level constructions) are sufficiently expressive and specific to capture usage in documentation. It is not clear how Bybee or McClelland would expect documentary fieldwork to be carried out at all, without the use of an analytical framework which included at least words, morphemes, and phonemes as described here.

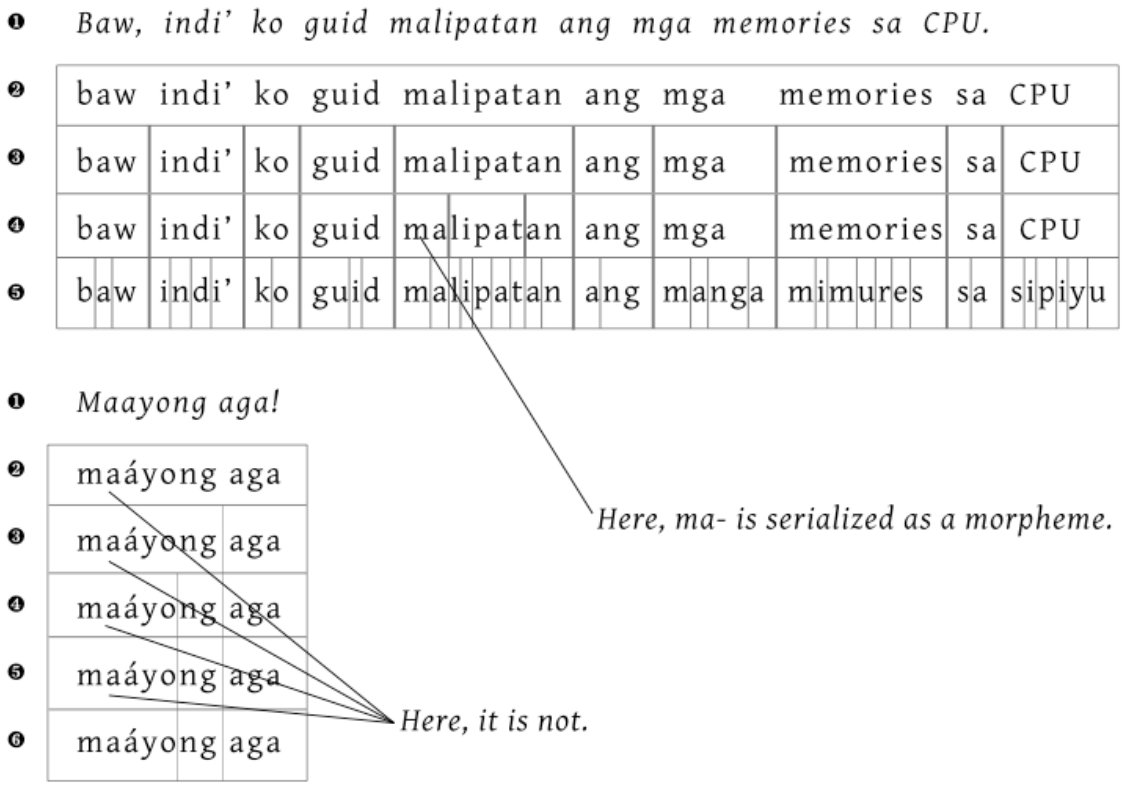
Again, it is not suggested that the boundaries between these levels do not blur, or that they alone suffice to fully describe usage. To the contrary, it is precisely the blurring which occurs between linguistic categories which is the locus for grammaticalization — a major force in language change.

What these levels of description do describe is a workable set of subdivisions of the speech stream, which, in a correctly serialized form can serve to capture those aspects of speech from which grammatical analysis is constructed, and to do so in a way which allows for the recovery of that structure from the record. Given that the pragmatic use of these levels of linguistic analysis is in fact part of common practice in usage-oriented fieldwork, it seems reasonable to suggest that they constitute what might be called a "working model" of linguistic structure — that is to say, as linguists record usage during fieldwork, they

constantly make references to these levels of analysis, and aim to capture them in the archive.

Complete Search

The important point about this view of linguistic data is that it enables *complete search* over the archive. Consider the following tiny "corpus," which consists of the union of the previously analyzed utterance together with another short utterance:



The hierarchical component of a corpus containing the previous utterance plus another, which means: 'Good morning!'

This sample corpus contains two utterances (represented here in a slightly normalized orthography which disregards case and punctuation):

1. "baw, indi' ko guid malipatan ang mga memories sa cpu."

2. “maayong aga”

Now, if one considers a typical query which might be carried out over this corpus:

Query: “Find all instances of a prefix with the form *ma-*.”

In an unstructured corpus (text transcribed into a text file without indication of source language and metalanguage, for instance) such a query can only be carried out as a *string seach* over the full transcription. Consequently, the prefix ma- will be found in both of the example sentences, once on the word *malipatan* and once on the word *maayong*.

However, the linguistic status of the prefix ma- in both of these words is assigned by the linguist’s analysis of the speaker’s intuitions. In this case, the linguist (this author) interpreted the speaker’s understanding of the word maayong as having just two morphemes. This contradicts earlier description of the language, such as the lexicography in Kaufmann (1935), That work analyzes *maayong* as having three morphemes, *ma-ayo-ng*. The “correct” analysis of this term will not be resolved here, but the point is that a linguist’s evaluation (defensible in the grander scheme of research or not) should be unambiguously encoded in the record.

If there were no serialized representation of the prefix ma- as a prefix then that analysis is not recoverable to future users of the corpus. Logically speaking, this problem is identical to the one posed earlier about the serialization of the velar nasal /ŋ/ as «ng»: the linguist’s knowledge about the nature of this linguistic unit is not captured *explicitly* in the archive..

Thus, the importance of hierarchical serialization is apparent at both the phonemic and morphological levels.

In a corpus where the degree of granularity of serialization of the entire hierarchy in a corpus is known, for instance, where every utterance is verified to have structured analysis of words, morphemes, and phonemes, then a complete search over that corpus can be said to truly constitute a search experiment.

In Hiligaynon, a series of perhaps a dozen verbal prefixes are implicated in a range of effects on various verbal categories. These include: *ga-*, *guin-*, *guina-*, *i-*, *ipa-*, *ka-*, *ma-*, *mag-*, *maka-*, *na-*, *nag-*, *naga-*, *pa-*, *pag-*. Such prefixes interact with tense, aspect, transitivity, voice, valence, and mood (and probably other categories). The motivating factors behind the choice of a particular prefix in any given instance, then, are clearly complex, and must be described with plentiful exemplification.

A linguist who is unfamiliar with Hiligaynon would be justified in asking whether this set of prefixes really should be thought of as a set in any meaningful sense. One response to such a query would simply be to point out that all six linguists who participated in the project agreed that the prefixes constitute a class. But strictly speaking, of course, that is no argument — or rather, it is an argument from authority. Additionally, it's not a reproducible without other linguists repeating an in-depth exposure to the language which would mirror the initial project. It is unrealistic to expect that other linguists — be they Austronesianists, typologists, Philippinists, etc. — could

If the point of view of linguists who did not participate in the project is assumed, interesting constraints on "relevance" emerge. Is it the case that some of these prefixes are far more common than others? Such an assumption seems likely for any such class of morphemes. A simple comparison in frequencies might be a starting point: in other words, can we count them?

The answer, in an archive where linguistic data is stored without structured relationships in a recoverable way, is "No." There are several reasons why this is the case. A complete search over a morpheme take *pa-* as an example, is quite simply impossible in an archive without a hierarchical model which extends to the linguistic level of the morpheme. Why this is so is apparent enough: *pa-* is not only the prefix of many Hiligaynon words, it

happens (coincidentally, not by historical relationship) to also be a prefix of another prefix. So, any search for the string serialized as *p*, a will fail on two counts: (1) it will return words without affixal morphology such as *pamilia* 'family' (a monomorphemic noun borrowed from Spanish *familia*, with once-regular sound change f > p) (2) it will return words which are prefixed with other prefixes which simply happen to begin with the sequence «p, a»: such as the prefix *pag*-, such as *pagká'on*, 'to feed'.

The Fragmented Archive

It is the inconsistency between the linguist's working model of linguistic relationships and the software's semi-structured representation of those relationships in the digital archive which is at the root of what might be called "fragmented" archives, where the data in disparate physical records (files) are incompatible with each other, and thus not amenable to what will be described below as "search experiments" — queries over all recorded data in the fieldwork archive. Incompatible data from distinct software packages cannot be searched simultaneously. However, linguists know that a valuable annotation or observation, once forgotten, is lost. Thus, if a given piece of software does not support a particular kind of annotation at a particular point in the fieldwork process, they will resort to ad-hoc annotation in an unstructured format.

During the Hiligaynon project, an interesting workflow emerged in group meetings, where recordings made by the speaker were collaboratively transcribed by the whole class. Inevitably, a thenceforth unencountered grammatical phenomenon or construction emerged that would catch the attention of the group. The construction in question may have been a single fleeting instance, but between the process of hearing the construction and working through an understanding and gloss of the construction with the speaker, the group would decide that the phenomenon was worth further exemplification. At such points, the focus

turned from transcription of a recording to elicitation of variants of the content of the newly discovered construction. It was at such moments that a revealing habit arose: the group's appointed scribe would usually switch out of ELAN and into a word processing program.

This shift in software is directly traceable to the fact that ELAN has no way to add annotations to linguistic units. When the ensuing elicited examples are serialized to the word processing files, their link to the original occurrence of the construction in natural speech is lost. While it is possible to add distinct tiers explicitly for notes to an ELAN transcription (this was successfully done in the Hiligaynon project), there is little support in such an arrangement for longer commentary on the transcribed material. This is in part a question of user interface, but it is also reflective of the fact that ELAN does not consistently serialize the distinction between linguistic content and annotation.

A More Extensive Example

Having discussed the general notion of fragmentation in archives, I proceed to a more extensive investigation of an actual research question carried out on the Hiligaynon documentation.2

In the wake of the previous discussion of the status of the Hiligaynon verbal prefixes as a "set" of some kind, I ask a more specific, question: how can the usages of *two* such prefixes be usefully contrasted, taking into account a complete search of the entire Hiligaynon corpus? The prefixes considered will be *na-* and *guin-*. That the distinction between these two prefixes be explained was suggested by the speaker himself as a test for

2In fact, the investigation is carried out on a subset of the archive, as : the data considered in this section includes texts extracted from the time-aligned content produced with the ELAN time-aligned transcription package. Even so,

the linguist.3 He was aware that the meanings of these prefixes overlapped somewhat, and was curious to see if the linguists could reasonably explain them. The current discussion, then, traces out an attempt to track the evidence from within the archive which might serve as evidence in establishing the grammatical and semantic differences between these two forms. The goal here is to show just how difficult it is to query an archive with nothing more than string searches over undifferentiated linguistic material. As I shall attempt to show below, without a serialization of data which makes specific levels of linguistic granularity addressable, searching is both inefficient and inaccurate. (Which is of course not to say that it is impossible.)

A first point to note is that the forms of both *na* and *guin* (occasionally spelled *gin* in the archive) are also parts of other, perhaps equally common prefixes: *nag*, *naga*, and *guina* or *guiná* are all also frequent members of the verbal prefix class. Additionally, there are numerous roots which begin with *na*: *na'* (with the apostrophe representing a glottal stop) is a common deictic pronoun, the adjective *namí* 'delicious', the pronominal form *namon*, *nanai* 'mother', and so forth.

Such ambiguity, of exactly the sort described above for *ma* as a prefix or as part of a root, may be inordinately slanted toward the inappropriate, partial match. And in fact, in the data considered here, roughly half of the instances of *guin-* are *not* false positives for the other prefix, *guiná*. Even in this moderately sized subset of a medium-sized archive, there

---

3The speaker's question ran:

ANG PAMANGKOT....ANO ANG DEPRENSYA SANG: (1) ***guin*** *obra nila (2)* ***na*** *obra nila. ano IMO sabat sa pamangkot?*

Which means:

"The question: what is the difference between *guin obra nila* and *na obra nila*? What is YOUR answer to the question?"

are more than four times as many false positives as relevant results. (In the terminology of information retrieval (Manning *et al* 2008:5), the *recall* in this is 100%, but the accuracy is a quarter of that).

frequency	annotation
43	*guin* and *guina*
22	*guin* (but not guiná)

Results for a search for guin-.

This represents an obvious loss in productivity for the linguist. Consider that these half of the results are irrelevant, and the only means of distinguishing false positives from true hits is to read them individually.

The accuracy for a query over *na-* is not carried out here, but considering the likelihood of such a short sequence, and the fact that it occurs in the common words listed above, it is likely to be considerably worse.

Note that at this point the original goal, to juxtapose a representative sample of the usages of *guin-* versus *na-*, has still not been approached.

There are numerous ways in which a hierarchically serialized documentary format could have alleviated these inefficiencies.

Declarative Data, Imperative Software

To this point, the discussion has described a theoretical representation of an optimally structured archive. But documentary linguists currently make use of several software packages — perhaps on the order of a half-dozen major specific applications. Reviews of recent literature (e.g., Thieberger and Berez (2011), Austin (2006)) tend to focus on the use

and troubleshooting of specific software packages, rather than engage in analysis of the ways in which the sorts of data that linguists normally make use of should determine the functionality of software.

Rather than engage in a detailed critique of individual software packages, it is helpful to think of groupings of commonly used software, so that the goals and capabilities of each may help to show where they fit in to the larger goals of digital linguistic fieldwork. The software which is currently in use by documentary linguists may be classified by the sorts of data that the software is designed to handle.

Analyzing the goals of existing software, and the ways in which those applications are made to interact (even when such interactions were not planned by the designers) may help elucidate junctures in current workflow where fractures in the underlying data model have costly consequences for the processes of linguistic documentation and linguistic description.

The definition of hierarchy in both the contexts of both linguistic structure and serialization is strictly constrained to a representation where every subunit is a child of a single parent at a level.

Consider the following very simple Hiligaynon sentence, (chosen because it has just a few subdivisions at each level of the documentary hierarchy):

Maayong aga. 'Good morning.'

The following is a possible representation of this sentence in a Toolbox-style annotation (using s: sentence, m: morphology, g: gloss, t: translation):

```
s)  Maayong aga.
m)  ma-áyoŋ ága
g)  IRR-good morning
t)  Good morning.
```

Note that the form maayong is probably historically *maayo-ng*, if not *ma-ayo-ng* 'IRR-good-LINKER', although the speaker seemed to interpret the form as monomorphemic. At first glance, this representation appears almost identical to the "icicle chart" above. But it

differs from that representation in a crucial feature: this Toolbox file is a serialization format. In other words, the data are to be interpreted as a string. But despite the fact that each line appears to have the same structure, it is clear that the Toolbox software does not treat these lines as such, because Toolbox does provide such functionality as complete search over morphemes and words and autocompletion of word-level annotations.

Thus, it is clear that whatever hierarchical model is built into Toolbox, that model is not embedded in the file format. In this sense, it may be said that the data serialization format used by Toolbox is implicitly hierarchical, whereas a bracketed tree notation of the sort described above is a declarative file format.

Further evidence that the linguistic hierarchy is not serialized by the Toolbox approach is shown by considering what would be required in order to answer the question "What is the Hiligaynon word for 'morning'?" In this serialization, each line corresponds to a linguistic level. (For this reason the serialization may be correctly referred to as "interlinear.") But logically speaking, the fact that ma-áyoŋ and the Leipzig-style morphological gloss IRR-good are related to each other is not serialized in the file. A linguist reading this notation will realize immediately that that is what is intended. But in order to make this relationship "understood" by a machine, the one of two steps must be followed:

Intimations of this characteristic of the Toolbox format are faint but detectable in the literature on language documentation, as in this prescient observation in Austin (2006):

> "The Shoebox/Toolbox tool automatically creates the appearance of vertical alignment in its interlinear text function, though it actually stores spaces in the data files to do so. Note that it does not store the relationships between the aligned information and rather relies on the user's implicit knowledge to interpret these." (Austin 2006:112fn3, emphasis in original)

Effectively this observation is in agreement with the idea that Toolbox does not serialize linguistic hierarchy in the sense described here.

The process of language documentation confronts the conflicting values of completeness and accuracy. On the one hand, linguists strive to create as varied and representative a record of the language as possible, while on the other, they strive to ensure that the content of that record should be as free from error and misrepresentation as possible. As there is not sufficient time during the actual work with speakers for linguists to delve into fully elaborated statements on the theoretical status of every morpheme of every token which is encountered, as it is encountered, documentary linguists must be pragmatic in the degree to which they represent theoretically fraught linguistic relationships. And in fact documentary linguists do exercise pragmatism when it comes to conceptualizing the structure of language in use.

Glossing and Annotations on the Hierarchy

Obviously the model is incomplete without a specification of how to add explanatory content. Considering the content of documentatary data from the viewpoint of hierarchical structure, all annotations may be considered to have the same structure: annotations on the hierarchy may be defined as strings with an associated type label, represented as children of any node in the hierachy which contains linguistic transcription.

To investigate this process, we will consider a single intonation unit. The following extract is somewhat complicated in detail, and it is thus instructive as a point of reference for evaluating hierarchical representation. This particular example is chosen precisely because it is so typical of real fieldwork data: it ends with a hesitation, and it so happens that that hesitation is code-switched into English.

We may begin by unpacking the data types implied by standard interlinear gloss notation:

malakat ako may—,
ma-lakat ako may—
IRREALIS-walk 1.SG.ABS may-
'then I'll walk'

[hil034:772.30-773.80]

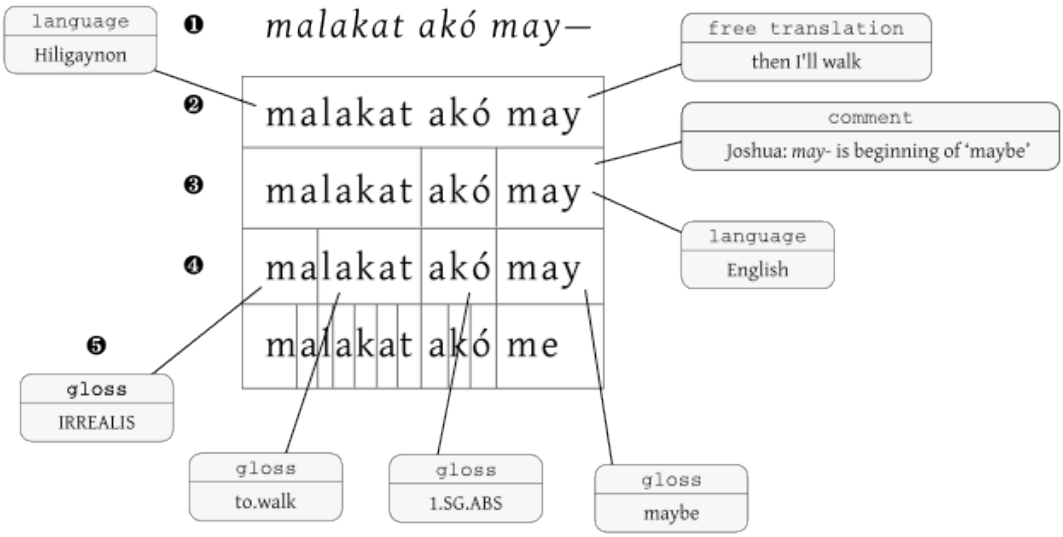
These elements constitute the containment hierarchy component of the representation (that is, all of the Hiligaynon material). This content may be taken as a whole in opposition to aribtrary annotative material, be that data metadata (timestamps, file references…), speaker's meta-commentary on the grammar of a sentence, the linguist's grammatical observations, etc.:

hierarchy	annotation
malakat ako may—,	*free translation*: 'then I'll walk'
ma-lakat, ako, may—	*fileID*: hil034.wav
ma-lakat ako may—	*start*: 772.30
	stop: 773.80
	comment: *may* is beginning of maybe
	gloss: IRREALIS-walk 1.SG.ABS may-

The free translation 'then I'll walk' is *not* an annotation of the sequence of words *malakat*, *ako*, *may-*. It is rather an annotation of the utterance "*Malakat ako may-*" as a unity.

The translation annotates the *whole* utterance: neither the subdivisions of the whole utterance, nor its analyzed components as such.

This seemingly inane distinction plays a key role in the incorporation of annotation into the linguistic structure.



A hierarchical data structure with annotations

Note that the selection of annotations visible in this diagram (language, gloss, comment, and free translation) are not prescribed by the model, although certain annotations are almost always present (particularly gloss and free translation). Note also that annotations which are attached at lower levels of hierarchy may override nnotations at higher levels of annotation. Thus, the English word maybe, (uttered here in hesitation) is marked Language: English at the word level, overriding the value Language: Hiligaynon at the sentence level. (This "cascading" behavior on particular annotation keys should be adjustable by the annotator; a Comment annotation on a morpheme, for instance, should not override a Comment on its parent word.)

This approach to annotation has another benefit: it is possible to annotate incompletely. In the Toolbox serialization below, an incompletely transcribed sentence, will break the software's ability to correctly map words to glosses:

```
s) ako ya gin atinán ko ya CPU.
m) ako  ya  gin atin-án ’ko ya  CPU
g) 1.SG.ABS EMPH  PERF 1.SG.ABS Central.Philippine.University
t) As for me, I attended Central Philippine University.
```

It is quite difficult to determine visually that the second instance of ya in this Toolbox serialization is not glossed. The abstract method of annotating described above avoids such mismatches by "anchoring" annotations to a particular node in the hierarchy. It might be made more visually apparent to a human by vertically aligning words and glosses, but such reading aids are irrelevant to the computer. This very fact is what Austin (2006) is referring to by the "appearance" of vertical alignment.

The Hiligaynon Documentation Project

Exemplification of these challenges will be taken from a recently produced fieldwork archive. As part of an intensive Field Methods project at UCSB, the Hiligaynon language was studied and described by a team of six linguists. Data were recorded with both ELAN time-aligned transcription software and the Toolbox package for morphological analysis. Additionally, a significant proportion of notes were recorded in semi-structured plain text or word processing documents. In all, several hours of time-aligned, transcribed content were produced. A corpus designed along the lines specified in this thesis is currently in development.

The Hiligaynon archive in its current instantiation is not constructed around a containment hierarchy in the sense outlined here. Consequently, while the archive was produced with current practices, it is still to some extent "fragmented" in the sense that its contents do not currently constitute a structured record of the evidence about the language

gathered during the course. This is in no sense due to the participating linguists' understanding of the structure of the Hiligaynon language, which grew to considerable extent, as did the archive itself. As of the time of writing, the transcribed portion of the material in the archive contains some containing over 20,000 words.

There were seven participants in the class: Marianne Mithun, instructor; Joshua De Leon, speaker; and five graduate students in linguistics (including the author). Members of Joshua's family and his circle of friends were also included in the documentary record through recordings that Joshua produced in his own home. Additionally, some external materials were sourced from the internet, including materials from social networking sites and content from home-made video blogs.

While most *in situ* linguistic fieldwork projects may involve fewer linguists and more participants from the speech community, this particular project is nonetheless fairly representative of the sorts of data collected in most usage-based approaches to documentation. In fact, it may be viewed as being representative of a longer documentation project, given the fact that so many participants were working in parallel. This is important because many of the most difficult challenges for language documentation emerge only when a critical amount of content is acquired.

Constraints on Detail in Fieldwork

The realities of fieldwork demand recognition of temporal and economic constraints: we only have so much time and resources to dedicate to working with sepakers. As a result, the linguist must aim to balace coverage of the linguistic levels. Instrumental phonetic analyssis, for instance, is usually relegated to later stages than the enumeration of basic phonemic oppositions. This is a workable approach in part because of existing well-developed phonetic terminology. Simplifying rather drastically, the table of consonants and the vowel

chart of the IPA (or another, equally useful system such Americanist phonetic notation) encompass a closed set of terminology, which provides a complete enough set of oppositions to capture all phonemic systems with a language. The utility of this conventional terminology is not that it provides a completely unambiguous representation of all sounds in all recorded utterances, but rather, that the terminology helps to set up a set of oppositions in a set of sounds. This line of thinking is very much in line with structuralist approaches to linguistic description: linguistic terminology was developed because it has been shown to succeed in capturing those oppositions in language which are used by speakers to distinguish constructions of all types.

The defininining feature of the pragmatic model of linguistic structure is compositionality. While linguists certainly disagree about the best way to understand constituency relationships of language at high levels, during the process of fieldwork (as mentioned above), pragmatic constraints demand that a few truly compositional categories be treated as axiomatic. Said differently, an archive which did not contain at least a representative sample of the most important elements at the traditional levels of analysis (syntactic, lexical, morphological, phonological, phonetic) would be unlikely to be recognized as a linguistic archive.

Interestingly, the archival representation described here offers benefits in reuse in areas generally held to be unrelated. For instance, applications in both quantitative analysis and language revitalization can be enabled by hierarchical serialization.

Evaluating the Approach

Does the model described here suggest pathways toward implementable software for doing documentation?

The serialized representation of hiererchical linguistic data described here, while useful for capturing hierarchy and interrelationships among linguistic units, as well as being more directly legible than current data formats, will nonentheless not be edited directly. Rather, the lessons learned from the design, testing, and implementation of user interfaces for existing software packages could be reused, and, more importantly, combined into a unified interface which stores its data in a hierarchical way. Such an interface could allow for arbitrary annotation in the sense described above, in a way that avoids the problems of fragmented archives.

It might be countered that a compositional model of the content collected in linguistic fieldwork is irrelevant, or retrograde, or too evocative of disproven approaches like strict structuralism. It might even be advanced that the current discussion is a matter of software development rather than of linguistics proper. It is, rather, a model of linguistic data itself, of the sort which is produced and used during linguistic fieldwork and the grammar writing. I suggest that while linguists may hold contradictory grammatical theories, even the most theoretically (or methodologically) opposed linguists will share at least basic a outline of the shape of the linguistic data which should be documented in fieldwork.

To be precise, assuming that linguists take as axiomatic the belief that recordings of linguistic usage are to be valued in theories of language, then the theories built on data collected from fieldwork may be built on structures which are more or less compositional in nature, but even so, documented instances of language use — the linguistic data — which are to serve as the evidence for theories at any point on this cline will necessarily make reference to data which is compositional.

Conclusions

The way in which linguistic content and relationships are encoded in the documentary archive may constrain analysis. Recommendations have been made on the basis of this evalution for a more productive model of linguistic fieldwork data. These recommendations also serve as an initial point of discussion for a a specification of a more software application for linguistic fieldwork which would map more closely onto linguistic documentary structures as they are used and have been used by linguists. The linguistic structures which have been visually encoded in annotative devices such as the “interlinear” gloss themselves constitute a sort of “technology” of language documentation.

However, this progress is primarily being represented digitally at the level of the document. All finer levels of linguistic structure — constructions, turn-taking, morphology — are only being committed to archives implicitly, necessitating the inaccurate forms of string searches whose drawbacks and inefficiencies are described above..

Consequently, they are not accumulating into a form which can constitute a true database. Ironically, it is the very ease with which such technology may be used to create documentation that is source of some of the biggest challenges for modern linguistics: as data quantities increase, the task of transferring those data into a meaningfully searchable form becomes ever more difficult, not less so.

The functionalist commitment to usage, and more generally to spoken language, requires that claims about the grammar of a language be supported by a non-trivial amount of evidence drawn from real usage — a corpus. But as has been shown here, a corpus is only the beginning — if that corpus does not allow the linguist to refer to the same classes of linguistic entities and relationships between entities that are part and parcel of linguistic

theory, then the corpus is not fulfilling its requirements as a tool for carrying out linguistic analysis. As archives grow in size, such problems will only worsen.

The frustrations that field linguists face in dealing with content of their work can only be overcome by facing the hard question: "What is the best, most extensible, most durable way in which to store our descriptive material in a computer?"

References

Austin, P. K. 2006. Data and language documentation. Essentials of language documentation. 87–112.

Austin, P. K. 2013. Language Documentation in the 21st Century. http://www.slideshare.net/pkaustin/language-documentation-in-the-21st-century (28 April, 2014).

Berez, Andrea L. & Nicholas Thieberger. 2011. Linguistic data management. The Oxford Handbook of Linguistic Fieldwork. 90–118.

Bird, S. & M. Liberman. 2001. A formal framework for linguistic annotation. Speech communication 33(1-2). 23–60.

Bird, S. & G. Simons. 2003. Seven dimensions of portability for language documentation and description. Language 79(3). 557–582.

Bybee, Joan & James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. The Linguistic Review 22(2-4). 381–410. (29 March, 2014).

Cormen, Thomas H. 2009. Introduction to Algorithms. MIT Press.

Finck, Franz Nikolaus. 1909. Die Haupttypen des Sprachbaus,. Leipzig: B.G. Teubner.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for. Essentials of language documentation 178. 1.

Kaufmann, John. 1935. Visayan-English Dictionary: Kapulúñgan Binisayá-Ininglís. La Editorial.

Lehmann, Christian. 2004. Directions for interlinear morphemic translations. Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung, vol. 2, 1834–1857. (Handbücher Der Sprach- Und Kommunikationswissenschaft 17). Berlin: W. de Gruyter. http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Publ/IMG.PDF.

Manning, C.D., P. Raghavan & H. Schutze. 2008. Introduction to information retrieval. . Vol. 1. Cambridge University Press Cambridge.

Nordhoff, Sebastian. Electronic Reference Grammars for Typology: Challenges and Solutions. Article. http://scholarspace.manoa.hawaii.edu/handle/10125/4352 (24 September, 2010).

Pumain, Denise. 2006. Hierarchy in natural and social sciences. Vol. 3. Springer. http://link.springer.com.proxy.library.ucsb.edu:2048/content/pdf/10.1007/1-4020-4127-6.pdf (28 April, 2014).

Robinson, Laura. 2006. Archiving directly from the field. Sustainable data from digital fieldwork 2005. 23–32. (8 March, 2014).

Thieberger, Nicholas. 2011. The Oxford Handbook of Linguistic Fieldwork. Oxford University Press.