Rapid word collection and historical linguistics

Ken Manson

4 September 2013

Introduction

With the rapid decline in the number of languages and language families, linguists have increased their documentation efforts to gather as much language data as possible before a language's demise. Himmelmann's description of language documentation as providing "a comprehensive record of the linguistics practices characteristic of a given speech community" (1998 §3.1) focuses primarily on natural texts, however, linguistic knowledge is not only found in communicative events but also in the words that compose these texts.

Issues in historical analysis of minority languages

There are a number of issues facing historical linguists working in lesser studied language families.

1. Written records
2. Wordlist comparison
3. Corpus size
4. No a prior cut-off in the historical-comparative method
5. Language loss

Written records

Written records have been foundational for historical-comparative research. However, over 90% of all languages do not have a writing tradition that is more than 3-5 generations old. This means that historical linguists are limited to present-day materials and this may lead to not fully reconstructing the protolanguage. For example, reconstructing Proto-Romance without the help of Latin fails to reconstruct all the nominal cases that are evident in Classical Latin.

Wordlist comparison

Historical-comparative textbooks imply that the basis for analysis is the wordlist. By collecting a list of words regular sound patterns across the languages being compared are identified. The unanswered question is "how are these lists compiled?" Is it from translation of a standardized wordlist, from a text corpus, from a speaker's intuitions?

By using the wordlist approach, regular sound patterns can be missed. For example, English *dog* and German *hund* are the most frequent/common word for "dog". Yet there is the English *hound* which is cognate to the German, and this is missed using a translated wordlist.

Corpus size

Very few languages have corpus of over 2,000 words. Many languages are poorly documented and have only 100-200 words, typically a Swadesh-like word list. To be able to have a text collection that is representative of a language there needs to be at least a 1 million word text collection. The Brown Corpus of American English (approximately 1 million words) contains roughly 16,000 unique words. A vocabulary of only 1,000 words covers 72% of the corpus.

1 million words is somewhere between 80 and 170 hours of spoken text, assuming a speaker produces 100-200 words per minute.

Boerger (2011: 228), based on actual field research, notes that it takes 16 hours to orally transcribe and translate one hour of recording. This is much quicker than written transcription and translation. A 1 million word text collection would require somewhere between 40 and 90 weeks to orally transcribe/translate, and even longer to transcribe orthographically. Although, Newell (1995:43) suggests that "one trained text gatherer working also as a computer keyboarder can collect, keyboard, and do a spelling edit on about one million words of text in one year." But this also is a huge investment of time.

Collecting enough textual material and commenting/annotating/glossing to be useful for semanticists and historical linguists is a prohibitive undertaking for an individual, and even more so when there are several languages needing to be compared which have no or limited records.

No a prior cut off

How many words or sound changes prove a relationship? And when there are conflicting isoglosses which ones are the more significant? The answer seems to be "we don't know/can't tell, but more is better." This implies that quantity is a substitute for quality. What constitutes a significant sound correspondence?

Brown et al (2012) have identified frequently occurring sound correspondences across the languages of the world. This study was based on a 40-word list, so more research needs to be done to identify less frequent patterns and areal distributions.

Language loss

Language loss is apparently increasing. Whalen & Simons (2012) identify 50 language families/stocks that have become extinct since 1950. And of the current 372 language families/stocks, 102 are moribund (Whalen & Simons 2012) and all of these stocks have inadequate archived materials. Looking lower down at the language level 25% of languages are moribund.

Even for language stocks that are considered to be safe from extinction, languages and branches within the stock are endangered.

There is no way of identifying languages that are currently viable that will become moribund in 2-3 generations. But a key indicator is that the smaller the language the greater chance of becoming moribund. If a language has less than 100 speakers it is guaranteed to be moribund (currently 443 languages), of languages with less than 1,000 speakers ⅓ are moribund (380 languages) and with less than 10,000 speakers ¼ are moribund (473 languages).

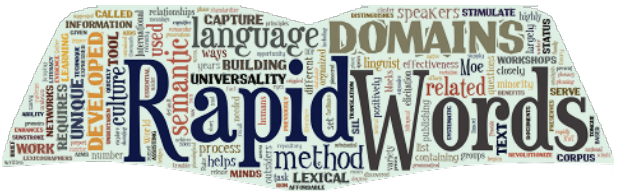
Refocusing language documentation

The emphasis on naturally occurring speech in language documentation to "provide a comprehensive record of linguistic practices" (Himmelmann 1998) leads to a skewed archive of language data. Significant linguistic knowledge is stored in the mental lexicon of speakers. Capturing this knowledge would significantly extend the usefulness of a language's archive. Not only for translation of recorded materials, but also for researchers focusing on the mental lexicon.

By incorporating the collection of the mental lexicon to the current practice of natural speech a more comprehensive archive can be developed.

The question is how do you collect all the words in a language, or at least a significant proportion of them?

Rapid word collection – a possible solution



The aims of a rapid word collection workshop are to help native speakers in their roles as writers and translators, to serve as a starting point for a dictionary and to provide a resource for those interested in the language.

Why?

1. A one million word text collection is the minimal collection to have a comprehensive list of unique words
2. Cannot transcribe/interlinearise all the recordings for a language
3. Meaning is not only associated with texts, but also words
4. Identifies the relationships between words
5. Provides a "wow factor" for the community
6. Validates the language and community
7. Synergy of working together
8. Community oriented, rather than linguist oriented
9. Produces a dictionary/lexicon **and** a thesaurus
10. Task oriented
11. Helps with standardising of the language
12. Aids literacy
13. Aids translation between languages
14. Helps with the analysis of texts
15. Resource for people interested in the language

How?

> By using community members' inherent semantic maps of how words are linked, linguists can leverage this ability into collecting, glossing and printing over 10,000 words within one month using a group workshop approach.

Words in the mental lexicon aren't listed alphabetically like in a published dictionary. Instead the brain organizes them into a multi-dimensional web of relationships that it uses when constructing sentences and conversations. This mental web is incredibly complex and powerful, clustering around key concepts, 'semantic domains'.

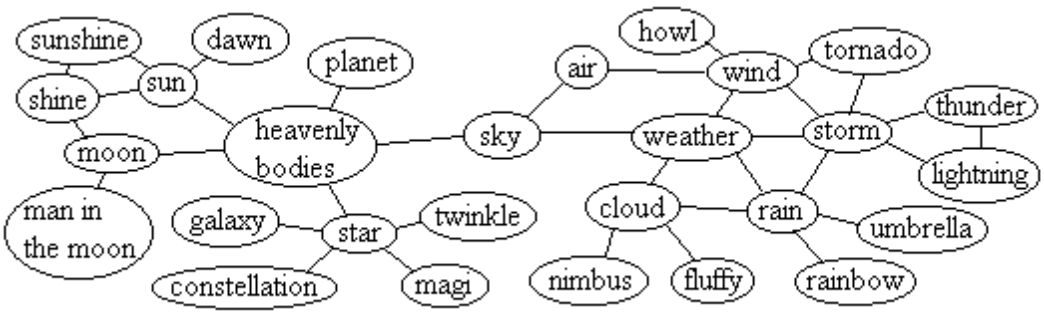


Figure 1: The domain 'Sky' and its subdomains

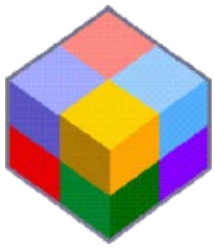
Instead of collecting and describing words one by one, the method collects the words all at once and then expand the dictionary one field at a time. At the heart of the program is a list of semantic domains. The list is exhaustive, universal and flexible, so it can be used to classify any word in any language.

Currently there are about 1,800 semantic domains, an example is below:

> **1.1.3.6 Lightning, thunder**
>
> Use this domain for words related to lightning and thunder.
>
> (1) What words refer to lightning? *lightning, lightning bolt, thunderbolt, lightning storm*
> (2) What does lightning do? *(lightning) strike, be struck by lightning, flash (of lightning), streak (of lightning), light up*
> (3) What do people use to protect themselves from lightning? *lightning rod*
> (4) What words refer to thunder? *thunder, thunderstorm, thundercloud, thunderhead*
> (5) What words describe the sound thunder makes? *peal (of thunder), clap, rumble, boom, crack of lightning, roll*

In a workshop setting small teams of native speakers work together to collect words. This synergy results in more words collected and at a more rapid pace. The procedure requires someone to read the domain template. The people working on the domain then think of all the vernacular words that belong to the domain (not translating the examples). These people do not have to be literate. Someone writes the words down. When they are finished, someone adds a simple gloss in the national language for each word. Then the domain is given to a typist for data entry. Teams spend about 10 minutes per domain, and can cover all the 1,800 domains within 2 weeks.

The program uses WeSay and FLEx to enter and manage the data.



Experience

Kayan dictionary.

Expanding beyond a single language: Karenic

The Karenic languages form a distinct cluster within the Sino-Tibetan family. Sino-Tibetan is the language family with the largest language (Mandarin Chinese). Whalen & Simons list 449 languages as part of this family. The Sinitic branch has 14 languages with over 1 million speakers per language. The remaining Tibeto-Burman branch has 435 languages distributed among 38 clusters.

- 3 clusters are extinct
- 4 clusters have a language with <2500 speakers
- 13 clusters have a language 15-83,000 speakers
- 12 clusters have a language 100-999,000 speakers
- 6 clusters have a language with $>$ 1,000,000 speakers

The Karenic cluster is a distinct group of between 20-30 languages with significant dialectal diversity.

The 6 largest languages range in size from 133,000 to 1.5 million. Five of these languages have writing traditions derived from missionary endeavours in the 19th century. There are significant numbers of speakers within these groups who can write.

By developing a lexicon for several languages within a language family branch, patterns of sound correspondence can be identified with fewer gaps in the data.

Once the framework of sound correspondences has been constructed, language data from smaller language samples can be added.

References

Boerger, Brenda. 2011. To BOLDly go where no one has gone before. *Language Documentation and Conservation* 5:208-233.

Brown, Cecil H., Holman, Eric W. & Søren Wichmann. 2013. Sound correspondences in the world's languages. *Language* 89:4-29.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–195. Online: http://corpus.linguistics.berkeley.edu/~ling240/himmelmann.pdf or http://hrelp.org/events/workshops/eldp2005/reading/himmelmann.pdf.

Newell, Leonard E. 1995. *Handbook on lexicography, for Philippine and other languages with illustrations from the Batad Ifugao Dictionary*. Manila: Linguistic Society of the Philippines.

Rapidwords. http://www.rapidwords.net/ [Accessed 2 September 2013].

Whalen, D. H. & Gary Simons. 2012. Endangered language families. *Language* 88:155-173.

Software

WeSay. http://wesay.palaso.org/

FLEx. http://fieldworks.sil.org/