Annotating archival recordings of Hocank (Winnebago)

Nancy Hall, Elica Sue & Irene Orellana

California State University Long Beach

We describe and discuss a project to annotate archival recordings of Hocank for linguistic analysis. Hocank, also known as Winnebago, is a Siouan language spoken in Nebraska and Wisconsin. It is highly endangered, with almost all fluent speakers over age 60.

Hocank has figured prominently in phonological theory, due to having several interesting and typologically unusual features. These include a third-mora accent pattern; a process of copy vowel epenthesis; a limitation of codas to obstruents; a complex interaction between accent and epenthesis; and different phonetic characteristics for lexical and epenthetic vowels. The challenges of explaining these patterns have spurred at least 20 analyses of Hocank phonology, including long treatments in seminal works like Halle & Vergnaud's *Essay on Stress* and Hayes' *Metrical Stress Theory*.

However, most analyses have been based on written data from a small number of sources, particularly Miner 1979 et seq., Susman 1943, and Hale & White Eagle 1980. There have been no acoustic phonetic descriptions. This gap is problematic, because impressionistic descriptions of phonetic patterns have heavily influenced some analyses. For example, Stanton & Zukoff's 2017 proposal of prosodic correspondence constraints hangs on the assumption that Hocank epenthetic vowels have a different pitch pattern than lexical vowels: a claim made tentatively by Miner 1979, but never backed up with acoustic data.

This gap could be addressed if linguists worked with existing recordings of Hocank. At least three field linguists have deposited collections of recordings in university libraries: Amelia Susman, Kenneth Miner, and Gerd Fraenkel. While these recordings are publicly available (with appropriate permission from archivists), they are not easy to work with, because the only documentation of their contents is scans of hand-written field notes. For a linguist interested in a particular phonological phenomenon, it requires considerable time and effort to make use of these. For example, suppose a linguist were interested in the accentual pattern of words with reduplication of epenthetic syllables (a topic discussed in Miner 1979). There are fewer than a dozen such words among the 7000+ word tokens Miner recorded. To find them, one would need to read hundreds of pages of hand-written notes to spot the words, determine which of the 28 tape reels those sections of notes belong to (the notes don't say), and finally locate the relevant few seconds within each hour-long reel. This is obviously impractical.

Annotation

Our goal in this project is to make words with particular phonological characteristics easily searchable, through phonologically annotated Praat textgrids. At the current stage, each textgrid consists of three tiers. Tier 1 contains a brief gloss, and the entry number from Miner's Winnebago Lexicon, if the word could be located there. Tier 2 is a phonemic transcription, using a version of ARPABET modified for the Hocank phonemic inventory. Tier 3 contains notes by annotators.

The ARPABET transcriptions make it possible to write scripts (for example, in Perl or Python) to locate words with particular characteristics. Using regular expressions, one can search for specific sequences of syllable weights, consonant / vowel combinations, etc, and extract the relevant snippets of recordings for analysis. With this resources, phonetic projects like the hypothetical reduplication study mentioned above become entirely practical.

We have so far annotated two corpora in this way. They are:

The Miner corpus (1974-5)

These recordings consist of 28 ¼ inch audio reels, most around one hour in length, mostly of one speaker. They are archived at the University of Wisconsin-Milwaukee Libraries as Tapes 108–135 in the Wisconsin Native American Language Project collection (copyright Great Lakes Inter-Tribal Council. The tapes were digitized in 2008. The library also holds photocopies (not originals) of Miner's notebooks, archived as UWM Manuscript Collection 20.

In 2017-2018, we annotated over 7000 Hocank words from these recordings, as described above. The tapes mostly contain elicitation of words or paradigms in isolation, with few examples of sentences or narratives. Although the corpus is a rich source for phonetic analysis, the fact that words are in isolation means that utterance-final intonational falls affect the accentual patterns.

The Fraenkel corpus (1959)

We are currently annotating an earlier set of tapes, recorded by Gerd Fraenkel in 1959 and archived at Indiana University. About 5 hours of digitized tapes are available, featuring 3 speakers. As Fraenkel never published on Hocank, the focus of his fieldwork is unknown, but the tapes seem to contain material complementary to that of the Miner tapes. There are many sentences and short narratives. Much is a re-elicitation of stories previously rcollected by Paul Radin.

Issues in annotation

In annotating these materials, we have encountered a number of methodological issues.

1. How should phonologically ambiguous features be annotated? There are many cases where Miner's field notes and dictionary disagree on vowel length and/or the location of accent. We have chosen not to transcribe accent (which is supposed to be predictable). For vowel length, we transcribed ambiguous tokens by ear, but this could create misleading data if we were wrong.
2. What other linguistic features would be most useful to annotate in future stages of the project? Ideas include Hocank orthography, morphological parses, intonational patterns, accent patterns, and the epenthetic / lexical status of vowels.
3. What is the best format making sure that annotations will be accessible in the long term? We used Praat textgrids because of their simple format---plain text files from which the labels and timestamps can easily be read manually. But some linguists have expressed a preference for receiving annotation in ELAN format.

References

Hale, K., & Eagle, J. W. (1980). A preliminary metrical account of Winnebago accent. *International Journal of American Linguistics*, *46*(2), 117-132.

Miner, K. L. (1979). Dorsey's law in Winnebago-Chiwere and Winnebago accent. *International journal of American linguistics*, 45(1), 25-33.

Stanton, J & Zuckoff, S. (2017) Prosodic identity in copy epenthesis: evidence for a correspondence-based approach. *Natural Language & Linguistic Theory*

Susman, A. (1943). *The accentual system of Winnebago*. Columbia University.