Jejueo talking dictionary:
A collaborative online database for language revitalization

Moira Saltzman

University of Michigan

`moiras@umich.edu`

Abstract

This paper describes the ongoing development of the Jejueo Talking Dictionary, a free online multimedia database and Android application. Jejueo is a critically endangered language spoken by 5,000-10,000 people throughout Jeju Province, South Korea, and in a diasporic enclave in Osaka, Japan. Under contact pressure from Standard Korean, Jejueo is undergoing rapid attrition (Kang, 2005; Kang, 2007), and most fluent speakers of Jejueo are now over 75 years old (UNESCO, 2010). In recent years, talking dictionaries have proven to be valuable tools in language revitalization programs worldwide (Nathan, 2006; Harrison and Anderson, 2006). As a collaborative team including linguists from Jeju National University, members of the Jejueo Preservation Society, Jeju community members and outside linguists, we are currently building a web-based talking dictionary of Jejueo along with an application for Android devices. The Jejueo talking dictionary will compile existing annotated video corpora of Jejueo songs, conversational genres and regional mythology into a multimedia database, to be supplemented by original annotated video recordings of natural language use. Lexemes and definitions will be accompanied by audio files of their pronunciation and occasional photos, in the case of items native to Jeju. The audio and video data will be tagged in Jejueo, Korean, Japanese and English so that users may search or browse the dictionary in any of these languages. Videos showing a range of discourse types will have interlinear glossing, so that users may search Jejueo particles as well as lexemes and grammatical topics, and find the tools to construct original Jejueo speech. The Jejueo talking dictionary will serve as a tool for language acquisition in Jejueo immersion programs in schools, as well as a repository for oral history and ceremonial speech. The aim of this paper is to discuss how the interests of diverse user communities may be addressed by the methodology, organization and scope of talking dictionaries.

1. Introduction

The purpose of this paper is to present the ongoing development of the Jejueo Talking Dictionary as an example of applying interdisciplinary methodology to create an enduring, multipurpose record of an endangered language. In this paper I examine strategies for gathering extensive data to create a multimodal online platform aimed at a wide variety of uses and user groups. The Jejueo Talking Dictionary project is tailored to diverse user communities on Jeju Island, South Korea, where Jejueo, the indigenous language, is critically endangered and underdocumented, but where the population's smart phone penetration rate is 75% (Lee, 2014) and semi-speakers are highly proficient users of technology (Song, 2012). The Jejueo Talking Dictionary is also intended for Jejueo speakers of varying degrees of fluency in Osaka, Japan, where up to 126,511 diasporic Jejuans reside (Southcott, 2013). A third aim of the Jejueo Talking Dictionary is to create extensive linguistic documentation of Jejeuo that will be available to the wider scientific community, as the vast majority of existing documentary materials on Jejueo are published in Korean. The Jejueo Talking Dictionary will serve as an online open-access repository of over 200 hours of natural and ceremonial language use, with interlinear glossing in Jejueo, Korean, Japanese and English.

2 Background

2.1 Language context

Very closely related to Korean, Jejueo is the indigenous language of Jeju Island, South Korea. Jejueo has 5,000-10,000 speakers

located throughout the islands of Jeju Province and in a diasporic enclave in Osaka, Japan. With most fluent speakers over 75 years old, Jejueo was classified as critically endangered by UNESCO in 2010. The Koreanic language family consists of at least two languages, Jejueo and Korean. Several regional varieties of Korean are spoken across the Korean peninsula, divided loosely along provincial lines. Jejueo and Korean are not mutually intelligible, owing to Jejueo's distinct lexicon and grammatical morphemes. Pilot research (Yang, 2013) estimates that 20-25% of the lexicons of Jejueo and Korean overlap, and a recent study (O'Grady, 2015) found that Jejueo is at most 12% intelligible to speakers of Korean on Korea's mainland. [1] Jejueo conserves many Middle Korean phonological and lexical features lost to MSK, including the Middle Korean phoneme /ɔ/ and terms such as *pizʌp* : Jejueo *pusʌp* 'charcoal burner' (Stonham, 2011: 97). Extensive lexical and morphological borrowing from Japanese, Mongolian and Manchurian is evident in Jejueo, owing to the Mongolian colonization of Jeju in the 13th and 14th centuries, Japan's annexation of Korea and occupation of Jeju between 1910 and 1945, and centuries of trade with Manchuria and Japan (Martin, 1993; Lee and Ramsey, 2000). Several place names in Jeju are arguably Japonic in origin, e.g. Tamna, the first known name of Jeju Island (Kwen ,1994:167; Vovin, 2013). Moreover, several names for indigenous fruits and vegetables on Jeju are borrowed from Japanese, e.g. *mi̱ʨaŋ* 'orange'. Mongolic speakers left the lexical imprint of a robust inventory of terms describing horses and cows, e.g. *mɔl* 'horse'. Jejueo borrowed grammatical morphemes from the Tungusic language Manchurian, e.g the dative suffixal particle **de < ti* 'to' (Kang, 2005).

2.2 Current status of Jejueo

The present situation in Jeju is one of language shift, where fewer than 10,000 people out of a population of 600,000 are fluent in Jejueo, and features of Jejueo's lexicon, morphosyntax and phonology are rapidly assimilating to Korean (Kang, 2005; Saltzman, 2014). Recent surveys on language ideologies of Jejueo speakers (Kim, 2011; Kim, 2013) show that a roughly diglossic situation is maintained by present day language ideologies. In a series of qualitative interviews on language ideologies, Kim (2013:33) finds common themes suggesting that Korean is used as a means of showing respect to unfamiliar interlocutors, as Korean "...is perceived as the language of distance and rationality". Likewise Jejueo is considered appropriate to use whenever interpersonal boundaries, such as distinctions within social hierarchies are perceived less salient than the intimacy and mutual trust two or more people share. (Kim, 2013).

Yang's (2013) pilot survey on language attitudes finds that while community members recognize Jejueo as a marker of Jeju identity worth transmitting to future generations, few speakers feel empowered to reverse the pattern of language shift to Korean. There are no longer monolingual speakers of Jejueo on Jeju or in Osaka. The examples below are samples of the same declarative construction produced by a fluent Jejueo speaker in (1), a typical younger Jejueo semi-speaker in (2), and the Korean translation (3). Jejueo morphemes in (2) are in boldface.

(1)
harmang -jʌŋ sontɛi -jʌŋ mi̱ʨaŋ
grandmother-CONJ grandchild-CONJ orange-
-ul tʰa -m -su -ta
ACC pick-PRS[PROG]-FO-DECL
"The grandmother and grandchild are picking oranges."

(2)
harmang -koa sontɛa -oa kjul
grandmother-CONJgrandchild-CONJ
orange-
*-ul ̱ʨa -ko i -**su-ta***
ACC pick-PROG-EXIST[PRS]-**FO.DECL**
"The grandmother and grandchild are picking oranges."

(3)
harmʌni -oa sontɛa -oa kjul
grandmother-CONJ grandchild-CONJ orange-
-ul ̱ʨa -ko iş -ʌjo
ACC pick-PROG EXIST[PRS]-FO.DECL

[1] In a 2015 study O'Grady and Yang found that speakers of Korean from four provinces on the mainland had rates of 8-12% intelligibility for Jejueo based on a comprehension task of a one-minute recording of Jejueo connected speech.

"The grandmother and grandchild are picking oranges."

While examples (1) and (3) have several cognate forms, the majority of grammatical particles are genetically unrelated. The accusative particle -ul is shared by Korean and Jejueo, although in Jejueo the nominative and accusative markers are most commonly dropped. In example (2) the construction the Jejueo morphemes have been replaced by Korean morphemes, save 'grandmother' and the verbal ending, a pattern typical of non-fluent speakers of Jejueo (Saltzman, 2014).

3 Jejueo lexicography and sustainability

Because most Korean linguists view Jejueo as a conservative dialect of Korean (Sohn, 1999; Song, 2012), lexical documentation of Jejueo has not been a scientific priority. The few Jejueo lexicographic projects have been carried out in the last 30 years by linguists native to Jeju Island and are all bilingual in Korean and Jejueo. Two large-scale Jejueo-Korean print dictionaries were published (Song, 2007; Kang, 1995), though Kang's oft-cited dictionary was given a small distribution to local community centers and libraries, and was not made commercially available. In 2011 Kang and Hyeong published an abridged Korean-Jejueo version of the dictionary. The remaining lexicographic studies of Jejueo are a handful of dictionaries tailored to individual semantic domains, such as 재주어 속담 사전 [Jejueo Idiom Dictionary] (Ko, 2002), 무가본풀이 사전 [Jeju Dictionary of Shamanic Terms] (Jin, 1991), and 문학 속의 제주 방언 [Jeju Dialect in Literature] (Kang et al., 2010), an alphabetized introduction to the Jejueo lexicon through Jeju folk literature. No major reference materials on Jejueo's lexicon or other linguistic features provide English glossing, although an English sketch grammar of Jejeuo is currently in development (Yang, in preparation).

It is well established that lexicographic materials can contribute significant symbolic support to a given language variety (Corris et al., 2004; Crowley, 1999; Hansford, 1991), particularly for unwritten non-prestige codes. Bartholomew and Schoenhals (1983) note that

publication of lexicographic materials may even help indigenous languages be perceived as 'real languages' in the sociolinguistic marketplace. The lexicographic materials alone, however, do not engender sufficient motivation for a speech community to maintain the use of their heritage language. Fishman (1991) warns against dictionary projects that become 'monuments' to a language rather than stimulating language use and intergenerational transmission.

A recent study by O'Grady (2015) found that the level of Jejueo transmission between generations shows a drastic decline. Given the task of answering content questions based on a one-minute recording of Jejueo connected speech, heritage speakers in the 50-60 age bracket demonstrated a comprehension level of 89%, while heritage speakers between 20 and 29 showed just 12% comprehension, equal to that of citizens of Seoul. In my previous field work in Jeju I found fluent Jejueo speakers and most semi-speakers unmotivated to access available Jejueo lexicographic materials. While these lexicographic works provide extensive data for the scholarly community, they arguably contribute to a growing body of Jejueo documentation and revitalization projects which are discrete, temporary and organized from the 'top-down' without community collaboration.

Sun Duk Mun, a Jejueo linguist with the Jeju Development Institute (JDI), reasons that Jeju parents must take pride in Jejueo and use it in the home, as Jeju teachers should allow Jejueo in classrooms, in order to expand Jejueo's declining domains of use (Southcott, 2015). However, in a highly competitive society where the majority of classroom hours are allocated to the Seoul-based national standard language (Song, 2012), and even entertainment media reflects the nation's emphasis on 'correct' usage of Korean, status planning for Jejueo revitalization is crucial. Beyond Kim's (2013) and Kim's (2011) studies on Jejueo language ideologies, no sociolinguistic research on Jejueo has been conducted, leaving issues like bilingualism, domains of use and the socio-historical factors for language shift speculative at most in the

literature. A successful campaign for the reversal of Jejueo language endangerment will hinge on the development of tools for documentation and language learning which reflect the socio-historical background and desires of the speech communities involved. To initiate such a campaign, ideological clarification and collaboration between the Jeju provincial government, Jejueo scholars, native speakers and educators is needed. The aim of the Jejueo Talking Dictionary is to match the diverse desires of Jeju users with collaborative methodology for data collection, as we will see in the next section.

4 Methodology

4.1 Community-based data collection

A primary goal of the Jejueo Talking Dictionary project is to train language activists in field linguistics to create a sustainable infrastructure for data collection, analysis, publication and archiving. In this way, Jeju community members will drive the scope of the Jejueo Talking Dictionary, in terms of adding the types of linguistic data that are found most useful to Jejueo-speaking communities in Jeju and Osaka. By training community members in linguistic documentation, Jejueo speakers and semi-speakers will have a foundation in field linguistics from which to build collaborative networks for crowdsourcing and status planning with Jejueo scholars, the provincial government, educators and elderly fluent speakers. At present, the team of foreign and local linguists developing the Jejueo Talking Dictionary is training local college students and activists at Jeju Global Inner Peace. Members of the team record elderly fluent speakers of Jejueo, annotate the recordings, and upload files into an open-access working corpus of data using Lingsync, a free online program for sharable audio and video files of annotated linguistic data. Linguists from Jeju National University, Jejueo specialists from the Jejueo Preservation Society, and I analyze the Jejueo data and check its accuracy with native speakers, ensuring the quality of the

corpus. In September, 2016 we will develop the corpus into a free online program and an Android application for smartphones.

4.2 Building an interdisciplinary network

A second goal of the Jejueo Talking Dictionary project is to build an interdisciplinary network for data collection. Our team of linguists from Jeju and abroad, language preservationists, activists and community elders have recruited ethnomusicologists, historians, experts in the indigenous religion and anthropologists to lend their expertise to the collection of lexemes and texts of various genres. In this way we aim to create a methodology of interdisciplinary data collection that builds a multidimensional record of Jejueo, to serve a wide range of uses and user groups. At present we have incorporated approximately 200 hours of previously unpublished annotated video data including Jeju oral history, shamanic rituals, indigenous music and cuisine preparation. Our team aims to enlist the support of ethnobotanists and ethnozoologists who can assist the team in collecting data on the indigenous flora and fauna of Jeju Island. In the future, data collection can be connected to language revitalization programs, such as master-apprentice programs (see Hinton, 1997), where semi-speakers join speakers of Jejueo in their usual activities farming, foraging for roots, herbs and vegetables in the mountains, picking seasonal fruit, and diving for seafood near Jeju's shores.

5 Contents of Jejueo Talking Dictionary

The Jejueo Talking Dictionary is intended to serve as a tool for both cultural education and language acquisition. With this in mind, we give equal attention to the collection of archaic and ceremonial speech, and the most frequently used lexemes and expressions. The Jejueo Talking Dictionary will compile existing annotated video corpora of Jejueo songs, conversational genres and regional mythology into a multimedia database,

supplemented by original annotated video recordings of natural language use. Lexemes and definitions will be accompanied by audio files of their pronunciation, listings of frequent collocations, and occasional photos, for items native to Jeju. The audio and video data will be tagged in Jejueo, Korean, Japanese and English, allowing users to search or browse the dictionary in any of these languages. Like Korean, Jejueo features complex agglutination of case, TMA and discourse register particles on verbs and nouns (Sohn, 1999). Videos showing a range of discourse types will have interlinear glossing, so that users may search Jejueo particles as well as lexemes and grammatical topics, and find the tools to construct original Jejueo speech. At the time of writing, we have recorded and annotated approximately 500 audio files of individual lexemes and 300 hours of video data. Below I itemize genres of Jejueo speech we have collected.

6.2 Inclusiveness versus usability

As the Jejueo Talking Dictionary is intended to serve a variety of uses, creating a cultural and linguistic repository of Jejueo stands somewhat at odds with developing a user-friendly dictionary for language education. We have found one solution to be to develop separate modules for the dictionary (see Vamarasi, 2013), so that it may be viewed according to individual purposes. In addition to viewing the dictionary page translated in Jejueo, Korean, Japanese and English, users can access separate modules from the main page. At present, these include modules for language lessons, browsing cultural topics, browsing photos, and browsing conversational genres and grammatical features. In the language education module, users access Jejueo lessons based around the most frequently used lexemes in the language, illustrated with photos. For this module we are also developing language-learning games and a 'word of the day' feature. All of the lexical entries in the dictionary and lexical items used in the narrative videos are tagged, so that searching a Jejueo word from any of the modules brings up a textual sample of the

lexeme in a grammatical construction, and videos featuring the lexeme in natural language use, where that data is available. Transcripts of all of the videos may be downloaded and printed, and users may select a transcript with interlinear glossing, the Jejueo transcription only, or a translation in Korean, Japanese or English.

6.3 Language standardization

It is important to note that the standardization of Jejueo orthography is still ongoing, and among the several regional varieties of Jejueo, none has been designated as the standard. For the Jejueo Talking Dictionary project we adopt the orthographic preferences of the most recent Jejueo lexicographic materials (Kang and Hyeon 2011; Kang, 2007), and list headwords and regional variants in the order assigned in those materials.

7 Conclusion

An open-ended lexicographic resource such as an online talking dictionary lends itself well to incorporating a variety of data designed to serve diverse uses and user groups. The Jejueo Talking Dictionary can be continuously and cost-effectively modified as we obtain more data and gain feedback on the usability of the dictionary. We aim for the Jejueo Talking Dictionary to be an accessible multipurpose repository of the Jejueo language, where the content and the collection of linguistic data are both driven by the Jeju community. With Jejueo in a state of critical endangerment, incorporating community members in the development and dissemination of language-learning materials is key.

Audio recordings	Video recordings
anatomy	conversations focusing on discourse markers
common expressions	game playing
cuisine	oral histories of 4.3 Massacre, Japanese occupation, marriage rituals, farming, fishing
cardinal directions	preparation of native cuisine
flora	preparation of rituals (Buddhist, shamanic)
fauna	shamanic rituals: major public rituals for lunar year, family rituals
geography	stories of indigenous mythology (regional varieties)
grammatical suffixes	work songs, chants (regional varieties)
high frequency nouns, verbs, adjectives	
ideophones	
idioms	
kinship terminology	
mythological and shamanic terminology	
weather terminology (polysemic terms for wind, rain)	

Table 1. Jejueo Talking Dictionary corpus

References

Bartholomew, Doris and Louise Schoenhals. 1983. *Bilingual dictionaries for indigenous languages.* Hidalgo, Mexico: Summer Institute of Linguistics.

Corris, Miriam, Christopher Mannig, Susan Poetsch and Jane Simpson. 2004. How useful and usable are dictionaries for speakers of Australian indigenous people. *International Journal of Lexicography*, 17(1): 33-68.

Crowley, Terry. 1999. *Ura: A disappearing language of Southern Vanuatu.* Canberra: Pacific Linguistics.

Fishman, Joshua. 1991. *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Bristol: Multilingual Matters.

Hansford, Gillian. 1991. Will Kofi understand the white woman's dictionary? Some ways to make a bilingual dictionary more usable to a new literate. *Notes on Linguistics*, 52, 17-28.

Harrison, K. David and Gregory D.S Anderson. 2006. Tuvan Talking Dictionary. Retrieved from http://tuvan.talkingdictionary.org

Hinton, Leanne. 1997. Survival of endangered languages: The California master-apprentice program. *International Journal of the Sociology of Language,* 123, 177-191.

Jin, Seong-Gi. 1991. *Dictionary of Jeju Shamanic Terms*. Jeju: Minsokweon.

Kang, Jeong-Hui. 2005. *A study of morphological change in the Jeju dialect*. Seoul: Yeokrak.

Kang, Yeong-Bong. 1994. *The language of Jeju*. Jeju: Jeju Munhwa.

Kang, Yeong-Bong. 2007. *Jeju Language.* Seoul: National Folk Museum, Jeju Special Self-Governing Province.

Kang, Yeong-Bong, Kim, Dong-Yun and Kim Sun-Ja. 2010. *Jeju Dialect in Literature*. Jeju: Gukrimgukeoweon.

Kang, Yeong-Bong. & Hyeon Pyeong-Hyo. 2011. *Standard Language- Jejueo Dictionary*. Jeju: Doseochongpan Gak.

Kim, Soung-U. 2013. *Language attitudes on Jeju Island - an analysis of attitudes towards language choice from an ethnographic perspective.* Master's thesis, School of Oriental and African Studies, London.

Kim, Sun-Ja. 2011. *A Geolinguistic Study on the Jeju Dialect*. Ph.D. dissertation, Jeju National University, Jeju.

Ko, Jae-Hwan. 2002. *Jejueo Idiom Dictionary.* Jeju: Minsokweon.

Kwen, Sangno. 1994. *A diachronic dictionary of Korean place names*. Seoul: Ihwa Munhwa Chwulphansa.

Lee, Iksop and Robert Ramsey. 2000. *The Korean language.* Albany: SUNY Press.

Lee, Minjeong. 2014, December 12. Smartphone usage overtakes PCs in South Korea. *Wall Street Journal.*

Martin, Samuel E. 1993. *A reference grammar of Korean: A complete guide to the grammar and history of the Korean language.* Rutland, Vermont: Charles E. Tuttle.

Nathan, David. 2006. Thick interfaces: mobilizing language documentation with multimedia. In J. Gippert, N.P. Himmelmann, & U. Mosel (Eds.), *Essentials of Language Documentation* (pp. 363-390). Berlin: Mouton de Gruyter.

O'Grady, William. 2015. *Jejueo: Korea's other language*. Paper presented at ICLDC 5, Honolulu.

Saltzman, Moira. 2014. *Language contact and morphological change in Jejueo*. Master's thesis, Wayne State University, Detroit.

Sohn, Ho-Min. 1999. *The Korean Language*. Cambridge: Cambridge University Press.

Song, Sang-Jo. *Big dictionary of Jeju language*. Seoul: Hanguk Munhwasa.

Song, Jae-Jung. 2012. South Korea: language policy and planning in the making. *Current Issues in Language Planning,* 13. 1-68.

Southcott, Darren. 2013, September 10. The story of Little Jeju: Jaeil Jejuin. *The Jeju Weekly*, pp. 2A.

Southcott, Darren. 2015, May 13. Parents must take pride in Jejueo. *The Jeju Weekly,* pp. 3B.

Stonham, John. 2011. Middle Korean △ and the Cheju dialect. *Bulletin of the School of Oriental and African Studies,* 74. 97-118.

UNESCO Culture Sector. 2010. Concerted efforts for the revitalization of Jeju language. Retrieved from http://www.unesco.org/new/en/culture/themes/endangered-languages/news/dynamic-contentsingleviewnews/news/concerted_efforts

_for_the_revitalization_of_jeju_language/#.U1 WeFFeZi4Y. Accessed 10 February 2015.

Vamarasi, Marit. 2013. The creation of learner-centred dictionaries for endangered languages: a Rotuman example. *Journal of Multilingual and Multicultural Development.* doi:10.1080/01434632.2013.825266

Vovin, Alexander. 2013. From Koguryo to Tamna: Slowly riding to the South with speakers of Proto-Korean. *Korean Linguistics, 15*. 222–240.

Yang, Chang-Yong. 2013. *Reference grammar of Jejueo*. Manuscript in preparation.