Dialectologia. Special issue, IV (2013), 199-211.
ISSN: 2013-2247

Received 22 May 2013.

Accepted 3 September 2013.

MAKING A DIALECT DICTIONARY FROM A SENTENCE-BASED CORPUS

Yumi NAKAJIMA
Hitotsubashi University, Tokyo
yumi.nakajima@r.hit-u.ac.jp

Abstract

I describe a project making a dialect dictionary (Tokunoshima, Amami archipelago, Japan) from a sentence-based corpus, which my colleagues Motoei Sawaki, Chistsuko Fukushima and I have been engaged in these 13 years. The Tokunoshima dialect is in a highly critical situation, so it is urgent that we now describe the local speech of the island, which until now lacked full-scale dictionaries. We have a well-trained informant Takahiro Okamura, who completed his original translation of "Two Thousand Sentences of Japanese" by Shigeo Kawamoto into his Asama dialect with us. "Two Thousand Sentences of Tokunoshima dialect" so vividly reflected the local life that it led us to make full use of the data as a digital dictionary of sentences with various searching functions, about which we have previously reported at the International Society for Dialectologists and Geolinguists (SIDG) conferences. Here I sketch some episodes from our cooperative works with Okamura, introducing how we started it and what problems we experienced.

Keywords
endangered dialects, Tokunoshima dialect, "Two Thousand Sentences", sentence-based corpus

ELABORANDO UN DICCIONARIO DIALECTAL A PARTIR DE UN CORPUS BASADO EN FRASES

Resumen

Se describe un proyecto de elaboración de un diccionario dialectal (Tokunoshima, archipiélago de Amami, Japón) a partir de un corpus basado en frases, en el cual han participado Motoei Sawaki, Chitsuko Fukushima y Yumi Nakajima en los trece últimos años. El dialecto de Tokunoshima está en una situación muy crítica, por lo que es urgente que ahora se describa el habla local de la isla, que hasta ahora carecía de diccionarios a gran escala. Tenemos un informador bien entrenado, Takahiro Okamura, quien completó la traducción original de "Dos mil frases japonesas" de Shigeo Kawamoto en su dialecto Asama. Las “Dos mil frases del dialecto de Tokunoshima” reflejan tan vívidamente la vida local que hemos hecho pleno uso de esos datos para elaborar un diccionario digital de las frases obtenidas que permite varias funciones de búsqueda, y sobre el que se ha informado en conferencias anteriores de la International Society for Dialectologists and Geolinguists (SIDG). Aquí se esbozan algunos episodios de nuestros trabajos cooperativos con Okamura, describiendo como se inició el proyecto y los problemas que han surgido.

Palabras clave
dialectos amenazados, dialecto de Tokunoshima, “Dos mil frases”, corpus basado en frases

1. The road to the Tokunoshima dialect

Up until now, I have already reported on our project with Motoei Sawaki and Chitsuko Fukushima — that is, an attempt to make a dictionary through the use of sentence data — more than once at the International Society of Dialectologists and Geolinguists (SIDG),[1] and would like to refer readers to those reports in regards to the overview and the structure of the data. Within this article, however, I would like to introduce again how the project began, and the path that we followed thereafter, summarizing along the way the advantages and problems that we discovered.

We three — Sawaki, Fukushima and Nakajima — began our research activities towards the construction of a dictionary of the Tokunoshima dialect in 2000. However, efforts to record the Tokunoshima dialect can be traced back to two linguistic geographical surveys conducted in 1975 and 1976 under the leadership of Takesi Sibata. Having taken on the challenge of linguistic-geographical surveying all of the villages two times with over 20 researchers, these surveys confirmed the reality of the dialectal differences within the island, primarily on the phonemic and lexical levels, and yielded many great results. However, looking back upon it now, while there were some attempts to look partially at things like kinship terms and phonemic distributions etc., quite wastefully they never realized into a full description of the dialect, with time

[1] M. Sawaki, Y. Nakajima, C. Fukushima, “Dialect Corpus as a Resource for Dialect Dictionary”, 4th SIDG, Riga (2003); “The Making of a Multimedia Dialect Dictionary as a Database with Indexes and Cross-references”, 5th SIDG, Braga (2006); “Making the Most of a Dialect Corpus: Development of the Tokunoshima Dialect Dictionary of Two Thousand Sentences”, 7th SIDG, Vienna (2012).

alone passing on. Even amongst that, we owe our continued relations with the island entirely to Takahiro Okamura, a junior high school teacher from Tokunoshima, Asama village.

Having awoken to the fascination of his own dialect through Sibata's encouragement, Okamura came from his post in Amami Oshima to Tokunoshima to cooperate with the 1975-1976 surveys. Okamura supplied us with invaluable information as an informant for the conveniences of the surveys, and with his naturally cheerful character we soon found ourselves hitting it off swimmingly. Afterwards, Okamura took full advantage of his posts in the different islands of Amami, and deepened his research through his own surveys, thus continuing our collegial exchange.

2. A new challenge

In the spring of 1995, Okamura reached the mandatory retirement age from his position as a junior high school teacher, and began a free life of retirement in his hometown of Asama. It was then that Sawaki brought students with him to embark upon practical dialectal survey training to Asama, and his report posed quite a shock to Fukushima and myself.

When we went around the island in the 1970s, the dialect was still firmly rooted in daily life on the island. The older generation did not understand standard Japanese well, to the degree that one would need to have their grade-school age grandchildren act as interpreters to finally make communication. However, the social changes that occurred in the years afterwards led to the transformation of life on the island, which naturally also had an influence on the language as well. The unique phonetic points that characterize the Ryukyu and Amami dialects such as the glottal stops that accompanies the beginning of vowels, the glottalization of consonants, and central vowels were all but completely gone from the language of young speakers, and the ability to command the dialect in daily life beyond the use of some limited vocabulary and phrases was also decreasing even amongst people of Okamura's generation.

Born of our sense of crisis concerning the demise of the dialect, we and Okamura were in accord in feeling that it was necessary to document the dialects there while it was still possible, and began our activities the following year. As regards the Amami islands, there is one superior dictionary on the Yoron island dialect compiled by the invaluable speaker Chiyo Kiku, as well as on the Amami Oshima dialect by Suma Nagata et.al., but aside from a few vocabulary lists, there was no comprehensive dictionary whatsoever on the Tokunoshima dialect. One of the reasons that it was difficult to compile a single dictionary of the dialect is the great linguistic differences within the island itself, but the crisis was already right in front of us. The first thing that we had to do was to document the dialects using Okamura's Asama dialect as a base-point. Even so, what we had been thinking of at the time was something more along the lines of traditional methods, like collecting vocabulary or building up a bank of word interpretations. With that, over our ever-anticipated summer vacation we settled down into the island, starting by redoing the basic lexical surveys while getting ourselves covered in sweat.

Okamura is a great lover of languages; having also been raised by his grandmother as a youth, he remembered many of the words from the older generations. Conversations would start with the differences between the dialects in neighboring villages and go on to those with other Amami islands, finally expanding outwards to touch upon past customs. For example, the story on the word *head*, from the list of body parts right at bat in the beginning of the list of 1,000 basic words, went like this. (A key to the phonetic descriptions may be found below, footnote 2.)[2]

"*Head* is [kara:zI]. There's also [cIbu:ru], which used to be used for people's heads but is now used to refer to pumpkins. [CjuNcIbu:ru] only refers to human skulls that have become skeletal. As a remnant of the old custom of disposing of dead bodies by exposing them to the elements before we start cremation, even 60 years ago you could still see many of them left in [jo:] or cliff-holes. The word [jo:] maybe comes from the word *iori* or a hermitage? But the same skull is called a [ko:bI] when it's been

[2] We introduced a rough phonemic transcription of the Tokunoshima dialect in Roman characters using only the keyboard characters, i.e., capital I and E for the two central vowels to distinguish the others, capital consonants for the glottalized [k], [m], [n], [t], [t] etc., except "N", which is for the syllabic [n]. "' " is for the initial glottal stops before vowels, " : " for the long vowels.

reburied, and reburials themselves are called [ko:bItui], a taking of the head. Foreign pumpkins are called [to:cIbu:ru], which reminds me that wasp nests, since they're shaped like pumpkins, are called [cIburIba:cI]. For head, there's also ['ukkaN], and I think this is from *okanmuri*, or a crown, but people with big head are called ['ukkaNnu futE:hai] (="'ukkaN" is big). Talking about whether someone is/is not intelligent, though, we do not say [kara:zI], but say [no:] (ex: "no:" is good/bad.)."

Roughly like this, one could use his descriptions to determine what words they were related to in classical Japanese and the mainland dialects, when they spread to Amami, and when they became more semantically limited, making it extremely fascinating. These descriptions have great value in regard to the history of Japanese, and the cultures and customs which form their background could lead to ethnographic discoveries. (The notes that I sorted through for body parts starting with word No. 1, *head*, to word No. 49, *wound*, came to 16 A4-size pages of descriptions, and would easily go over 20 pages including Okamura's lexical list draft descriptions.) Another time, at the word *cow,* the discussion moved on to the fighting bulls that are the pride of Tokunoshima, leading to the assertion that bulls "will not follow standard Japanese", and to a never-ending explanation of the commands used towards bulls.

While the work is fun, how many years will it take to collect and organize the 1,000 basic words like this? To what level should we expand the list of related words, and build interpretation banks for them? How much can we really do, given how much time we need with our abilities just to determine the form of the words themselves? This is how we were feeling when our savior, *Two Thousand Japanese Sentences*, appeared.

3. Marvelous sentence data

It was a coincidence that led us to using sentence-based data in compiling our dictionary. In the middle of our basic vocabulary survey, Sawaki introduced Okamura to *Two Thousand Japanese Sentences*, it having suddenly come to mind. While based on Henri Frei's *Le Livre des Deux Mille Phrases* (1951), *Two Thousand Japanese Sentences* was written in Japanese with its own free ideas; when we were students, we used it for things such as comparing sentence patterns in various languages, but we had forgotten about it after that. We brought it out thinking we could use it as a reference to make example sentences for each index word.

While there are several collections of example sentences from which one could obtain sentence data, their principle objectives are to cover all of the grammatically necessary expressions, and aside from being limited in vocabulary, are often not a bit unnatural in setting. In comparison, starting with the various workings of humans, *Two Thousand Sentences* was prepared with many settings (8 categories over 150 items), and its contents are colorful, making it interesting to read as well.

Okamura, having encountered *Two Thousand Sentences*, found himself drawn in by the interesting examples, and got straight to work rewriting them into dialect. Who knows how much concentration he drained into it, but after just two months passed *Two Thousand Sentences* of the Tokunoshima Dialect was completely done. Overflowing with vivid dialectal expressions, we found ourselves just filled with admiration in front of this masterpiece — but that was not the only surprise. There were also expressions within the original Japanese version that, naturally, were difficult to express, as is in dialect. In those cases, he made full use of the meaning of the sentences, changing them to expressions that would be natural in dialectal life. There were also cases where just the vocabulary was not appropriate, as well as examples were the settings themselves would be difficult to imagine within life on the island. In order that we could know how the process went, Okamura also supplied literal retranslations back into standard Japanese from the dialect examples that he made.

We began to grope about for ways that we could utilize these great results, and the first thing that was needed for that was to decide on how to transcribe the examples — which necessitates work on a computer using an enormous amount of data. We decided then to use a modified phonemic transcription using roman characters on the basic data, setting up a system for transcribing the dialect only using symbols that could be typed in using the keyboard. We understood that the Asama dialect could, due to its phonemic structure, be transcribed without problems if one differentiated between the capital and lower case letters, and expressed glottalization using the symbol “ ’ ”. On the other hand, for ease of reading it was also desirable to have a transcription using kanji (Chinese characters) and kana (the Japanese syllabary): Since our goal was to make the dictionary into something people on the island could use themselves, transcriptions in roman letters alone would be problematic. In the end, we decided to standardize the data into the following four forms.

1) Standard Japanese sentences (in kanji and kana)

2) Japanese literal translations of the Tokunoshima dialect examples (in kanji and kana)

3) Tokunoshima dialect examples in roman characters

4) Tokunoshima dialect examples in kanji and kana

At first, we had only been thinking of including 2) as a reference. However, to us researchers whose understanding of the dialect was incomplete, our understanding of how the dialectal expressions corresponded to standard Japanese was much better having looked at these translations, which should also be the same for young people on the island, as well. Is there no way to use these translations?

It turns out that the data from 2) would itself come to have a huge role in the utilization of sentence data later on. That was because we realized that by organizing 2) as a key with their corresponding dialectal sentences, we could create a simple dictionary. This was the advantage of using literal translations. However, there are problems with this as well; as opposed to European languages, Japanese does not divide text into words when writing, meaning that unless one divided up the sentences into some sort of elements, one could not ascertain correspondences between the two sets of data.

Thus we decided to utilize *phrases* made up of *substantives + ancillary* words. Phrases can said to be the parts that make up a sentence in agglutinative languages like Japanese (example below).

Two Thousand Sentences: example (No. 351)

１）だれかがドアをノックしているよ。(Somebody is knocking on the door.)

２）誰かが扉を打っているよ。

３）誰（たる）がだん　やーどぅ　打っちゅんだー。

4）　[tarugada]N ja:[du] [‘uc]cjuN[da:].

phrases:	tarugadaN	ja:du	‘uccjuNda:
structure:			
	taru-ga-daN	**ja:-du**	**’uccjuN-da:**
	somebody-nominative-question	house door-direct object	beat-progressive-sentence final

Of course, one must have a grasp of the morphemic and semantic functions of non-substantives in order to understand the grammar. However, at this stage the analysis of the endings for verbs, adjectives and auxiliaries had really not been at all developed yet, and it was also possible that within this dialect, nouns and particles would reduce into one word, changing the morphemes. Analyzing sentences by phrases seemed to be the best and safest method we could attempt at that time, and we settled it by deciding that it would be fine to go back to those detailed elements later on.

4. A bright — or hard? — future looming

Breaking up the text into phrases required a lot of manual labor while referring to the punctuation added in by Okamura. When the results from this process came back from Sawaki, who had been devoted to it, we once more found ourselves completely filled with the value of the sentences that Okamura created. With just this varied data we could make KWIC search tools, reverse searches and the ability check for word frequencies. Fukushima, who was in charge of the conjugation analysis, found that this allowed her to extract the verb and adjective conjugations easily, which overjoyed her with the sudden advancement of the analysis. *Ahh, the bright future ahead of us!* We were really feeling good when we decided to start planning for a multimedia dictionary. In addition to the kinds of data mentioned above, we also really wanted audio data for the Two Thousand dialectic sentences. *We definitely want to utilize the basic vocabulary data we organized — and wouldn't it be ideal if we could have videos showing how Okamura was pronouncing them?* Our dreams got bigger and bigger, and I started preparations to record audio and video versions of Okamura reading the sentences.

I was in the midst of these preparations, however, when I was shocked to find that there were inadequacies in the *Two Thousand sentences* themselves. When I collated the phrases together, there were cases where the literal translations and the dialects failed to correspond, or that there were phrases without anything to correspond to altogether. There were some incomprehensible things and some simple mistakes, too, but partially due to our lack of understanding, I could not always determine whether the correspondences were accurate. I had been in charge of the data checks, and I felt like everything around me went black — but this was all just the fact that I did not come face to face with the Tokunoshima dialects coming back to haunt us. What should I do? Clearly, there was nothing else for me to do but to go carefully back over the original data one by one with Okamura.

Then I came upon a plan to break up the 2,000 sentences into blocks of 500, and, after reading them together with Okamura and reconfirming them, record the audio, which we could then listen to and make more cross-checks. As the transcriptions in the Roman alphabet had accent information on them, we also needed to check that the audio recordings and the transcriptions were consistent. While we reread the data and analyzed them one by one, checking whether the Japanese and the dialect examples were consistent in meaning and whether the dialect examples were natural, going over them like this we also discovered new problems: In comparison with words, it is very difficult to determine what is a corresponding sentence in the dialect.

However, these labors were extremely stimulating, and most importantly, they were effective in bringing up our understanding of the dialect. How did we examine the meanings of the sentences themselves? For example, the sentence *koutsuu ga* *todaeteiru* (traffic is at a standstill; 0391: all transportation is down) is perfect for talking about when typhoons hit the island, so one could also say *fune mo hikouki mo nani mo konai yo* (no boats or planes will come). However, in regards to *furo agari no sappari shita kanji* (how you feel when you get out of the bath; 0792: feeling fresh and pleasant), which seems perfectly fine in standard Japanese, Okamura made the following comment: "On the island in the summer, you never stop sweating and your body is constantly wet, and the temperatures don't change very much at night, so it'd be really wretched to get into a hot bath — getting out of a hot bath can't be *sappari* (fresh)". In actuality, when I first went to the island, I myself had the experience of not listening to my host's words of caution, jumping right into the freshly filled bathtub for first dibs, only to find my body glowing throughout the night, unable to sleep. Okamura selected the onomatopoeic expression [kasIkasIsjI:], which, if on the mainland, would mean something along the lines of "so dry there's no wetness at all", saying that that was the way that you could talking about being *sappari* on the island.

There were also cases that no matter how many times we tried to rewrite them, we could not get examples that seemed natural in the dialect. For the phrase *gakkari suru* (to be disappointed; 0911: I was disappointed in him). Okamura claimed that it was difficult to say that one is *gakkari* (disappointed) in regards to a human. So, we tried several times to come up with examples that would express that one is disappointed with some unexpected thing, but nothing really felt right. When we finally came up with the expression *kotoshi wa mikan ga taberareru to omotta no ni, kaze ni yararete gakkari shiteiru* (I thought that I would be able to eat mandarin oranges this year, but they got destroyed by the wind and I'm disappointed), we put our hands together in joy. Actually, this very situation came to reality in 2011. That May, an out of season typhoon lingered over the Amami islands, leading to strong winds blowing the mandarin orange flowers away right when they were to bloom; the mandarin orange orchards were completely destroyed, striking great damage to the island industry. That winter when I went to Okamura's house and visited his garden, all of the mandarin orange trees were cut down, making me think that this must have been what he was talking about.

Aside from those kinds of difficult questions, sometimes Okamura jumped the gun on things during the first stages of work. Once, the expression *soutou hakaga ikimashou ne* (let's make progress fast) became *soutou ka ga iru deshou ne* (there really are many mosquitoes). Perhaps because he was not used to the standard Japanese expression *hakaga iku*, he evidently mistook where to break it into phrases. When we had fun discoveries like that, we would laugh out loud for a change of pace.

5. Next task

In this way, the data to be used was settled, and in 2006, we were able to make a paper-version which included the *Two Thousand Sentences* with the four different versions, a list of the phrases in the dialect and their corresponding standard Japanese — e.g., a simple dictionary — as well as the results of the verb declensions that Fukushima had been working on. In 2009, the electronic data versions that Sawaki had devised and the audio and visual data were all put together, and included as a DVD attached to the revised paper-copy.

Our "dictionary" completely evolved into something we never anticipated, but of course, this is just one step on the road towards getting the final dictionary. As described above, our next job is to divide the phrases, and determine the morphemic and phonemic functions of the non-substantives. Sawaki, feeling that a tagged corpus was essential, is now wrestling with XML programming in order to find the best method for doing so. I am currently transcribing the smallest morphemic units and their semantic functions while confirming with Okamura's sense as a native speaker. Fukushima is also moving forward with her work while keeping the determinations of paradigms for verbs, adjectives and also auxiliaries. All three of us are in the midst of a fight to move forward from the provisional units of phrases.

Today, it is perfectly fine for there to be many different types of dictionaries. Especially now that IT skills have become so much more familiar to us all, there is no reason to not use that accessibility. The electronic version of our *Two Thousand Sentences Dictionary* now allows one to search within the sentences themselves so as to see what a particular phrase corresponds to in standard Japanese or how it is used in other sentences, regardless of how it appeared within the original sentence. Lately, there are many online dictionaries which use databases; such dictionaries allow users to search for all of the word forms at once, making it possible to compare recent usages, idiomatic phrases as well as even things like how conjugations and variant forms are being used. These offer conveniences that a paper dictionary cannot, but there are good points about a paper dictionary, too. To taste the unfettered expressivity of the dialect, one finds oneself wanting a paper copy.

We gave our best effort in creating *Two Thousand Sentences of the Tokunoshima Dialect* to make sure that a sense of life on the island came through, that it was not some tiring expressions of the old days, but rather a reflection of current life on the island, where people go to Tokyo by airplane and watch television every day. If the sentences reflect that life well, and are interesting, then they should be easy to remember; and if one remembers the phrases, then it is easy to make new sentences on one's own by putting them into other sentences with the same structures. For example, if one remembers *atama ga itai* (one's head hurts), and one also knows other words referring to body parts from the basic vocabulary list, then it is simple to come up with sentences like *me ga itai* (one's eyes hurt) or *ashi ga itai* (one's legs hurt). In this way, a sentence dictionary can be used as a set with data on the elements that make up the parts of sentences to be used to the fullest. This is actually something that we ourselves experienced personally, and while we still have a long way to go, our understanding of the dialect is much higher now than it was when we first started the project in 2000.

In 2012 and 2013, we did a simple survey amongst individuals in the prime of life to see how much the dialect was currently being used throughout the island.[3] There are many people in their 40s who still understand the language of their elders, and who hope to spread the island's dialect down to the children. However, young people feel that "they don't know what to say" or "it's embarrassing to say it wrong", and appear to be hesitant to use the dialect. Our hope is that we can make a dictionary that will not just be a record of what words that used to be used meant, but rather a

[3] The report is published on the following website: <http://www.ninjal.ac.jp/socioling/nwavap02/working~papers.html>, Motoei Sawaki, Yumi Nakajima, and Chitsuko Fukushima, "Standardization and Dialect Leveling in Tokunoshima".

dictionary that will be useful for young people when trying to create sentences in the dialect themselves.

References

FREI, Henri (1953) *Le Livre des deux milles phrases*, Geneve: Droz.

KAWAMOTO, Shigeo (1971) *The Two Thousand Sentences of Japanese Language*, Tokyo: Waseda Institute of Language and Culture.

SIBATA, Takesi *et al.*: Research Group of Linguistics Department, Tokyo University (1972) *Dialect in Tokunoshima*, *Amami*, Tokyo: Akiyama Shoten.

Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, *The Linguistic Questionnaire for Asian and African Languages* 1 (1966), 2 (1967), Tokyo University of foreign Studies.