Dialectologia. Special issue, IV *(2013), 213-219.*
ISSN: 2013-2247

Received 1 October 2013.
Accepted 20 October 2013.

DIALECT DICTIONARY IN A DVD

Motoei SAWAKI
Shinshu University, Matsumoto
msawaki@shinshu-u.ac.jp

Abstract

A team of three researchers published *Two Thousand Sentences of Tokunoshima Dialect* with a DVD whose main content includes KWIC (Key Word In Context) reference data in Tokunoshima dialect. KWIC is a very powerful tool for analyzing the dialect. Moreover, the raw data from which the KWIC was made and the programs for processing the data are also included in the DVD. Thus, how the final result was obtained is made clear in this article.

Keywords

Tokunoshima dilalect, raw data, hypertext, KWIC (Key Word In Context), programs

UN DICCIONARIO DIALECTAL EN DVD

Resumen

Un equipo de tres investigadores ha publicado el volumen *Two Thousand Sentences of Tokunoshima Dialect* con un DVD cuyo contenido principal incorpora los datos de referencia del dialecto de Tokunoshima en KWIC (palabra clave en contexto). KWIC es una herramienta muy poderosa para analizar el dialecto. Por otra parte, los datos en bruto, a partir de los cuales se elaboró KWIC y los programas para procesar los datos, también se han incluido en el DVD. De este modo en este artículo se hace patente la obtenció del resultado final.

Palabras clave

dialecto de Tokunoshima, datos en bruto, hipertexto, KWIC (Key Word In Context), programas

1. Introduction

We, Motoei Sawaki, Yumi Nakajima and Chitsuko Fukushima, published *Two Thousand Sentences of Tokunoshima Dialect*, which was accompanied with a DVD. The DVD was something which can be called a hypertext. Instantly after you load the DVD into your PC, a program automatically starts as a portal to various contents. When you click on a speaker icon, you can hear recitation of the dialectal text of *Two Thousand Sentences*. When you click on a sentence number, you jump to a new page showing the sentence the number indicates and so on.

As to the "Two Thousand Sentences", refer to Nakajima's paper in this issue. Here I use an acronym TTS for "Two Thousand Sentences" and TTST for *Two Thousand Sentences in Tokunoshima Dialect* which was published in a book format.

2. Contents of the TTST DVD

The main menu of the DVD is shown in Figure 1.

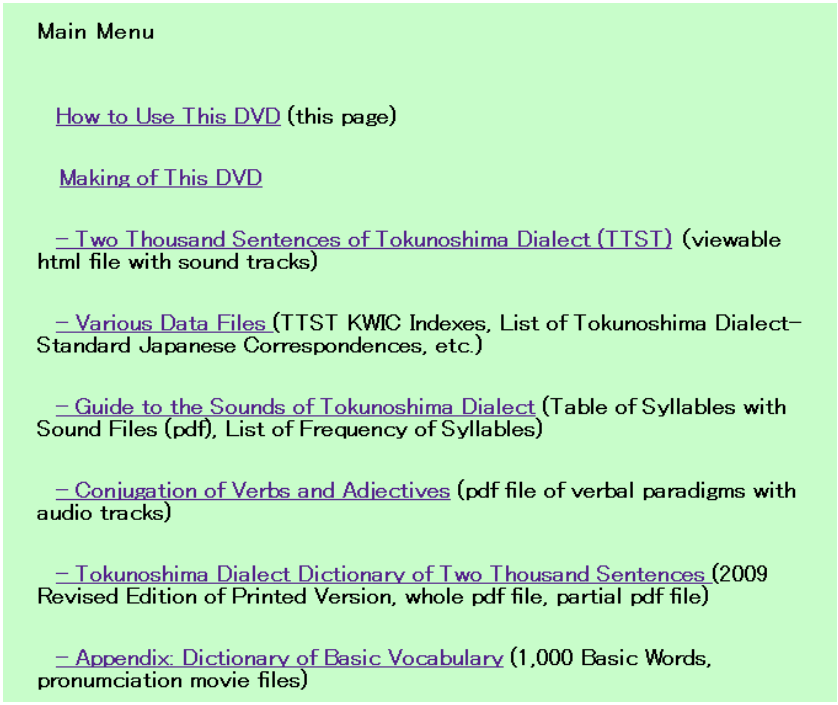

Figure 1. Main menu of *Two Thousand Sentences in Tokunoshima Dialect*

The contents of the DVD are following:

2.1 TTS in the Tokunoshima dialect and standard Japanese

Standard Japanese translation has two versions: one is literal translation and another is free translation. The texts are compiled in viewable html file with sound tracks

2.2 Lists of sets of phrases

Each phrase (as to "phrase", also refer to Nakajima) in Tokunoshima dialect is paired with one in standard Japanese.

Each pair of sentences in Tokunoshima dialect and standard Japanese are divided into phrases with strict one-to-one-correspondence. That is to say when a sentence in Tokunoshima dialect is broken up into phrases A, B, C, D... and the corresponding sentence in standard Japanese is also broken up into A', B' ,C' ,D'..., A and A' have approximately the same meaning and so are B and B', C and C', D and D'... We did this way of breaking up sentences intentionally, so that we can safely match a Tokunoshima dialect phrase with a standard Japanese counterpart.

Thus we obtained a set of pairs of phrases which is sorted in four different orders; normal (dictionary) order using Tokunoshima dialect as the key, reverse order using Tokunoshima dialect as the key, normal (dictionary) order using standard Japanese as the key, and reverse order using standard Japanese as the key.

The reverse order is one as if you looked key words from tail to head. For example, "acd" is placed after "bba" in the reverse order.

2.3 KWIC (Key Word In Context) data of TTS

Sentences in Tokunoshima dialect are paired with those in standard Japanese. Each phrase in a sentence is used as a key, so that if a sentence consists of four

phrases, it is repeated four times.

As you can easily guess, KWIC reference data is very voluminous compared with original data. If an average sentence has four phrases, KWIC reference data is four times larger than the original data.

Here in our DVD, key words are sorted in four different orders, which make more data. It is impossible for KWIC reference data to fit into book format. DVD is a very convenient solution.

The phrase level KWIC reference data along with sets of phrases is a very powerful tool. We can obtain nouns and verbs from the normal order KWIC. It is very convenient for deriving verb conjugation. Fukushima completed her study of verb conjugation with the aid of KWIC reference, too.

With the reverse order KWIC, we can obtain a list of postpositions. We have discovered three points about postpositions.

1: Postpositions have two alternate forms.

2: The postposition *du* is used only once in a sentence or never.

3: The accusative case is rarely designated by postpositions.

2.4 Phonetics

a. A list of syllables with mp3 sound files

b. A frequency list of syllables that appeared in TTS in Tokunoshima dialect

Theoretical occurrences of syllables are described in Sibata's work (1960), but I know none of frequency list of syllables in Tokunoshima dialect ever described in academic papers.

The result looks very interesting. Syllables which have no cognate in standard Japanese are sparsely used.

2.5 Conjugation table of verbs and adjectives

Fukushima's laborious work which was completed with the aid of KWIC reference

data. The pdf files include audio tracks for some forms.

2.6 List of substantives

Japanese substantives appeared in standard Japanese part of TTS with numbers of sentences. This list can be used like a Japanese-Tokunoshima dictionary. For example, if you need to know where the word for fish is used, you only have to look for *sakana* 'fish' in standard Japanese and you will know the word is used in sentences number 1215, 1232, and 1233.

2.7 Making of DVD

The "d_and_p" folder of the DVD includes various data, computer programs, and tutorials for programs and the programming language AWK in which the programs are written.

a. TTS with at signs (@) as delimiters which break up sentences into phrases
This is the source data for the KWIC reference data and the list of phrases.

b. Computer programs for making data described in §2.1, 2.2, 2.3, 2.4, and 2.6
Texts are converted into html files in §2.1 and 2.2. Syllables are extracted by computer and made into files in §2.4. Key words in standard Japanese are analyzed with a ready-made program which gave their pronunciation. An original program was necessary to produce the result in §2.6.

Individual programs are not too big or complex. I intended to work with combination of simple programs. The operations are not fully automatic. Sometimes a last finish of human touch was necessary. The goal is to make linguistically important results with minimum efforts in programming, not to make beautiful programs.

c. How to use the programs
Concise description of programs and execution order of programs is added here.

d. Tutorials for Programming language AWK

This was written for myself as memos but is also useful to understand how ask was used to make programs to mas TTST.

2.8 Mp3 data of TTS in Tokunoshima dialect

The “mp3” folder includes sound files linked with data files.

2.9 Appendix

a. Dictionary of Basic Vocabulary

1,000 Basic Words with pronunciation movie files for some of them

b. PDF files of printed version of TTST

Tokunoshima Dialect Dictionary of Two Thousand Sentences (2009 Revised Edition of Printed Version, whole PDF file, partial PDF file)

In addition to the data files of TTST, the files of the book format TTST are added as PDFs. Also, the research results on Basic Words are included.

3. Pros and cons of DVD

All the data and programs used to make the DVD are open to everyone. You can get the same final products (KWIC reference data, for example) following instructions in the DVD. To guarantee repeatability of the process is very important. Therefore, I am satisfied that we were able to show both the data and the programs.

I had to be content to break up sentences into phrases under several restrictions. First, we had only limited knowledge of Tokunoshima dialect. We were not sure we got morphemes accurately. Second, semantic correspondence between Tokunoshima dialect and standard Japanese is maintained only when we deal with phrases. Things are different at the morpheme level. For example, two postpositions in Tokunoshima

dialect often correspond to one in standard Japanese. Therefore, the morpheme level one-to-one semantic correspondence cannot be expected.

Phrase-level KWIC reference has shortcomings and limitations. Typical phrases consist of a substantive and postposition(s). Thus we have a group of phrases containing the same substantive gathering in one place in normal order KWIC reference and a group of phrases containing the same postposition in reverse order KWIC reference. But when a standard Japanese phrase contains two postpositions, we can get only one postposition following another in Tokunoshima dialect. The postposition preceding another is hidden inside. Also, when we try to find syntagmatic relations of morphemes, the phrase level KWIC reference is ineffective.

4. On-going task

Now I am in the process of making the xml-like tagged data. We have accumulated the knowledge on Tokunoshima dialect grammar, so I can extract morphemes with little hesitation. A tag is used to indicate parts of speech, number of the morphemes (we have listed up all the morphemes appearing in TTS and numbered them), conjugation (in case of verbs and adjectives), and so on.

I believe that making xml-based tagged data will give us a good insight into Tokunoshima dialect.

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.

OKAMURA, Takahiro; Motoei SAWAKI; Yumi NAKAJIMA; Chitsuko FUKUSHIMA and Satoru KIKUCHI (2009) *Tokunoshima hogen nisenbun jiten kaiteiban* [Tokunoshima Dialect Dictionary of Two Thousand Sentences: Revised edition]. Matsumoto: The Association of Tokunoshima Dialect.

SIBATA, Takesi (1960) "Phonology of Tokunoshima Dialect", *Kokugogaku*, 41.