THE DESIGN OF THE NEWSPAPER-BASED JAPANESE LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION CORPUS

Katunobu ITOU [1] *Mikio YAMAMOTO* [2] *Kazuya TAKEDA* [3]

Toshiyuki TAKEZAWA [4] *Tatsuo MATSUOKA* [5] *Tetsunori KOBAYASHI* [6]

Kiyohiro Shikano [7] *Shuichi ITAHASHI* [2]

[1] ETL, [2] Univ. of Tsukuba, [3] Nagoya Univ., [4] ATR, [5] NTT, [6] Waseda Univ., [7] NAIST, Japan

ABSTRACT

In this paper we present the first public Japanese speech corpus for large vocabulary continuous speech recognition (LVCSR) technology, which we have titled JNAS (Japanese Newspaper Article Sentences). We designed it to be comparable to the corpora used in the American and European LVCSR projects. The corpus contains speech recordings (60 hrs.) and their orthographic transcriptions for 306 speakers (153 males and 153 females) reading excerpts from the newspaper's articles and phonetically balanced (PB) sentences. This corpus contains utterances of about 45,000 sentences as a whole with each speaker reading about 150 sentences. JNAS is being distributed on 16 CD-ROMs.

1. INTRODUCTION

In the USA and Europe, effort such as ARPA (NAB)[1] and SQALE [2] have resulted in a large technology push in speaker independent, continuous speech recognition.

In Japan, the Acoustical Society of Japan (ASJ) Continuous Speech Corpora (ASJ-PB)[3] which contain about 10,000 PB sentences, have been widely used as a public resource for LVCSR research. However, we do not have a large text database; the main reason is that Japanese texts are written without spacing between words, and we do not have an adequate automatic word segmentation tool. For this reason, Japanese LVCSR systems are not well developed. Recently, however, progress with morphological analysis systems has enabled automatic segmentation to be used for learning of the statistical language models(SLM), and thus some LVCSR systems have begun to develop[4].

In Japan, we have been unable to compare different recognition methods and systems, because we did not have any common Japanese speech corpus for LVCSR research. To stimulate Japanese LVCSR research, we designed a Japanese speech corpus for LVCSR technology that is comparable to the corpora used for NAB and SQALE.

In developing the text database, we still have some language-dependent problems with training the language model. The main problem is that we do not have a general rule to separate text into words. One of the other problems is that, because Japanese text consists three character systems (*hiragana*, *katakana*, and *kanji* (Chinese characters)) there are a lot of variations of spelling.

These problems cause variations between morphological systems (grammar, lexicon, and so on) in Japanese natural language processing (NLP) research. Differences between morphological systems affect the word-frequency lists. For referential comparison, we need to normalize the morphological system or prepare a sharable referential tool as a public standard. We designed and developed the corpus after careful consideration of these points.

In constructing JNAS, the Large Vocabulary Continuous Speech Database Working Group (LVCSD-WG) of the Special Interest Group of Spoken Language Processing (SIG-SLP) of the Information Processing Society of Japan (IPSJ) designed and developed text sets for recording from 1995 to 1997, and the Speech Database Committee of the ASJ developed speech corpus in 1997. We do not plan any formal project for competitive evaluation such as ARPA or SQALE in our community, but we intended the corpus to be used as common reference data. **Fig. 1** shows the construction flow of JNAS. In this paper, we describe the specification and development of JNAS.

2. DESIGN AND DEVELOPMENT OF THE CORPUS

2.1. Japanese Language-Dependent Problems in SLM Training

Japanese text is not divided by white space at word boundaries. In Japanese text, we use *kanji* (Chinese orthography) and *kana* (phonetic characters). There are two types of *kana*: *hiragana* and *katakana*. *Kanji* consists of ideograms. When writing with ideograms, since the objects of linguistic expression are innumerable, there are also an extremely large number of characters. For example, the four years of newspaper text we used, contained more than 5,000 different characters. Many *kanji* have multiple readings: one reading is derived from the Chinese pronunciation, and the other is the "Japanese" reading of the Japanese word that corresponds to the meaning of the Chinese character–what is called *wago*. Many *kanji* have the same Chinese pronunciation. Therefore, it is very hard to disambiguate readings in the *"mixed kana-kanji"* style of writing. Moreover, in the Japanese language, we do not have general rules for what constitutes a single word, and verbs, adjectives, and other inflected words have many inflections.

Therefore, it is difficult to define a word unit, and there are a lot of variations between morphological systems. In making referential comparisons, we need to considered these points carefully.

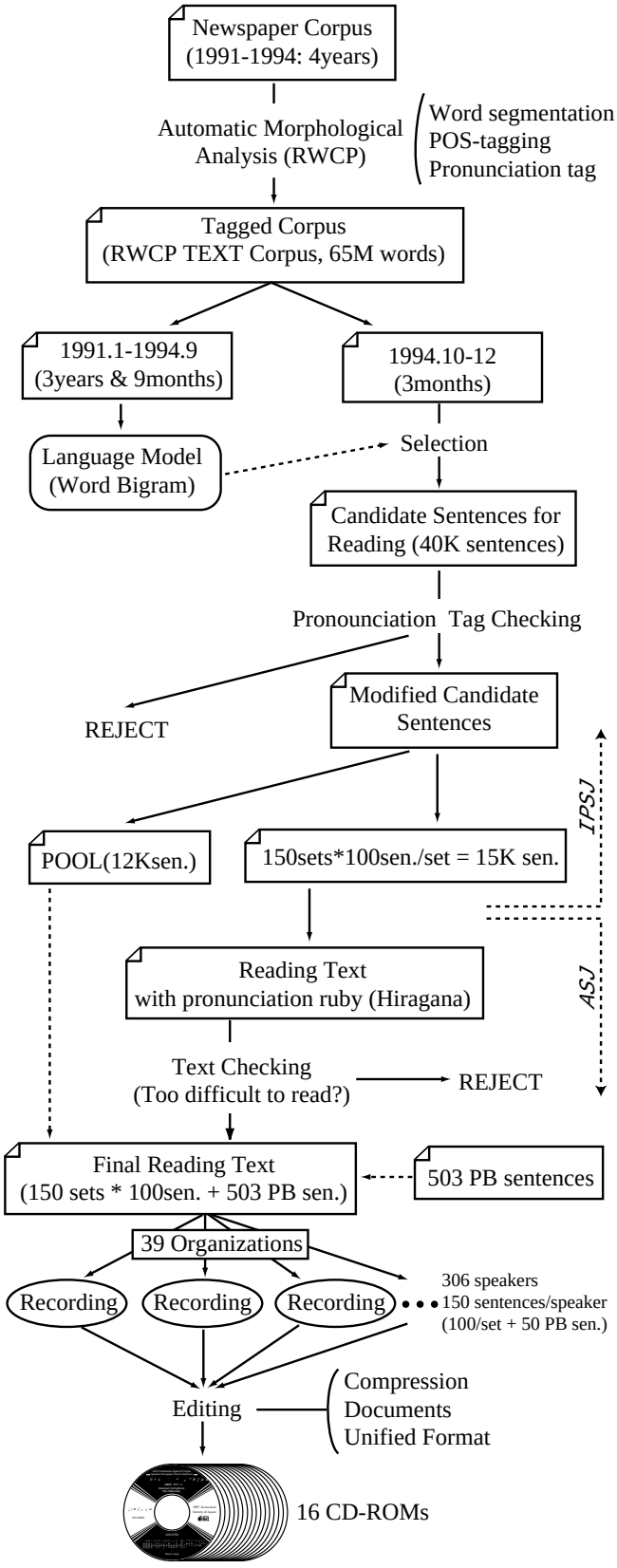

Figure 1: The construction flow of JNAS.

2.2. Text Preprocessing

First, we discussed which text to select for training and evaluation material. After we discussed which paper to select among some major dailies and a business paper, we decided to use the Maninchi Newspaper, one of the major Japanese dailies, because its copyright permission is most suited to our purpose of releasing the resultant corpus to the public.

Ideally, the text preprocessing should divide the sentence into words and resolve the ambiguities with all of the readings of the words. This preprocessing is similar to the type that might be used in a text-to-speech system. A text-to-speech systems' preprocessor, however, can only give the readings, and can not give any grammatical and/or morphological information, such as segmentation of words or the part of speech of the word, which is useful for constructing a language model for speech recognition. In the research community for Japanese natural language processing, a system called "morphological analysis system" is widely used, and the system estimates word segmentation, part of speech, and inflection.

However, estimation of the reading of the word is beyond the ability of the current morphological analysis system, because it is developed for text processing which needs not estimate the reading. Moreover, in Japanese we don't have any standard general rule for the definition of vocabulary, morphological grammar, or a system for parts of speech. Therefore, we fixed the goal of text preprocessing as analysis of a sentence by a morphological analysis system.

There were no public morphological analysis systems which had the ability to construct a language model, and so we decided to use the morphological tagged corpus of the Mainichi Newspaper which is distributed as the RWCP-Text-Corpus (RWC-DB-TEXT-95-1) by the RWCP (Real World Computing Partnership) as a standard text database for training of the language model.

An example of the RWCP-Text-Corpus is shown in **Fig. 2**. In the example, each line stands for a morpheme. The first column contains the notation for the morpheme, the second contains the basic form (dictionary form) of the morpheme, and the third contains the ID number of a part of speech (POS).

notation	original form	POS ID
G'<1	G'<1	1
$7	9k	63
J1$l	$J$$	445
$P	$P	422
Ji	Jk	276
$J$$	$J$$	443
!#	!#	468

Figure 2: An example of the data of the RWCP-Text-Corpus. The example sentence is "G'<1$7$J1lPJiJ $$!#" (I have to recognize).

The RWCP proposed a POS system called THiMCO (Tagset of High quality for Integrated Multi-usage Corpus Openly available to public). In the RWCP-Text-Corpus, THiMCO95 was used. THiMCO95 is a relatively detailed Japanese POS system and contains about 500 parts of speech.

The first step in preprocessing was to extract all of the article paragraphs as fields from the original CD-ROM with RWCP-Text-Corpus. An article has a specific ID number and consists of paragraphs. After extraction, each paragraph was numbered in order automatically (this number was used as the document control number) and collected into a file by month.

Next, paragraphs which had no period were removed automatically for readability filtering. Such paragraphs included poems, recipes, tables, lists, and so on. As another readability filter, sequences of morphemes between special symbols (for example, round brackets), which were automatically estimated as "unread" expression was removed. Finally, the paragraph was divided into sentences at periods or equivalent symbols. After sentence segmentation, each sentence was numbered in order. Thus, the sentence number was dependent on the morphological analysis system.

2.3. Sentence Selection for Recording

Next it was necessary to divide the text data into a training section and an evaluation section. The most recent three months' data (about 10% of the whole data) were selected for testing, and the rest of the articles covering 45 months (about 90% of the data) were reserved for training. The size of the corpus is shown in **Tab. 1**.

	91/1–94/9	94/10–94/12
# sentence	2,372K	194K
# paragraph	1,438K	139K
# article	282K	21K
# morpheme	65,347K	4,936K
vocabulary size	291K	97K

Table 1: text corpus

For classifying sentences, it was necessary to form a language model for selection of the sentences for recording. The first step was to form a word-frequency list (WFL, a frequency-ordered morpheme unigram list) from all of the training text with their morphological information.

To form a WFL, we needed to define a counting unit for word. As we mentioned above, it is reasonable that we treated a morpheme as a word. In this case, we have several choice from many definitions of counting units. We considered the following definitions: i) by notation, ii) by original form iii) by reading.

It seems that the third method is better one than any other choice, because we do not need to consider the treating of a word which has multiple readings. However, any sharable adequate text-to-reading translation tool or any large text corpus which has the information of reading is not available.

It seems that the second method is better one from the linguistic viewpoint. The first method is the most simple one. However, in this method, we do not distinguished the morphemes which have the same notation and the difference POSs or readings.

Finally, we defined the counting unit as a morpheme distinguished by all of the morphological attributes given by RWC Corpus (shown in **Fig. 2**), because the definition is the most precise and it is advantageous as the reference corpus.

We counted using this counting unit, it yielded a list of 291K words from all of the learning data. The frequency-weighted word coverage of the WFL is shown in **Tab. 2**.

Size	Coverage (%)
5K	85.8
8.1K	90.0
20K	95.7
27.6K	97.0
291K	100.0

Table 2: Frequency-weighted word coverage (from the word-frequency list)

Next, a word bigram language model was constructed to calculate the test-set perplexity of the sentence. The word bigram language model was generated using the CMU SLP Toolkit [5]. The language model was an open vocabulary backoff word bigram which was constructed with the cutoff set to 2, the discount strategy specified as "Good Turing discounting," and a vocabulary size of 20K words.

A statistically-controlled text set consists of 90 sentences (SC-sentences) collected from the 30 categories according to **Tab. 3** and about 10 sentences taken from a few paragraphs which consisted of only the three or more sentences which were satisfied in any category in **Tab. 3**.

Five other text sets were "article" sets. An "article" set consisted of three articles. Each article included 10 or more paragraphs that consisted only of sentences classified into any class in **Tab. 3**. Each paragraph contained 3-10 sentences. We didn't check for duplication of sentences between "sentence" sets and "article" sets.

vocaburary class	sentence length
MID	$6 \leq normal < 20 \leq long < 39$
LARGE	$6 \leq normal < 30 \leq long < 39$
	perplexity
MID	$0 \leq L < 40 \leq M < 85 \leq H < 400$
LARGE	$0 \leq L < 70 \leq M < 130 \leq H < 400$

sentence length	normal			long		
perplexity	L	N	H	L	N	H
MID	2	6	2	1	3	1
MID+	2	6	2	1	3	1
LARGE	4	12	4	2	6	2
LARGE+	2	6	2	1	3	1
LARGE++	2	6	2	1	3	1

MID = 5k voc.
MID+ = 5k voc. with one unknown word
LARGE = 20k voc.
LARGE+ = 20k voc. with one unknown word
LARGE++ = 20k voc. with two unknown words

Table 3: Distribution of perplexity (pp), sentence length, vocaburaly class for 90 sentences as selected for each speaker.

2.4. Recording

The speech data were recorded in collaboration with 39 sites, so the recording conditions and AD conversion char-

acteristics, including low-pass filter characteristics, were not unified. Each recording site collected data sets for 4-10 speakers (equal numbers of male and female speakers chosen). Each speaker read one set (about 100 sentences) from SC sentence sets, and one subset (about 50 sentences) from the ATR PB sentence sets using in ASJ-PB[3].

From the 150 SC sentence sets, 138 sets were read by both one male speaker and one female speaker, 4 sets were read by both of two male speakers, 4 sets were read by both of two female speakers, 2 sets were read by one male speaker, and 2 sets were read by one female speaker. At each of the recording sites, all of the speakers read the same PB sentence subset.

The utterances were recorded with two microphones simultaneously: a standard close-talk microphone (Sennheiser HMD410/HMD25-1 or equivalent) and a desktop microphone which was selected independently at each site (Sanken, Sony, and similar). The two versions of the data were stored in separate files.

Reading text was printed out to papers and the speakers read the text. Only kanji characters in reading sentences had ruby (readings) for easy to read. They were automatically generated from morphologically analysed sentences with readings using 'diff' command of UNIX which may use dynamic programming.

Each utterance was checked at each recording site. In the prompting text, each word was given a single reading. However, in Japanese, there are words which have several readings (i.e., "Japan" has two readings: *nihon*, and *nippon*). An orthographic transcription was created of any changes to readings made at each recording site. No changes to the content of the newspaper articles were permitted under the copyright permission, and the orthographic transcription was not modified for any other errors or variations. A list of these errors was collected in the check list file at each recording site.

3. SPEECH CORPUS: JNAS

The data described here was compiled into 16 CD-ROMs and titled JNAS (Japanese Newspaper Article Sentences). 9 Gigabytes by the "shorten."

The CD-ROMs have been released to the public. **Tab. 4** shows the final specification of JNAS. Since JNAS is a relatively large corpus, it is not error free. At this time, we know some errors such as reading error, A/D conversion error, disfluencies, lack of files, and so on. We plan to maintain these error information about JNAS on the WWW (`http://www.milab.is.tsukuba.ac.jp/jnas/`).

4. CONCLUSION

The JNAS corpus and its components have been designed and developed for LVCSR research by the joint efforts of the LVCSD-WG IPSJ and the Speech Database Committee of ASJ.

The Speech Database Committee of ASJ are now selecting text sets for referential evaluation. It plans a training set that contains 100 speakers × 100 sentences and an evaluation set that contains 25 speakers × 4 sentences.

To promote both research of component technologies and development of systems for LVCSR, we have recognized the necessity of a sharable software repository which includes recognition engines, acoustic models and language models. Thus, we are developing a Japanese Dictation Tool Kit[6], sponsored by the Information-Technology Promotion Agency (IPA), in Japan. In the project, we will also develop a tool to normalize the differences between morphological analysis systems.

#Reading text sets	Newspaper	155sets(16,176 sen.) !Jabout 100 sen./set)
	PB sen.	10sets (503 sen.) (about 50 sen./set)
#Speakers		306 (153 fe/males)
#Utteraces	Newspaper	31,938
	PB sen.	15,372
Recorded time of newspaper sentences		215,247 sec. (about 59h 47m)
#Recording site		39
Microphone	headset	common
	desktop	inconsistent

Table 4: The Specifications of JNAS.

ACKNOWLEDGMENTS

The prompting texts and the bigram language models for the Mainichi Newspaper article sentences were prepared by Akinori Ito (Yamagata Univ.), Takehito Utsuro (NAIST), Tatsuya Kawahara (Kyoto Univ.), Toru Shimizu (KDD), Masafumi Tamoto, Kazuhiro Arai (NTT), and Nobuaki Minematsu (TUT). We used the NIST SPHERE package to attach headers to the wave files and for the "shorten" compression technique used to reduce the number of CD-ROMs. The NIST SPHERE package was implemented by the Spoken Natural Language processing group, National Institute of Standards and Technology, U.S.A. The 'shorten' compression technique was developed by Tony Robinson at Cambrigde University and SoftSound Limited, UK. The speech data was collected by the efforts of many volunteers at the 39 research institutes.

We would like to thank all of the above groups and individuals.

REFERENCES

[1] D. B. Paul, et. al. The design for the wall street journal-based csr corpus. In *Proc. DARPA SNL Workshop*, pages 357–361, 1992.

[2] H. J. M. Steeneken, et. al. Multilingual assessment of speaker independent large vocabulary speech recognition system: the SQALE-project. In *EUROSPEECH*, pages 1271–1274, 1995.

[3] S. Hayamizu, et. al. Design and creation of speech and text corpora of dialogue. *IEICE Trans. Inf. & Syst.*, E76–D(1):17–22, 1993.

[4] T. Matsuoka, et. al. Japanese large-vocabulary continuous speech recognition using a business-newspaper corpus. In *Proc. of ICSLP*, pages 22–25, 1996.

[5] Ronald Rosenfeld. The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation. In *Proc. ARPA SLS Technology Workshop*, pages 47–50, January 1995.

[6] T. Kawahara, et. al. Sharable software repository for japanese large vocabulary continuous speech recognition. In *Proc. ICSLP-98*, 1998.