

‘BonTen’ – Corpus Concordance System for ‘NINJAL Web Japanese Corpus’

Masayuki ASAHARA◇

Hideto MASUOKA♠

Toru MORII♡

Sachi KATO◇

Kazuya KAWAHARA♠

Yasuko OHBA♡

Yuki TANAKA♡

Yuya TAKEI♠

Yuki TORII♡

Kikuo MAEKAWA◇

Hikari KONISHI◇

◇ National Institute for Japanese Language and Linguistics,
National Institutes for the Humanities, Japan
♠ Retrieva Inc., ♡ Everyleaf Corporation

Abstract

The National Institute for Japanese Language and Linguistics, Japan (NINJAL) has undertaken a corpus compilation project to construct a web corpus for linguistic research comprising 25 billion words. The project is divided into four parts: page collection, linguistic analysis, development of the corpus concordance system, and preservation. This article presents a corpus concordance system named ‘BonTen’, which enables a ten-billion-scaled corpus to be queried by string, a sequence of morphological information or a subtree of the syntactic dependency structure.

1 Introduction

The National Institute for Japanese Language and Linguistics, Japan (NINJAL) has compiled a ten-billion-word scale Japanese web corpus named ‘NINJAL Web Japanese Corpus’ (hereafter ‘NWJC’)(Asahara et al., 2014). This paper presents the web-based corpus concordance system ‘BonTen’ – *Brahman*¹ for NWJC. The system designs are based on the string search mechanisms of the web-based search system ‘Shonagon’², which is used to search ‘the Balanced Corpus of Contemporary Written Japanese’ (hereafter ‘BCCWJ’)(Maekawa et al., 2014). Shonagon enables a short unit sequence search to be carried out on the web-based corpus concordance system ‘Chunagon’ for BCCWJ, and a dependency search to be performed on the corpus management system ‘ChaKi.NET’ (Matsumoto et al., 2005; Asahara et al., 2016). Because the system functions as a web application, the user only requires a web browser to access the corpus. The user interface design³ is based on that of ‘ChaKi.NET’. The back-end search system is based on ‘Sedue for Bigdata’ and was developed by Retrieva Inc.⁴. The search system can effectively search the ten-billion-word scale corpora at practical speeds.

2 ‘NINJAL Web Japanese Corpus’ (NWJC)

NWJC is a web corpus for Japanese linguistic research comprising ten billion words.

Page collection is performed by employing remote harvesting using the Heritrix crawler.⁵ Our web crawler processes one hundred million URLs every three months to provide fixed-point observations for one year. The list of URLs is changed annually.

In linguistic analysis, we perform *normalisation*, *Japanese morphological analysis*, and *Japanese dependency analysis*. The crawled pages are normalised by nwc-toolkit-0.0.2⁶ to remove HTML tags and convert character encoding, after which they are split into sentences. The periods (Kuten), exclamation marks, and question marks are removed during the sentence splitting process. Sentences are collected according to types rather than by tokens to alleviate duplication issues on the web. Sentences are paraphrased for type unification purposes by using the `uniq` command (with `sort`). We use the MeCab-0.996

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹The name originated from a form of ‘Bengalese finch’.

²<http://www.kotonoha.gr.jp/shonagon/>

³the user interface is developed by Everyleaf Corporation <http://everyleaf.com/>

⁴<http://retrieva.jp/>

⁵<http://webarchive.jira.com/wiki/display/Heritrix/Heritrix/>

⁶<http://code.google.com/p/nwc-toolkit/>

morphological analyser ⁷ and the UniDic-2.1.2 dictionary ⁸ to conduct Japanese morphological analysis (word segmentation and POS tagging). We also use the dependency analyser CaboCha-0.69 ⁹ with UniDic head rule options ¹⁰ for Japanese dependency parsing.

We preserve the data collected for linguistic studies to monitor any future changes. The corpus is archived in WARC format (ISO 28500:2009) by the Heritrix crawler. The data will be harvested in online backup as open source wayback software and in offline backup as LTO tapes.

Japanese copyright law prevents us from making the corpus publicly available. However, the data will be accessible by search engine to enable the user to search for query strings, morphological information, and syntactic dependency subtrees. The result shows the links to the original pages. The search engine includes data crawled from October 2014 to December 2014 (2014-4Q) by the types of sentences. The statistics for the data are provided in Table 1.

Table 1: Statistics :2014-4Q

No. of URLs	83,992,556
Tokens of Sentences	3,885,889,575
Types of Sentences	1,463,142,939
No. of Bunsetsus	8,736,741,719
No. of Short Unit Words	25,836,947,421

3 Functions of the corpus concordance system ‘BonTen’

This section describes the functions of the BonTen corpus concordance system ‘BonTen’. We aimed to utilize existing user interface designs for query building. However, some functions such as regular expression matching are omitted in order to work on a ten-billion-word scale corpus.

Below, we introduce three query systems: string search, short unit search, and dependency search. We also present the display of retrieval results.

3.1 String Search

A string search is the most fundamental query function for accessing the corpus and returns sentences that include the exact same query string. The result can be refined by the last two parts of the URL domain.¹¹

BonTen cannot use regular expressions such as concatenation, alternation, and Kleene star, because the length of Japanese words tends to be shorter than other languages.

Because the data is stored in the form of sentences, we cannot throw a query across sentence boundaries. Although some preceding systems use a history function to list previous user queries, this function was not implemented in BonTen. The history function will not be implemented until the end of 2016.

We demonstrate the two formats in which retrieval results are displayed. Authenticated users will be able to use the rich display. We present the display itself in Section 3.4.

3.2 Short Unit Search

The short unit search is a function that is very similar to the short unit search function in Chunagon and tag search function in ChaKi.NET. The function can throw queries based on sequences of morphological information. Figure 1 shows the on-screen display of an example query. The boxes correspond to the morphemes (short units) in the sentences. We can determine the relative positions of the boxes. Relative position ‘0’ denotes a centred word in the KWIC. A relative position with a negative value indicates the left side of the centred word. A relative position with a positive value indicates the right side of the

⁷<http://mecab.googlecode.com/svn/trunk/mecab/doc/>

⁸<https://osdn.jp/projects/unidic/>

⁹<https://taku910.github.io/cabochoa/>

¹⁰`./configure --with-posset=UNIDIC`

¹¹Internet top-level domain such as .com and .jp, and second-level domain such as .co.jp and ac.jp.

centred word. The two numbers on the box denote the relative positions between the boxes, in which the number on the left is the minimum (left-most) relative positions and the number on the right is the maximum (right-most) relative positions.

The box can specify the following morphological information: Surface form: (〈 表層形 〉) the form appearing in the sentence; POS1, POS2, POS3, POS4: 〈 品詞 1 〉, 〈 品詞 2 〉, 〈 品詞 3 〉, 〈 品詞 4 〉 parts of speech in the hierarchical tag; Conjugation type: 〈 活用型 〉; Conjugation form: 〈 活用形 〉; Lemma – reading: 〈 語彙素読み 〉 the reading form of entry data in the UniDic, which should be transcribed using Katakana; and Lemma: 〈 語彙素読み 〉 the writing form of entry data in the UniDic.

The POSs, Conjugation type, and Conjugation form are listed by clicking the ▾ icon.

Because the data is stored in sentence form, we cannot throw a query across sentence boundaries. Although some preceding systems use a history function to list previous user queries, BonTen has not implemented a history function. The history function will be not implemented until the end of 2016.

The boxes can be expanded by clicking ‘+’, reduced by clicking ‘×’, or cleared by clicking the eraser icon.

In the example, we specify a centred word of the POS type ‘Noun, Proper Noun, Place Name’ followed by the lemma ‘語 (*language*)’.



Figure 1: Short unit search query



Figure 2: Dependency search query

3.3 Dependency Search

The dependency search function is nearly the same as that in ChaKi.NET. The function can be used to search for ‘Bunsetsu (*Japanese base phrase*)’-based dependency structures. The query can be specified by providing a subtree of the Bunsetsu-based dependency structure, such as morphological information, the relative position in a Bunsetsu, the relative positions between Bunsetsus, and the dependency relation.

Figure 2 shows a screen displaying an example query. The green and orange boxes specify the Bunsetsus. The numbers on the upper left side of the colored boxes specify the ID of the Bunsetsu (to the left of the colon) and the ID of the head Bunsetsu (to the right of the colon). The ^ sign in the figure indicates the left (Bunsetsu) boundary. The – sign in the figure indicates that the two (morpheme) units are adjacent. The < sign in the figure indicates that the two (Bunsetsu) units occur in this linear order. We can also use the \$ sign as the right (Bunsetsu or Sentence) boundary. The + sign is for increasing the size of a morpheme or Bunsetsu box.

In this example, we define two Bunsetsus: the first Bunsetsu includes a word of the POS type ‘Noun, Proper Noun, Place Name’ as the left-most word followed by the lemma ‘語 (*language*)’, whereas the other Bunsetsu includes a word of the POS type ‘Verb’. The two Bunsetsus appear in this order and have a dependency relation.

3.4 Displaying the retrieval results

The retrieval results show the number of query hits, and the example sentences. Fifty example sentences are displayed using pagination. We prepared the following two displays of retrieval results. The first is a simple example in which only the sentences are rapidly displayed. The second is a rich example (Figure 3) in which sentences are shown together with the morphemes and Bunsetsus segmented according to the top level of the POS tag. In the figure, the mouse cursor is on the morpheme/Bunsetsu ‘する’ (in the yellow background) of the fourth example. The morphological information appears in a pop-up box

4	空白 空白	副詞	代名詞 助詞	た	え	われ	われ	が	名詞 名詞 助詞	名詞 名詞	名詞 助詞	接辞類 名詞 助詞 助詞 助詞 補助記号	こと	が	[不 得 手 と は い え 、]
	空白 空白	副詞	補助記号	名詞 名詞 助詞	名詞	["pos1"=>"動詞","pos2"=>"非自立可能","pos3"=>"*","pos4"=>"*","c_type"=>"サ行変格";									

Figure 3: Rich display of retrieval results

Table 2: String Search Evaluation

query	English translation	Hit number	Response Time (simple display)	Response Time (rich display)
さくら	cherry blossoms	137,680	0.717 sec.	11.387 sec.
フランス語	French	15,214	0.692 sec.	9.376 sec.
国立国語研究所	NINJAL	106	1.239 sec.	5.420 sec.
じゅげむじゅげむ ごこうのすりきれ	a phrase (in Buddhist scripture)	13	0.460 sec.	0.998 sec.

and is displayed against a white background. The Bunsetsus that appear against a blue background are dependants of those highlighted in yellow, whereas a Bunsetsu highlighted in pink is the head of the Bunsetsu highlighted in yellow.

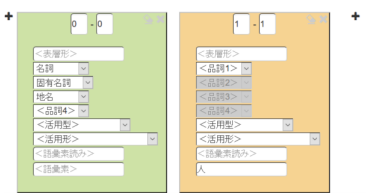
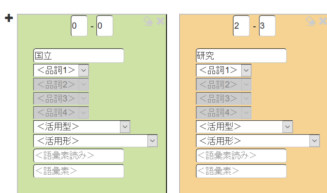

We also provide two services to download the results. The first enables 50 example sentences to be displayed with the morphological information and syntactic dependency structure in the format of the syntactic dependency parser CaboCha. The second service allows the user to download the retrieved sentences (max. 100,000 sentences) in tsv format (without any linguistic annotation). Both of these types of data are displayed with a URL list. The line ending code can be specified as CRLF (Windows), LF(Linux), or CR(Mac OS). The character encoding is fixed as UTF-8.

4 Evaluations

We evaluated the response time of the corpus concordance system using the Firefox 40.0.1 browser with HttpWatch Basic Version 10.0.44¹². The computer was connected to the Internet via optic fibre for home use and the evaluation was performed in March 2016.



Table 2 shows the results of the string search in two display modes. The simple display mode produces results reasonably quickly, with the network latency being the main cause of time loss. The rich display mode needs time to construct visualization of linguistic annotations, including morphological information and dependency structures. Tables 3 and 4 show the results of a Short Unit Search and Dependency Search, respectively. The dependency subtree search on the ten-billion-word scale web corpus takes less than one minute, and includes rich annotation information.

Table 3: Short Unit Search Evaluation

Query	+ [0-0] +		+ [0-0] [2-3] [3-4] +		
					
Hit number	412,763		2,067		
Response Time (rich display)	16.567 sec.		10.702 sec.		

¹²<https://www.httpwatch.com/>

Table 4: Dependency Search Evaluation

Query		
Hit number	15,641	2,704
Response Time (rich display)	47.904 sec.	9.304 sec.

The Demo Video: <https://youtu.be/jYxeLYbnd3k>

Both searches can only be carried out in the rich display mode. The first query in Table 3 is a pattern of words of the POS type ‘Noun, Proper Noun, Place Name’ as the left-most word followed by the lemma ‘人 (*people*)’. The second query in Table 3 involves a pattern containing one or two words (wildcards) between the word ‘National’ and the compound word ‘Research Institute’. Both queries need less than 20 seconds to fetch the results. The first query in Table 4 relates to a pattern consisting of two Bunsetsus with a dependency relation in the following linear order. The first Bunsetsu includes a word of the POS type ‘Noun, Proper Noun, Person Name’ as the left most word followed by the surface form ‘が³’ and the POS is a Case Particle (Subject marker). The second Bunsetsu includes a word of the POS type ‘Adjective General’. Because there are no lexicalized content words among the content words, the query takes nearly one minute to display a result. The second query in Table 4 again involves a pattern consisting of two Bunsetsus with a dependency relation in the following linear order. The first Bunsetsu includes the surface form ‘Prime Minister’ as the left-most word followed by the surface form ‘が³’ and the POS is a Case Particle (Subject marker). The second Bunsetsu includes a word of the POS type ‘Verb’. Thus, one lexicalized entry reduces the query time.

5 Conclusions

The paper presents the functions of the BonTen corpus concordance system. The system has the ability to process queries using strings, morphological information sequence, and by using a subtree of the dependency structure for the ten-billion scale web corpus.

Acknowledgments

The work reported here is a result of the “Choo-daikibo koopasu” (ultra-large scale corpus) project of the Center for Corpus Development, NINJAL (2011-2015).

References

- Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan. *Alexandria*, 25(1-2):129–148.
- Masayuki Asahara, Yuji Matsumoto, and Toshio Morita. 2016. Demonstration of ChaKi.NET – beyond the corpus search system. In *Proc. of COLING-2016 (Demo Session)*.
- Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371.
- Yuji Matsumoto, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Otani, and Toshio Morita. 2005. ChaKi: An Annotated Corpora Management and Search System. In *Proceedings from the Corpus Linguistics Conference Series*.