Keyaki Treebank segmentation and part-of-speech labelling

Stephen Wright HORN* Alastair BUTLER* Kei YOSHIMOTO†
* National Institute for Japanese Language and Linguistics
†Graduate School of International Cultural Studies, Tohoku University

horn.s.w@ninjal.ac.jp ajb129@hotmail.com kei@compling.jp

Abstract

This paper describes how the Keyaki Treebank, a corpus of Japanese language parsed for syntax, derives the segmentation and part-of-speech labelling used for generating trees. Aiming to expose basic functional structure while remaining fairly flat to ease search, the annotation policy selects for large terminal nodes, but not so large as to incorporate into purely lexical elements other elements with functional roles. This corresponds closely to, but also systematically deviates from, the LUW (Long Unit Word) standard of the Corpus of Spontaneous Japanese (CSJ; Maekawa 2003) and the Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa et al. 2014). The implementation applies post-processing to tools deriving LUW analysis (Mecab and Comainu), thereby feeding corpus construction. A new technique is employed to feed statistical parsing: A “scaffolding” script provisionally generates structures for the purpose of referring to structure for disambiguation of part-of-speech.

1 Introduction

This paper describes how the Keyaki Treebank, a corpus of Japanese language parsed for syntax, derives the segmentation and part-of-speech labelling used for generating trees. Originally proposed in Butler et al. (2012), this resource continues to grow as a component of the NINJAL Parsed Corpus of Modern Japanese (NPCMJ, `http://npcmj.ninjal.ac.jp`). In contrast to corpora designed primarily to capture morphological information, the Keyaki Treebank has the purpose of associating text with syntactic structure based on interpretations of the meanings of sentences, including such things as constituency, grammatical role, scope, anaphoric relation, focus, quantification, null pronominalisation, etc. The data model stresses economy in the use of categories and accessibility to search.

The first task in the production of a syntactically parsed corpus is the assignment to linguistic items of terminal node labels that indicate as closely as possible the syntactic functions of those items in a sentence. The approach starts with morphological analyses generated by machine parsers, and optimises the supplied information, adopting those analyses that serve to articulate structure, ignoring levels of analysis that have no syntactic consequence, and adding analyses where needed.

The paper is structured as follows. Section 2 presents policy followed for resolving questions of 1) segmentation, and 2) mappings between categories of segments, and sketches the implementation used for corpus construction. Section 3 addresses scenarios that limit the size of segments. Section 4 discusses some of the possibilities and implications suggested by the employment of the techniques introduced here.

2 Policy and implementation

The policy for segmentation and part-of-speech labelling assumed by the Keyaki Treebank follows the principle of using terminal nodes that are as large as possible, but not so large as to incorporate into purely lexical elements other elements with functional roles.

On the whole, such a policy corresponds closely with the LUW (Long Unit Word) standard of the Corpus of Spontaneous Japanese (CSJ; Maekawa 2003) and the Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa et al. 2014). An LUW is composed of at least one SUW (single-morpheme Short Unit Word), but complex LUWs containing more than one SUW are common.

The SUW based analysis is obtained with the parser Mecab (Kudo et al. 2004) using the UniDic dictionary (Den et al. 2008). The Comainu parser (Kozawa et al. 2014) adds an extra layer of analysis, “chunking” multiple SUWs into a single complex LUW depending on co-occurrence relations of the string in question.

The chunking from Comainu is not limited to complex nominal expressions or complex predicates: Heterogeneous strings that appear to have undergone grammaticalisation (e.g., some formal noun/particle pairs, some complex modal expressions, etc.) are chunked as well. Complex LUWs are usually incorporated into the Keyaki Treebank just as single segments. For example, numerals are analysed digit-by-digit into component SUWs by Mecab, but the Keyaki Treebank strings these into a single segment according to the unit containing them, which Comainu assigns.

While the chunking from Comainu is intended to identify units with significance in syntax, the information is not always rich enough to generate immediate constituency trees approaching descriptive adequacy for syntax. Depending on the circumstances, SUWs may need to be split, and LUWs may need to be concatenated under one terminal node label (an instance of further chunking). Furthermore, some finer distinctions in morphological analysis that have no consequence for syntax are sometimes ignored, while other distinctions deemed important are introduced. This is a consequence of the Keyaki Treebank aiming to expose the basic functional structure of the language, while keeping the structure fairly flat and easily searchable.

Results of the initial analyses from Mecab and Comainu are collected in the M-XML (morphology-based XML) format of the BCCWJ. A subsequent "rewrite" to strings of text paired to terminal node labels is accomplished with an XSL script, and then further manipulated by a script written in Tsurgeon language (Levy and Andrew, 2006) that performs "scaffolding". Based on the types and orderings of the terminal node labels, the scaffolding process builds phrase structures that allow for further decisions about segmentation and the assignment of terminal node labels to be implemented. Assembled phrase structure is subsequently removed to once again leave strings of text paired to terminal node labels. These results are then sent to a statistical parser. The full parser pipeline is available from: `http://www.compling.jp/haruniwa2`.

To illustrate the process just described, consider the following example:

(1) それにしても契約を結ぼう
"Nevertheless, let's sign a contract"

Sending (1) through Mecab and Comainu and changing to the M-XML format results in the following analysis (simplified here to show essential lemma, part-of-speech, and inflection ('cForm') information):

```
<sentence>
  <LUW l_lemma="其れ" l_pos="代名詞">
    <SUW lemma="其れ" pos="代名詞">それ</SUW>
  </LUW>
  <LUW l_lemma="にしても" l_pos="助詞-接続助詞">
    <SUW lemma="に" pos="助詞-格助詞">に</SUW>
    <SUW lemma="為る" pos="動詞-非自立可能"
         cForm="連用形-一般">し</SUW>
    <SUW lemma="て" pos="助詞-接続助詞">て</SUW>
    <SUW lemma="も" pos="助詞-係助詞">も</SUW>
  </LUW>
  <LUW l_lemma="契約" l_pos="名詞-普通名詞-一般">
    <SUW lemma="契約"
         pos="名詞-普通名詞-サ変可能">契約</SUW>
  </LUW>
  <LUW l_lemma="を" l_pos="助詞-格助詞">
    <SUW lemma="を" pos="助詞-格助詞">を</SUW>
  </LUW>
  <LUW l_lemma="結ぶ" l_pos="動詞-一般">
    <SUW lemma="結ぶ" pos="動詞-一般"
         cForm="意志推量形" >結ぼう</SUW>
  </LUW>
</sentence>
```

Note the chunking executed by Comainu in the second LUW in the example above, creating a conjunctional particle by concatenating the particle-verb-particle-particle sequence に-し-て-も. The XSL script inherits this. The XSL script also assigns the syncretic category "NV" (noun or verb) to 契約 (keiyaku, 'contract') as UniDic assigns this item @pos="noun-common.noun-'sa'.irregular.verbal.morphology.possible". For the scaffolding, NV will be a verb when followed by verbal morphology, and a noun when contained in a "PP" (particle phrase). The category "VB-VOL" (verb-volitional) is assigned to 結ぼう based on the cForm="volitional.suppositional.form". This reinterpretation of the M-XML is expressed as the following sequence of pairings:

```
それ        PRO
にしても    P-CONJ
契約        NV;契約する
を          P-CASE
結ぼう      VB-VOL;結ぶ
EOS
```

This sequence of pairings is converted to the basic tree structure below. At this stage all items are immediately dominated by a sentence node "IP" (inflectional phrase).

```
(IP (PRO それ)
    (P-CONJ にしても)
    (NV;契約する 契約)
    (P-CASE を)
    (VB-VOL;結ぶ 結ぼう))
```

This initial structure is modified with the scaffolding script, forming the constituency tree below based solely on the terminal node labels and their order of occurrence. Note that at this stage the scaffolding projects more phrases than just IP: The particle を projects a PP with the simple "NP" (noun phrase) 契約 as its complement: Then the script decides for 契約 the category of noun based on its containment in the PP projected by particle を. Also note how the sentence-initial sequence of (PRO それ) and (P-CONJ にしても) is further chunked by the scaffolding script. Finally, the script splits 結ぼう into two parts (VB, 'verb' and MD, 'modal') to reflect (to the extent that Japanese orthography allows) the fact that the whole word ends with a modal flective.

```
(IP (CONJ それにしても)
    (PP (NP (N 契約))
        (P-CASE を))
    (VB;結ぶ 結ぼ)
    (MD う))
```

In this way the scaffolding stage builds phrase structure by creating NP, PP, and IP projections; it disambiguates parts-of-speech by reference to structural position; it merges and splits words depending on their distribution or their morphology. In fact the sole reason for building phrase structure at this stage is to provide contextual information to disambiguate parts of speech more accurately. Context-sensitive manipulation of data at this stage is a powerful way to improve the performance of statistical parsers. To the best of the authors' knowledge, using a scaffolding script to feed statistical parsers is the first implementation of this kind. Once the desired terminal node labels are in place, all phrases are stripped away and part-of-speech labels are simplified, reducing the data to the following pairing:

```
それにしても  CONJ
契約          N
を            P
結ぼ          VB
う            MD
EOS
```

This is the form of the data that is passed to a statistical parser, with currently the Berkeley parser (Petrov and Klein 2003) being used. The Berkeley parser is capable of generating trees of great complexity and with deep embeddings. The statistically parsed result derived from our short example in (1) takes the form of a binarised tree with constituency calculated.

```
(IP-MAT (CONJ それにしても)
        (IML (PP (NP (N 契約))
                 (P を))
             (IML (VB 結ぼ)
                  (MD う))))
```

Further post-processing flattens the structure by removing "IML"(intermediate level) nodes, adds null pronouns (e.g., `*pro*`), adds functional/grammatical information for NPs (e.g., the extended `NP-SBJ`), and disambiguation information for PPs (e.g., `(NP-OB1 *を*)`), etc., to produce a structure signalling the grammatical contribution of every constituent.

```
(IP-MAT (NP-SBJ *pro*)
        (CONJ それにしても)
        (PP (NP (N 契約))
            (P を))
        (NP-OB1 *を*)
        (VB 結ぼ)
        (MD う))
```

Ultimately the sequence of bracketed trees produced by this pipeline are hand-corrected by annotators using interpretations of meaning and knowledge of grammatical patterns. The trees are organised along basic principles of linguistic structure and linguistic processes (projection of phrases, selection, modification, movement, pronominalisation, scope, etc.). Once the hand-annotation is completed, the trees and the text together instantiate a basic descriptive grammar of the Japanese language. They also provide the basis for a formal semantic representation to be generated.

3 Limitations on chunking

The policy for the Keyaki Treebank is to chunk as large as possible, but there are limitations to the chunking of strings. When there is clearly some constituency in a string that must be expressed by structure, or when there is a need to indicate the semantic effects of structure, chunking is not carried out. In fact, occasionally the Keyaki Treebank undoes the chunking executed by Comainu out of consideration for these two factors. To exemplify the former situation, consider the morpheme 中 (tyuu,'middle') which is frequently analysed as a suffix by Mecab and grouped together with a preceding string by Comainu: In contexts ambiguous for morphological parsers, reference to non-adjacent material and sentence meaning sometimes indicates that 中 is a formal noun.

Here is the LUW containing 旅行 (ryokou, 'travel') and 中 (tyuu,'middle') to form 旅行中 ('in the middle of travelling') as it appears in one BCCWJ analysis:

```
<LUW l_lemma="旅行中" l_pos="名詞-普通名詞-一般">
  <SUW lemma="旅行" pos="名詞-普通名詞-サ変可能"
     pron="リョコー">旅行</SUW>
  <SUW lemma="中" pos="接尾辞-名詞的-副詞可能"
     pron="チュー">中</SUW>
</LUW>
```

Referring to the UniDic dictionary, Mecab assigns to 中 @pos="suffix-noun.like-adverb.possible" and to 旅行 @pos="noun-common.noun-'sa'.irregular.verbal.morphology.possible," Comainu groups "suffix" 中 together with "noun" 旅行 in the same LUW, an analysis which results in the Keyaki Treebank's initially ambiguous assignment of NV 旅行 being decided as an N. However, it is clear from the syntactic context in (2) below that a verb of motion selects the accusative-marked object NP 海外を ('overseas ACC') with the semantic role of "path," and while verbal morphology is absent from 旅行, it is clear that 旅行 is what is doing the selecting.

(2) 現在は海外を旅行中だ
"(He) is presently travelling overseas."

Syntactic tests (e.g., negative concord) can demonstrate that an argument of a "verbal noun" in a construction like this is not local to the だ (da, 'is') that follows 中. Furthermore, the possibility of case-marking on 中 in analogous contexts shows that 中 heads a noun phrase. A principled explanation for the pattern in (2) above is that

VB 旅行 heads a gapless relative clause (IP-EMB) modifying N 中: 中 is not a suffix, but rather a formal noun with its own segment and terminal node label; 旅行 is not a noun but rather a verb. Neither morphological analysis alone nor (at present) scaffolding is equipped to carry out this analysis: Only hand annotation can decide for the structure that appears in the Keyaki Treebank below:

```
(IP-MAT (NP-PRD (IP-EMB (NP-SBJ *pro*)
                        (PP (NP (N 海外))
                            (P を))
                        (VB 旅行))
                (N 中))
        (AX だ)
        (PU 。))
```

This is one clear example of syntactic principles trumping morphological ones in a parsed corpus. A more sophisticated scaffolding script might be able to produce such analyses, reducing the burden on human annotators. The possibilities for the scaffolding technique have yet to be extensively mined.

4 Conclusion

To sum up, this paper has detailed segmentation and part-of-speech labelling for the Keyaki Treebank. This has included descriptions of policy as well as components of an implementation used for corpus construction. The rewriting of morphological annotation from the mechanical parsers has been discussed briefly. The new technique of scaffolding has been introduced, and some of its possibilities have been suggested. The Keyaki Treebank undertaking is still actively developing these techniques in order to take fullest possible advantage of the morphological analyses provided by Mecab and Comainu. This situation reflects the richness of those analyses as much as it does the incipient state of our battery of post-processing steps.

For the NPCMJ (an extension of the Keyaki Treebank, with an XML encoding to accommodate information that is additional to a core syntactic parse), UniDic morphological information created with Mecab is linked to each segment, and can be referred to with user interfaces (`http://npcmj.ninjal.ac.jp/interfaces`). In the process of recognising and expressing grammatical structure, apparent mismatches between the UniDic analyses and the part-of-speech analyses of the Keyaki Treebank arise, and these are easily retrievable from the NPCMJ. The ability to assign grammatical categories and structures to text at this level of delicacy is largely thanks to the morphological analysis, and mismatches in the Keyaki Treebank are more in the way of additions of functional information than they are indications of original mis-assignment. In the end, all the described decisions about segmentation and labelling are made in the service of exposing the roles of constituency and syntactic processes in the compositional expression of sentence meaning.

Acknowledgements

We thank Beatrice Santorini for the original idea of scaffolding. This paper describes work funded as a component of the NINJAL Parsed Corpus of Modern Japanese. We thank all project members for their help and advice, especially Shinya Okano for his assistance in developing the scaffolding script.

References

Butler, Alastair, Zhu Hong, Tomoko Hotta, Ruriko Otomo, Kei Yoshimoto, and Zhen Zhou. 2012. Keyaki treebank: phrase structure with functional information for Japanese. In *Proceedings of Text Annotation Workshop*.

Den, Yasuharu, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 1019–1024. Marrakech, Morocco: European Language Resources Association (ELRA).

Kozawa, Shunsuke, Uchimoto Kiyotaka, and Yasuharu Den. 2014. BCCWJ に基づく長単位解析ツール Comainu. In *Proceedings of the Twentieth Annual Meeting of the Association of Natural Language Processing*, pages 582–585. Kyoto, Japan: The Association for Natural Language Processing.

Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *In Proc. of EMNLP*, pages 230–237.

Levy, Roger and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structure. In *5th International conference on Language Resources and Evaluation*.

Maekawa, Kikuo. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*. Tokyo.

Maekawa, Kikuo, 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(2):345–371.

Petrov, Slav and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411.