Keyaki Treebank: phrase structure with functional information for Japanese

Alastair Butler[*][†], Tomoko Hotta[‡], Ruriko Otomo[†], Kei Yoshimoto[†][‡], Zhen Zhou[‡], Hong Zhu[‡]
[*]PRESTO, Japan Science and Technology Agency
[†]Center for the Advancement of Higher Education, Tohoku University
[‡]Graduate School of International Cultural Studies, Tohoku University
ajb129@hotmail.com, tomorc@hotmail.com, ruri660@gmail.com
zhuhong200812@yahoo.co.jp, kei@compling.jp, syusin3@yahoo.co.jp

1 Introduction

This paper describes our project of building a Treebank for Japanese following the *Annotation manual for the Penn Historical Corpora and the PCEEC* (Santorini 2010) with adaptations appropriate for Japanese. In addition to other motivations for building treebanks (to investigate syntactic phenomena, evaluate theories, extract grammars, train and test parsers, etc.), we have as a key goal the ambition to build a syntactic base able to support automatically deriving meaning representations for formal semantics research on a corpus linguistics scale.

The paper is structured as follows. Section 2 offers background on existing treebanks that have directly influenced our decisions. Section 3 introduces and illustrates the adopted parsing scheme. Section 4 sketches our method of treebank construction. Section 5 outlines current and planned content for our treebank. Section 6 considers maintaining compatibility with existing treebank resources for Japanese. Section 7 concludes.

2 Background

Trees parsed following the Treebank II Annotation Style for the Penn Treebank of English (Bies, Ferguson, Katz, and MacIntyre 1995) offer a level of representation which allows automatic determination of the main predicate, the logical subject, the logical object, as well as other arguments and adjuncts. This is aided by co-indexed null elements in "underlying" syntactic positions and notation for recovering discontinuous constituents.

Extending semantic representation beyond the predicate-argument level is more of a challenge with the Penn Treebank scheme, requiring construction specific work arounds to determine options such as whether structure is embedded or coordinated, a situation that is greatly alleviated with the modifications of the Penn Historical Corpora scheme outlined in the next section. Nevertheless the Penn Treebank scheme, together with related schemes developed for Arabic (Bies and Maamouri 2003), Chinese (Xue and Xia 2000) and Korean (Han, Han, and Ko 2001), serves as an excellent benchmark for the information a treebank must encode to enable the accurate recovery of semantic structure from parsed syntactic structure.

The dominant treebanks for Japanese, most notably the Kyoto Text Corpus (Kurohashi and Nagao 2003), are bunsetsu dependency based. Introduced by Hashimoto (1934) a bunsetsu is a phrasal unit consisting of one or more adjoining content words (noun, verb, adjective, etc.) and zero or more functional words (postposition, auxiliary verb, etc.). A bunsetsu dependency analysis involves segmenting the sentence into bunsetsu and establishing modifier (dependence on) relations between the bunsetsu to reveal information about sentence internal structure. In addition with version 4.0 of the Kyoto Text Corpus (Kawahara, Sasano, Kurohashi, and Hashida 2005) a subset of 5,000 sentences are annotated with case, anaphora and coreference information. The addition of case frame information into a bunsetsu dependency analysis offers essentially the equivalent in information content to what is found with the Penn Treebank, and extracting predicate-argument information is made as straightforward as reading off the case frame entries.

Problems with the bunsetsu dependency analysis begin as soon as one wishes to derive semantic information that is beyond the predicate argument level (see e.g., Butler, Zhou, and Yoshimoto 2012). While the dependency analysis is very good at telling us where structure should go, there is no information about how structure should be combined. Consequently every complex sentence is rendered structurally ambiguous in multiple ways. In principle it should be possible to harness case information to resolve some of the ambiguity. However being entirely indexed based, bunsetsu dependency structure is remarkably resistant to modification, with there being no (easy) systematic way to supplement information beyond the level of individual bunsetsu. A simple modification can begin a ripple effect that requires changes throughout the bunsetsu dependency structure. In short one can see a lot of information is present, but the notation blocks access.

3 Penn Historical Corpora scheme

Syntactic annotation depends on a clear parsing scheme to determine the standard for annotation. In creating the Keyaki treebank we aim to follow the *Annotation manual for the Penn Historical Corpora and the PCEEC* (Santorini 2010), hereafter referred to as the annotation system.

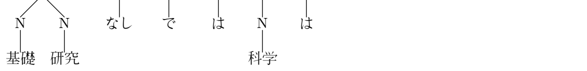
With appropriate cross-linguistic amendments, the annotation system has been applied to develop substantial treebanks for Old English (Taylor, Warner, Pintzuk, and Beths 2003, Pintzuk and Plug 2002), Early English (Taylor, Nurmi, Warner, Pintzuk, and Nevalainen 2006), Middle English (Kroch and Taylor 2000), Early Modern English (Kroch, Santorini, and Delfs 2004), Modern British English (Kroch, Santorini, and Diertani 2010), Historical French (Martineau, Hirschbühler, Kroch, and Morin 2010), Historical to Modern Icelandic (Wallenberg, Ingason, Sigurðsson, and Rögnvaldsson 2011), Historical Portuguese (Galves and Britto 2002), Ancient Greek and Yiddish. The existence of such a diverse body of languages annotated with essentially the same system is one more reason for wishing to see Japanese parsed in this manner, as this holds much promise of assisting with identifying and documenting crosslinguistic comparisons.

The annotation system is itself a modified Penn Treebank scheme, representing syntactic structure with labelled parentheses. All open parentheses have an associated label, representing nodes in a tree. These are either word-level labels (part-of-speech tags; N, ADJ, etc.) provided for every word, or phrase level labels that can indicate both form and function. In general, the basic label indicates the form of the constituent (NP, PP, ADJP, etc.), while additional labels (separated by a hyphen) indicate function (NP-SBJ = subject, ADVP-TMP = temporal adverb, CP-REL = relative clause, IP-INF = infinitive, etc.). Not all constituents are marked for function; in most cases there is at most one additional label, but there may be more (IP-INF-PRP = purpose infinitive, IP-IMP-SPE = direct speech imperative, etc.).

Phrasal labels are not included in every case in which a fully labelled tree would require them. Intermediate levels of structure in the sense of X' theory (N', ADJ', etc) are never represented explicitly, and there is typically no VP level. As a consequence parsed structure includes multiply branching nodes and is generally flat. This can be seen with a parse for (1), which is shown as a tree in (2) and with labelled bracket notation in (3). It can be seen that IP-MAT (a matrix IP) immediately dominates all verbs (to be understood in a broad sense, including modals and auxiliaries) and sentence level constituents.

(1) 基礎研究なしでは科学は発展しない。
Science would not develop without basic research.

(2)



(3)
```
(IP-MAT (PP (NP (N 基礎)
                (N 研究))
            (ADJ なし)
            (AX で)
            (P は))
        (PP (NP (N 科学))
            (P は))
        (NP-SBJ *は*)
        (VB 発展)
        (VB0 し)
        (NEG-HIGH ない)
        (PU 。))
```

In accordance with the annotation scheme, PPs are not marked for function. In (3) the function tag -SBJ is used in the construct (NP-SBJ *は*) to indicate that the prior PP headed by は has the subject grammatical role. Without this annotation it is very difficult to systematically determine the grammatical role of a は introduced NP. Also -HIGH has been added, which is an extension over the annotation system, to indicate a wide scope for negation. The syntactic information of (3) serves as the basis for recovering the meaning representation of (4).

(4)
$$\exists x(\text{科学}(x) \wedge \neg\exists y e_1(\text{基礎研究}(y) \wedge \text{発展し}(e_1,\ x) \wedge \text{なしでは}(e_1) = y))$$

Being は marked, 科学 'science' is rendered with a scope that is outside the negation.

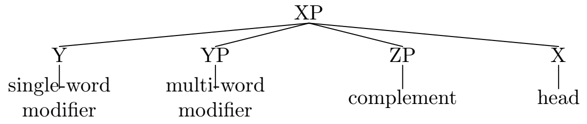
3.1 Advantages over the Penn Treebank scheme

The annotation system offers essentially the same level of information as the Penn Treebank to make the various aspects of predicate-argument structure easy to decode, including both function tags and markers of "empty" categories for displaced constituents.

Differences from the Penn Treebank scheme include cosmetic changes to give nodes of parsed trees labels that are more familiar to generative linguists. But there are also major changes: the VP level of structure is typically absent, the internal syntax of phrasal categories is fundamentally similar, and function is marked on all clausal nodes and all NPs that are clause level constituents, but not on PPs.

Having the internal syntax of phrasal categories fundamentally similar is a significant advantage in terms of allowing for a uniform exploitation of phrase structure. Barring some predictable exceptions, heads always project a phrasal node. In general the phrase head (N, P, ADJ, etc.) is overt and matches the category of the phrase level (NP, PP, ADJP, etc.). The phrasal node (NP, PP, ADJP, etc.) immediately dominates the phrase head (N, P, ADJ, etc.); that is, there are no intermediate bar-levels in the sense of X' theory. Thus with the annotation system, both modifiers and complements are sisters of the head, as pictured in (5).

(5)



There can be no fixed order to modifiers and complements. Rather required function tags allow for the determination of what is a multi-word modifier and what is a complement.

Most notable in regard to differences from the Penn Treebank is the function annotation that accompanies all clause nodes, which has been the killer reason for our adopting the annotation system. Thus matrix clauses are labelled IP-MAT, and they may be further characterised as direct speech (IP-MAT-SPE) or parentheticals (IP-MAT-PRN). Other IP clauses have their own labels, such as IP-IMP, imperative, IP-SMC, small clause, IP-PPL, participial, etc. All CPs also have extended labels to indicate type (CP-THT=that clause, CP-ADV=adverbial complement, CP-REL=relative clause, CP-QUE=question (direct or indirect), etc.).

As we are about to see with examples in section 3.2, marking clauses for function allows for a clear distinction between clauses that are integrated into a semantic representation as conjuncts (participial clauses, adverbial complements, relative clauses, etc.) and clauses that should be integrated as embeddings (infinitive complements, that complements, embedded questions, etc.). Having access to such information is essential when building semantic structures that go beyond the predicate-argument level. Having easy and systematic access is a huge bonus.

3.2 Examples

A complex sentence contains more than one clause, raising the issue of how the clauses combine to make up the sentence. Clauses may combine with coordinate conjunctions such as が 'but' or with the て-forms of verbs, adjectives or the copula meaning '∼ and', as in (6).

(6) 吉田さんは東京に行って鈴木さんに会った。
'Mr. Yoshida went to Tokyo and met Mr. Suzuki.'

A parse of (6) is given by (7), from which we can derive (8). This combines with conjunction the contributions of the clauses that make up (6). We know to combine with coordination because of the -PPL (= participial) function marking.

(7)
```
(IP-MAT (PP (NP (NPR 吉田)
                (NPR さん))
            (P は))
        (NP-SBJ *は*)
        (IP-PPL (PP (NP (NPR 東京))
                    (P に))
                (VB 行っ)
                (P て))
        (PP (NP (NPR 鈴木)
                (NPR さん))
            (P に))
        (VB 会っ)
        (AXD た)
        (PU 。))
```

(8) $\exists e_1 e_2$(past(e_2) ∧
行って(e_1，吉田さん) ∧
に(e_1) = 東京 ∧
会っ(e_2，吉田さん) ∧
に(e_2) = 鈴木さん)

A different way to combine clauses of a complex sentence is illustrated by (9), which receives the annotation of (10).

(9) 日本へ行きたいとトムは言っている。
'Tom says that he wants to go to Japan.'

(10)
```
(IP-MAT (CP-THT (IP-SUB (PP (NP (NPR 日本))
                            (P へ))
                        (VB 行き)
                        (AX たい))
                (P と))
        (PP (NP (NPR トム))
            (P は))
        (NP-SBJ *は*)
        (VB 言っ)
        (P て)
        (VB2 いる)
        (PU 。))
```

For (9) we want to build meaning representation (11). The function marking -THT with the CP determines that the clausal structures are combined with embedding, rather than coordination.

(11) $\exists e_1$言っている(e_1，トム，
$\exists e_2$(行きたい(e_2) ∧ へ(e_2) = 日本))

Note that we could not have relied on the presence of particle と in (9) to conclude the presence of an embedded clause, since と also has a subordinate conjunction function, as (12) demonstrates. The parse of (13) allows for the production of (14). Notably the -ADV (= adverbial complement) marking that accompanies the embedded IP indicates combining with と 'as' acting as a coordinating relation.

(12) その道を行くと彼に会った。
'As I went along the road, I met him.'

(13)
```
(IP-MAT (PP (IP-ADV (PP (NP (N その道))
                        (P を))
                    (VB 行く))
            (P と))
        (PP (NP (PRO 彼))
            (P に))
        (VB 会っ)
        (AXD た)
        (PU 。))
```

(14) $\exists x e_1 e_2$(past(e_2) ∧
その道(x) ∧
と(行く(e_1) ∧ を(e_1) = x，
$\exists y$(彼:y = ? ∧ 会っ(e_2) ∧
に(e_2) = y)))

In (15a) 昨日とった 'we took yesterday' is a relative clause that modifies 写真 'picture'. By contrast in (15b) 子供が泳いでいる 'a swimming child' is an embedded clause, and is the content of 写真 'picture'.

(15) a. 昨日とった写真がかかっていた。
‘The picture that we took yesterday was hung.’

b. 子供が泳いでいる写真がかかっていた。
‘The picture of a swimming child was hung.’

Parsings for the sentences of (15) are given in (16) and (17).

(16)
```
(IP-MAT (PP (NP (IP-REL (NP-SBJ *T*)
                        (NP-TMP (N 昨日))
                        (VB とっ)
                        (AXD た))
                (N 写真))
            (P が))
        (VB かかっ)
        (P て)
        (VB2 い)
        (AXD た)
        (PU 。))
```

(17)
```
(IP-MAT (PP (NP (IP-EMB (PP (NP (N 子供))
                            (P が))
                        (VB 泳い)
                        (P で)
                        (VB2 いる))
                (N 写真))
            (P が))
        (VB かかっ)
        (P て)
        (VB2 い)
        (AXD た)
        (PU 。))
```

The relative clause marking -REL of (16) leads to meaning representation (18a), while the embedded clause marking -EMB of (17) allows for producing (18b).

(18) a.
$$\begin{aligned}
&\exists t_1 e_1 x(\exists e_2(\text{past}(e_2) \wedge \\
&\quad \text{写真}(x) \wedge \text{昨日}(t_1) \wedge \\
&\quad \text{とっ}(e_2,\ x) \wedge \text{時間}(e_2) \sqsubseteq t_1) \wedge \\
&\quad\ \text{past}(e_1) \wedge \\
&\quad\quad \text{かかってい}(e_1) \wedge \text{が}(e_1) = x)
\end{aligned}$$

b.
$$\begin{aligned}
&\exists x e_1(\text{past}(e_1) \wedge \\
&\quad \text{写真}(x, \\
&\quad \exists y e_2(\text{子供}(y) \wedge \\
&\quad\ \text{泳いでいる}(e_2) \wedge \text{が}(e_2) = y)) \wedge \\
&\quad \text{かかってい}(e_1) \wedge \text{が}(e_1) = x)
\end{aligned}$$

The relative clause of (16) contains a trace denoted *T* that is function marked (-SBJ = subject). In deviation from the annotation scheme there is no CP layer as Japanese has neither overt complementisers nor relative pronouns, and so there is no coindexing either. Coindexing is however used with the annotation of internally headed relative clauses, where explicit marking of the internal head is required.

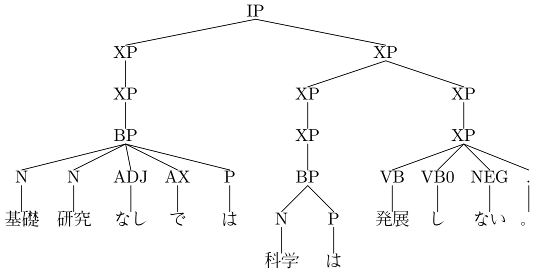
4 Treebank construction

We have developed a tool chain to reach parsed representations employing existing bunsetsu based parsing technologies. To achieve the parsed representation of (3) for the sentence in (1), we begin with the bunsetsu dependency analysis of (19), here produced by the dependency parser `cabocha` (Kudo and Matsumoto 2002), but we also utilise the data of existing bunsetsu dependency treebanks.

(19)
```
* 0 2D 2/4 4.983084
基礎    名詞,一般,*,*,*,*,基礎,キソ,キソ,,0
研究    名詞,サ変接続,*,*,*,*,研究,ケンキュウ,ケンキュー,,0
なし    形容詞,自立,*,*,形容詞・アウオ段,文語基本形,ない,ナシ,ナ
シ,なし/無し,0
で      助動詞,*,*,*,特殊・ダ,連用形,だ,デ,デ,,0
は      助詞,係助詞,*,*,*,*,は,ハ,ワ,,0
* 1 2D 0/1 0.000000
科学    名詞,一般,*,*,*,*,科学,カガク,カガク,,0
は      助詞,係助詞,*,*,*,*,は,ハ,ワ,,0
* 2 -1D 1/2 0.000000
発展    名詞,サ変接続,*,*,*,*,発展,ハッテン,ハッテン,,0
し      動詞,自立,*,*,サ変・スル,未然形,する,シ,シ,,0
ない    助動詞,*,*,*,特殊・ナイ,基本形,ない,ナイ,ナイ,,0
。      記号,句点,*,*,*,*,。,。,。,,0
EOS
```

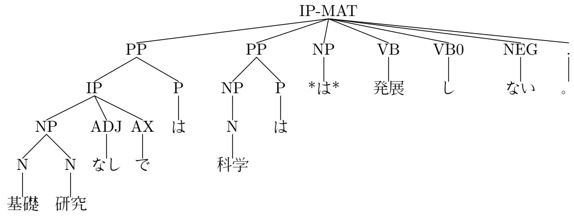
The bunsetsu analysis of (19) is processed to return the tree structure of (20).

(20)



The tree of (20) is further modified by tree transforming scripts using the `tsurgeon` tool (Levy and Andrew 2006) to return the tree of (21).

(21)



Additional modifications to reach the final parsed structure of (3) must be undertaken by a human annotator. This step involves correcting false part-of-speech assignments and mistaken attachments as well as adding functional information, particularly to NP and IP nodes. Every IP (or an immediately dominating CP) must be marked for function. Among clause level constituents (i.e., all phrases immediately dominated by IP), function is marked on all noun phrases (NP-SBJ = subject, NP-OB1 = object NP, NP-TMP = temporal NP, etc.). Bare NPs are either complements of a non-verbal head (e.g., a particle), or part of a conjunction structure, and so the need for NP functional information is easily determined. In the case of annotating (1) we also see the removal of an IP node to capture the functional role of なしでは 'without' as a combination particle.

5 Treebank content

Content for the Keyaki treebank is to include:

- Textbook examples, e.g., from the Tanaka corpus (Tanaka 2001)
- Mainichi Shimbun newspaper articles for 1995 (the data of the Kyoto Text Corpus)
- Law texts from `http://www.japaneselawtranslation.go.jp`
- Articles from *Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles* (MASTAR 2011)

6 Maintaining compatibility

With our plan to include data of the Kyoto Text Corpus as the biggest portion of the treebank, comes the significant desire to maintain as much compatability with the Kyoto Text Corpus as possible. This can be achieved with a tabular format that orients tree structure around parts-of-speech nodes. For example, the bracket representation of (3) can be reformatted as (22). When there is identical segmentation this allows aligning with tabular bunsetsu dependency structures, such as (19).

(22)
```
(IP-MAT_(PP_(NP_*  N         基礎
*)                 N         研究
*                  ADJ       なし
*                  AX        で
*)                 P         は
(PP_(NP_*)         N         科学
*)                 P         は
*                  NP-SBJ    *は*
*                  VB        発展
*                  VB0       し
*                  NEG-HIGH  ない
*)                 PU        。
```

With alignment we can automatically take information from a bunsetsu annotated corpus. For example, the NAIST Text Corpus (Iida, Komachi, Inui, and Matsumoto 2007) annotates zero pronouns for が, を and に arguments for the data of the Kyoto Text Corpus. This information can be integrated into our treebank, first at the lexical level of predicates and then distributed appropriately with, for example, structure adding `tsurgeon` scripts. Similarly information from our annotation can be fed into bunsetsu dependency treebanks, e.g., in the way version 4.0 of the Kyoto Text Corpus (Kawahara et al. 2005) adds case frame information.

7 Conclusion

To sum up this paper has described our project of building a treebank for Japanese, following the *Annotation manual for the Penn Historical Corpora and the PCEEC* (Santorini 2010), with the goal of enabling an automatic retrieval of meaning representations to support formal semantics research on a corpus linguistics scale. We highlighted advantages of the new treebank over existing treebanks for Japanese. These advantages were not so much in terms of the information content that the treebanks might contain in principle, but rather with how a given treebank is able to make information accessible for further processing. Most notably, the ability to mark clause level functional information makes it readily possible to build meaning representations that go beyond the predicate-argument structure level. Having an internal syntax where phrasal categories are fundamentally similar is also of great assistance.

References

Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for Treebank II style Penn Treebank project. Tech. Rep. MS-CIS-95-06, LINC LAB 281, University of Pennsylvania Computer and Information Science Department.

Bies, Ann and Mohamed Maamouri. 2003. Penn arabic treebank guidelines. Tech. rep., Linguistic Data Consortium, University of Pennsylvania. DRAFT.

Butler, Alastair, Zhen Zhou, and Kei Yoshimoto. 2012. Problems for successful bunsetsu based parsing and some solutions. In *Proceedings of the Eighteenth Annual Meeting of the Association of Natural Language Processing*, pages 951–954. The Association of Natural Language Processing.

Galves, Charlotte and Helena Britto. 2002. *The Tycho Brahe Corpus of Historical Portuguese*. Department of Linguistics, University of Campinas. Online publication, first edition, (http://www.tycho.iel.unicamp.br/ tycho/).

Han, Chung-hye, Na-Rae Han, and Eon-Suk Ko. 2001. Bracketing guidelines for penn korean treebank. Tech. Rep. IRCS Report 01-10, Institute for Research in Cognitive Science, University of Pennsylvania.

Hashimoto, Shinkichi. 1934. *Essentials of Japanese Grammar (Kokugoho Yousetsu)*. Iwanami. (In Japanese).

Iida, Ryu, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In *ACL Workshop 'Linguistic Annotation Workshop*, pages 132–139.

Kawahara, Daisuke, Ryohei Sasano, Sadao Kurohashi, and Koichi Hashida. 2005. Specification for annotating case, ellipsis and coreference. Kyoto Text Corpus Version 4.0. (In Japanese).

Kroch, Anthony, Beatrice Santorini, and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, (http://www.ling.upenn.edu/hist-corpora/).

Kroch, Anthony, Beatrice Santorini, and Ariel Diertani. 2010. *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, (http://www.ling.upenn.edu/hist-corpora/).

Kroch, Anthony and Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, (http://www.ling.upenn.edu/hist-corpora/).

Kudo, Taku and Yuji Matsumoto. 2002. Japanese dependency analyisis using cascaded chunking. In *Proceedings of 6th CoNLL*, pages 63–69.

Kurohashi, Sadao and Makoto Nagao. 2003. Building a Japanese parsed corpus – while improving the parsing system. In A. Abeillé, ed., *Treebanks: Building and Using Parsed Corpora*, chap. 14, pages 249–260. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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*.

Martineau, France, Paul Hirschbühler, Anthony Kroch, and Yves Charles Morin. 2010. *Corpus MCVF (parsed corpus), Modéliser le changement : les voies du français*. Département de français, University of Ottawa. CD-ROM, first edition, (http://www.arts.uottawa.ca/voies/voies_fr.html).

MASTAR, Project. 2011. Japanese-English bilingual corpus of Wikipedia's Kyoto articles. National Institute of Information and Communications Technology, Online publication, version 2.01, (http://alaginrc.nict.go.jp/WikiCorpus).

Pintzuk, Susan and Leendert Plug. 2002. *The York-Helsinki Parsed Corpus of Old English Poetry*. Department of Linguistics, University of York. Oxford

Text Archive, first edition, (http://www-users.york.ac.uk/˜lang18/pcorpus.html).

Santorini, Beatrice. 2010. Annotation manual for the Penn Historical Corpora and the PCEEC (Release 2). Tech. rep., Department of Computer and Information Science, University of Pennsylvania, Philadelphia.

Tanaka, Yasuhito. 2001. Compilation of a multilingual parallel corpus. In *Pacling2001*.

Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. 2006. *The York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC)*. Department of Linguistics, University of York. Oxford Text Archive, first edition, (http://www-users.york.ac.uk/˜lang22/PCEEC-manual/index.htm).

Taylor, Ann, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)*. Department of Linguistics, University of York. Oxford Text Archive, first edition, (http://www-users.york.ac.uk/˜lang22/YcoeHome1.htm).

Wallenberg, Joel, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC)*. Department of Linguistics, University of Iceland. Online publication, version 0.9, (http://www.linguist.is/icelandic_treebank).

Xue, Nianwen and Fei Xia. 2000. The bracketing guidelines for the penn chinese treebank (3.0). Tech. Rep. 00-08, Institute for Research in Cognitive Science, University of Pennsylvania.