# Coindexed null elements for a Japanese parsed corpus

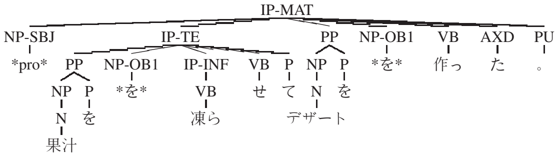
Alastair BUTLER[*] Shota HIYAMA[†] Kei YOSHIMOTO[*]
[*]Institute for Excellence in Higher Education, Tohoku University
[†]Graduate School of International Cultural Studies, Tohoku University

## 1 Introduction

To capture language phenomena such as pro-drop, wh-movement, control, and discontinuous structures, special tokens, called null elements, are used in Penn Treebanks (Marcus et al 1994, Xue and Xia 2000, Bies and Maamouri 2003). Concerned with resource creation for Japanese, this paper describes supplementing hand annotation of a parsed corpus with the generation of formal semantic representations which are used to obtain coindexing information to invest back into the hand annotation. This yields annotation with the full range of coindexed null elements available in Penn Treebanks.

## 2 Annotation scheme

Following a scheme adapted from the annotation manual for the Penn Historical Corpora series (Santorini 2010), our parsed corpus represents syntactic structure with labelled parentheses. A typical parse in tree form looks like:



Following the SUW/LUW standard of the BCCWJ (Maekawa et al 2014), words are tokenised and every word is labelled for part-of-speech (N=noun, P=particle, VB=verb, etc.). Phrasal nodes (NP, PP, ADJP, etc.) immediately dominate the phrase head (N, P, ADJ, etc.), so that the phrase head has as sisters both modifiers and complements, an arrangement optimised for querying annotated content (see e.g. Randall 2009). While this leaves no explicit representation of intermediate levels of structure in the sense of X-bar theory, information is present to transform to an alternative binary tree representation (e.g., to an arrangement optimised for parser training; Tanaka and Nagata 2013), as extended phrase labels marking function distinguish modifiers and complements (e.g., IP-TE above is a modifier, while IP-INF is a complement). The PP (particle phrase) label is never extended with function marking, but the immediately following sibling of a PP may be present in the annotation to provide disambiguation information for the PP. Thus, `(NP-OB1 *を*)` indicates the immediately preceeding PP (with `(P を)`) is the object of its clause.

## 3 Hand annotated null elements

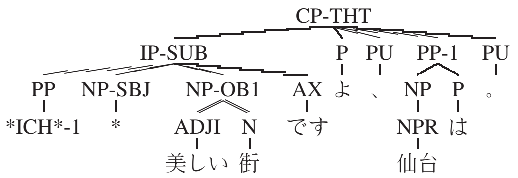
In its hand annotated state, null elements in our parsed corpus include trace markers of relative clauses, null expletives and dropped pronouns, none of which carry coindexing information. A list of elements used is shown in Table 1 along with their intended usage.

| null element | meaning |
|---|---|
| *T* | relative clause trace |
| *exp* | null expletive |
| *pro* | "small/baby pro" (i.e., pro-dropped subject/object) |
| *speaker* | refined *pro* with speaker as referent |
| *speaker+pro* | refined *pro* |
| *speaker+hearer* | refined *pro* |
| *hearer* | refined *pro* with hearer as referent |
| *hearer+pro* | refined *pro* |
| *arb* | generic impersonal reference |

Table 1: List of null elements

Notably our parsed corpus is more fine grained than Penn Treebanks in that *speaker*, *speaker+pro*, *speaker+hearer*, *hearer*, *hearer+pro* and *arb* are all more specified versions of *pro*, reflecting the productiveness of pro-drop in Japanese.

While the null elements mentioned above have no coindexing, one further hand annotated null element is coindexed: *ICH* (abbreviating "interpret constituent here"). This is used as a trace marking discontinuous structures, e.g., extraposition, scrambling, or other displacements that cross phrase levels without fitting into the A/A-bar dichotomy of generative syntax. Coindexing works by having an index number added to the label of the original constituent and incorporated into *ICH* to indicate where the original constituent should be interpreted, as in the following example:

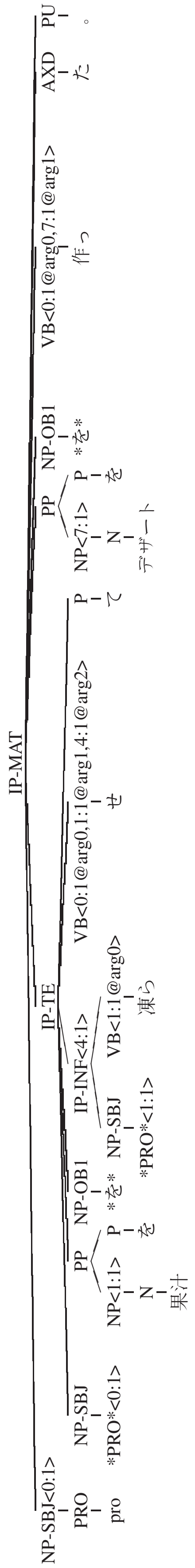


## 4 PRO automatically added

In its hand annotated state, there is no PRO in our parsed corpus, that is, marker of control dependencies. This has been a deliberate decision as control information is reliably calculated based on the arrangement of the annotation via generation of a full semantic representation to gain information which can be embedded back into the original hand annotation. With the Treebank Semantics method (`http://www.compling.jp/ts`; Butler, Otomo, Zhou and Yoshimoto 2013) of obtaining formal semantic representations from constituent tree annotations, the tree of section 2 is sufficiently specified to produce the following semantic representation:

∃$x_6 x_1 x_2 e_3 e_4 e_5$（
$x_6$ = pro ∧
果汁（$x_1$）∧
デザート（$x_2$）∧
て（せ（$e_4$,$x_6$,$x_1$,凍ら（$e_3$, $x_1$）），
作っ_た（$e_5$, $x_6$, $x_2$）））

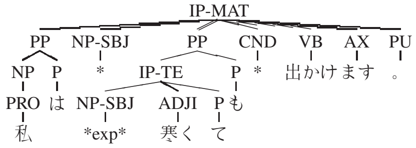
Such a semantic representation captures case frame information which can be integrated back into the annotation to create the following annotation:



Case frame information is here associated with constituents forming arguments. No corresponding constituent gives information to construct *PRO* along with its coindexing information. In deviance to Penn Treebanks, coindexing is given as "word:height" information (following Propbank; Palmer, Gildea and Kingsbury 2005) that specifies the token number of the first word and the number of levels up in the tree to go to find the root of the appropriate constituent.

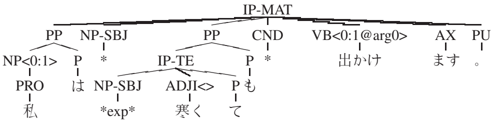
# 5 Hand annotated expletives

Control environments are created with IP-TE=て-clause, IP-ADV=adverbial clause, IP-INF=infinitival clause and IP-EMB=embedded clause with formal noun sister. The default is for control inheritance from the higher clause layer to occur favouring NP-OB2 over NP-LGS (logical subject with passive) over NP-OB1 over NP-SBJ when any of these is present and accessible. However there are cases, e.g., involving weather predicates, where there should be no inheritance. Prevention of a control relation is achieved with the hand annotation of *exp*, as demonstrated in the following example and its resulting semantic representation, with notably 寒く taking only an event binding.



$$\exists z_1\,(z_1 = \text{私} \wedge \forall e_2\text{て\_も}\,(\text{寒く}\,(e_2),\ \exists e_3\text{出かけ\_ます}\,(e_3,\ z_1)))$$

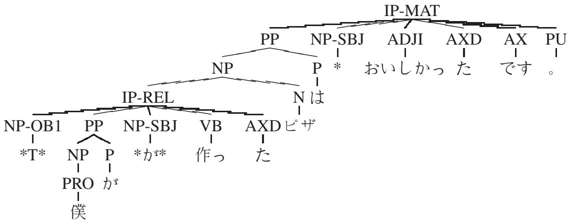
With case frame information from the semantic representation added to the original annotation, there is no creation of *PRO*:



It follows that hand annotation of *exp* is required exactly because there is no hand annotation of *PRO*. That is *exp* and *PRO* have a complementary distribution in control environments and the one can be calculated from the other. With *exp* occurring rarely and taking no indexing, it is the better candidate to hand annotate.

# 6 Coindexed wh operators

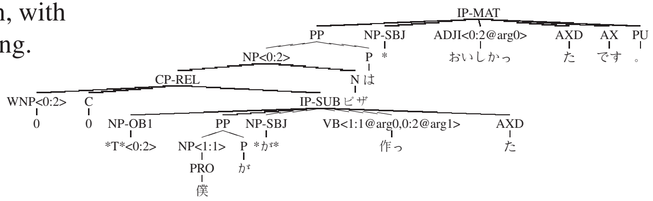
As a final example we consider adding an overt wh operator with relative clauses. The following example illustrates relative clause annotation as projecting an IP-REL constituent with an embedded trace *T* with a default clause initial placement and taking functional information (NP-OB1 to serve as the object inside the relative clause):



This annotation is specified to support deriving the following semantic representation:

$$\exists x_1 x_4 e_2 e_3\,(x_4 = \text{僕} \wedge \text{作っ\_た}\,(e_2, x_4, x_1) \wedge \text{ピザ}\,(x_1) \wedge \text{おいしかっ\_た\_です}\,(e_3,\ x_1))$$

This provides linking information for automatically recovering a coindexed wh operator, akin to such operators present in the Penn Treebanks, with an elaborated CP-REL projection:



# 7 Corpus Statistics

This section outlines statistics for 3915 annotated sentences to which PRO and wh operator information was automatically added. All sentences are from the newspaper domain. Table 2 details numbers involved of the different types of null element.

| | hand added | | | | | automatic | |
|---|---|---|---|---|---|---|---|
| sent | pro | arb | speaker | hearer | exp | PRO | wh op |
| 3915 | 2542 | 40 | 84 | 15 | 29 | 1945 | 1721 |

Table 2: numbers involved of the different types of null element

The increasing likelihood of meeting null elements with growing sentence lengths is illustrated by Figure 1, which gives a histogram of sentence lengths, alongside data for sentences of a given length restricted to containing one or more instances of the range of null elements, with data for the refined variants of *pro* collapsed under the “with pro” entry.

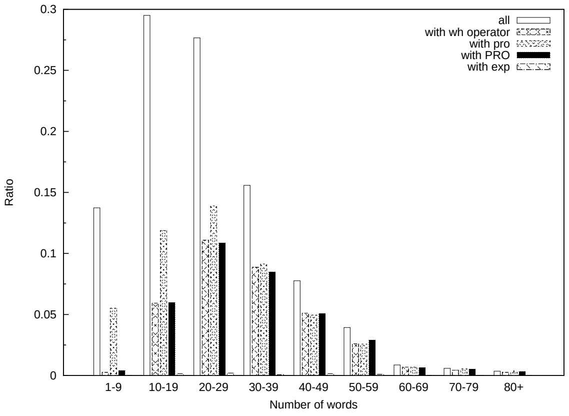


Figure 1: histogram of sentence lengths

# 8 Conclusion and applications

This paper has described an effort to capture null elements in a Japanese parsed corpus by supplementing hand annotation with an automated adding of control and coindexing information. Null elements are relevant for the linguistic community as they provide a means to disambiguate different types of structure, so it becomes possible to read off frequency counts for pro-drop, relative clauses, control, discontinuous structures, etc. The automated procedure for adding PRO information also serves as a cashing out of a theory of control in Japanese, testing the assumption that control relationships are governed by grammatical processes (clause type, passivisation, causativisation, etc). Tracking control also has a larger role to play in establishing syntactic properties, most significantly to informing levels of attachment and scope. The resulting annotation also makes predicate argument interpretation straightforward, e.g., null elements can simply be replaced by the constituent to which they are coindexed. Furthermore, Xiang, Luo and Zhuo (2013) demonstrate significant improvements with statistical machine translation when including null elements, and so depend on large-scale null element annotation as training data.

## Acknowledgements

We are very grateful to Masaaki Nagata for the discussion that prompted this paper. This research is supported by the feasibility-type collaborative research project entitled 日本語テクストのツリーバンクアノテーション法の開発 (Development of annotation scheme for Japanese text treebank) funded by the National Institute for Japanese Language and Linguistics (NINJAL), Japan.

## References

Bies, Ann and Mohamed Maamouri. 2003. Penn Arabic Treebank Guidelines. Tech. rep., Linguistic Data Consortium, University of Pennsylvania. DRAFT.

Butler, Alastair, Ruriko Otomo, Zhen Zhou, and Kei Yoshimoto. 2013. Treebank annotation for formal semantics research. In Y. Motomura, A. Butler, and D. Bekki, eds., *JSAI-isAI 2012*, LNAI 7856. Heidelberg: Springer.

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.

Marcus, Michell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the Human Language Technology Workshop, March 1994*. San Francisco, CA.: Morgan Kaufmann Publishers Inc.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Randall, Beth. 2009. CorpusSearch 2 Users Guide. (http://corpussearch.sourceforge.net).

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, Takaaki and Masaaki Nagata. 2013. Constructing a practical constituent parser from a Japanese treebank with function labels. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*.

Xiang, Bing, Xiaoqiang Luo, and Bowen Zhou. 2013. Enlisting the ghost: Modeling empty categories for machine translation. In *Proceedings of ACL*, pages 822–831.

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.