Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5195–5201
Marseille, 11–16 May 2020

Development of a General-Purpose Categorial Grammar Treebank

Yusuke Kubota, Koji Mineshima, Noritsugu Hayashi, Shinya Okano
NINJAL, Ochanomizu University, The University of Tokyo, idem.
{10-2 Midori-cho, Tachikawa; 2-2-1 Otsuka, Bunkyo; 3-8-1 Komaba, Meguro}, Tokyo, Japan
kubota@ninjal.ac.jp, mineshima.koji@ocha.ac.jp, {hayashi, okano}@phiz.c.u-tokyo.ac.jp

Abstract

This paper introduces ABC Treebank, a general-purpose categorial grammar (CG) treebank for Japanese. It is 'general-purpose' in the sense that it is not tailored to a specific variant of CG, but rather aims to offer a theory-neutral linguistic resource (as much as possible) which can be converted to different versions of CG (specifically, CCG and Type-Logical Grammar) relatively easily. In terms of linguistic analysis, it improves over the existing Japanese CG treebank (Japanese CCGBank) on the treatment of certain linguistic phenomena (passives, causatives, and control/raising predicates) for which the lexical specification of the syntactic information reflecting local dependencies turns out to be crucial. In this paper, we describe the underlying 'theory' dubbed ABC Grammar that is taken as a basis for our treebank, outline the general construction of the corpus, and report on some preliminary results applying the treebank in a semantic parsing system for generating logical representations of sentences.

Keywords: treebank, categorial grammar, Japanese, annotation

1. Introduction

This paper reports on the progress of the construction of a general-purpose categorial grammar treebank in Japanese that can be used as a resource in both theoretical linguistics and natural language processing. Categorial grammar (CG) is a linguistic theory that is well-known for its explicit and simple syntax-semantics interface and is regarded as one of the most influential linguistic theories that can be used for NLP research with 'deep' semantic interpretation that has gained renewed attention in recent years. For a further development of this line of work, it is essential to construct a linguistically valid treebank on CG. However, current corpora based on CG often do not take advantage of linguistically adequate analyses developed in the CG literature, mainly because these corpora are converted from existing resources which do not contain fine-grained annotation (Honnibal et al., 2010). We will see that this is also the case with the current Japanese CCGBank (Uematsu et al., 2015), a CG treebank converted from dependency treebanks (Section 4). To build a linguistically valid Japanese CG treebank, we use a constituency treebank (the Keyaki Treebank) as a source corpus, which has fine-grained annotation including predicate-argument structures and empty categories. In this paper, we report the progress of our project and compare our treebank with the current Japanese CCGBank.

2. ABC Grammar as a Categorial Interlangage

There are two major lines of research in categorial grammar: Combinatory Categorial Grammar (CCG; (Steedman, 2000)) and Type-Logical Grammar (TLG; (Morrill, 1994; Moortgat, 1997; Kubota and Levine, 2015)). CCG is well-known as one of the few implementable linguistic theories that enable deep semantic analysis (Steedman, 2000; Clark and Curran, 2007). TLG belongs to the tradition that seeks to understand the theoretical underpinnings of CG in terms of mathematical logic. TLG is also distinct from CCG in having a more transparent syntax-semantics interface that is closer in its core architecture to the Chomskian linguistic theory using the notion of 'syntactic movement' (Carpenter, 1997; Kubota and Levine, 2015).

Previous work on converting PSG and dependency treebanks to CG treebanks includes Hockenmaier and Steedman (2007), Moot (2015), and Uematsu et al. (2015). All these studies take a particular version of CG (i.e. a variant of either CCG or TLG) as the 'target' theory in conversion (see Table 1). The present study differs from these previous studies in its deliberate choice on being agnostic about the target CG theory. This makes it easier to adapt the output treebank for multiple purposes (in CCG/TLG parsing and even for the purpose of converting to other grammatical formalisms such as HPSG). For this purpose, we have chosen to encode the output in an intermediate framework 'ABC grammar' (described below), which basically consists of the common components of CCG and TLG.

Table 1: Comparison with related work

	original corpus	output
H&S	Penn Treebank	English CCG
Uematsu et al.	Kyoto Corpus (dependency)	Japanese CCG
Moot	French PSG Bank	French TLG
present work	Keyaki Treebank (= PSG+α)	Japanese ABC (= general-purpose CG)

The ABC Grammar consists of the so-called AB Grammar (Ajdukiewicz, 1935; Bar-Hillel, 1953) (which consists solely of the function application rules, shown on the left-hand side of Figure 1), plus (a certain subset of) function composition rules, specifically, the rules shown on the right-hand side of Figure 1.[1] This architecture provides a clear and concise analysis of a large subset of linguistically important phenomena, including causative predicates, passivization, raising and control, as well as nonconstituent coordination (Kubota, 2014; Kubota and Levine, 2015). These phenomena involve rearrangement of argument structure, which in the CG context amounts to the rearrangement of lexically assigned syntactic category specification. At a certain level of abstraction, function composition corresponds to the concept of 'syntactic movement' and it is one of the important concepts in categorial grammar. Function composition is a basic rule in CCG, and in TLG, it can be derived as a theorem from the more primitive rules of grammar. For this reason, an ABC Grammar treebank can be easily converted further to a CCG treebank or a TLG treebank.

[1] We use the TLG convention due to Lambek (1958) (instead of the CCG convention) in notating slashes. A/B is a category that looks for a category B to its right to form an A and $B\backslash A$ is a category that looks for a category B to its left to form an A.

Function Application	**Function Composition**
$A/B \;\; B \Rightarrow A$	$A/B \;\; B/C \Rightarrow A/C$
$B \;\; B\backslash A \Rightarrow A$	$C\backslash B \;\; B\backslash A \Rightarrow C\backslash A$

Figure 1: Rules of ABC Grammar

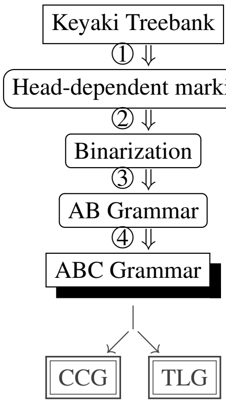


Figure 2: The pipeline of conversion

3. Converting the Keyaki Treebank

We choose the Keyaki Treebank [2], a resource related to the NINJAL Parsed Corpus of Modern Japanese (NPCMJ) [3], as the source corpus. The Keyaki Treebank is a treebank of Modern Japanese based on phrase structure grammar. The original text consists of open-access and commercially-licensed materials, and the PSG annotation is available for free. The treebank provides some 50,000 sentences (including both the open-access and commercially-licensed parts) annotated with phrase-structural constituency and fine-grained grammatical information (subject, object, etc.).
Figure 2 outlines the conversion process of Keyaki sentences to ABC Grammar trees. The details are exemplified by Figure 3.

4. Comparison with the Japanese CCGBank

One of the advantages of the ABC Grammar is that it can give a concise and linguistically appropriate analysis of the argument structure of predicates by using two grammatical rules: Function Application and Function Composition. In this section, the features of our treebank based on the ABC Grammar are introduced in comparison with the existing Japanese CCGBank (Uematsu et al., 2015), focusing on some linguistically interesting cases: passives and causatives (section 4.1) and raising and control predicates (section 4.2).

4.1. Passives and causatives

In so-called 'lexicalist' theories of syntax, passivization and causativization are generally considered to be operations that change the argument structures of predicates (Bresnan, 1982; Manning et al., 1999). In the Japanese CCGBank, however, a passive morpheme *-rare* and a causative morpheme *-sase* are given the category S\S, as in (1) and (2), respectively.[4] This analysis does not properly reflect their effect on the argument structure of the verb.

(1) a.

すでに	サイ	は	投げ	られ	た
Sudeni	sai	wa	**nage**	**rare**	ta
already	die	TOP	cast	PASS	PAST

'The die has already been cast' [5]

b.

投げ	られ		
nage (cast)	**rare**		
$NP_{ga}\backslash S$	$S\backslash S$	$\Rightarrow$	$NP_{ga}\backslash S$

(2) a.

全員	を	引き揚げ	させ	た。
Zen'in	o	**hikiage**	**sase**	ta
everyone	ACC	leave	CAUS	PAST

'They made everyone leave.' [6]

b.

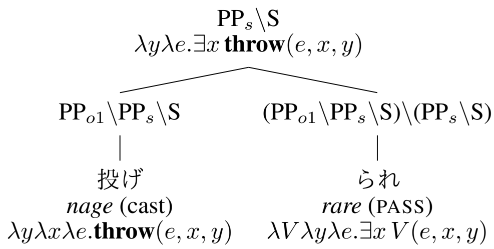
引き揚げ	させ		
hikiage (leave)	**sase**		
$NP_o\backslash S$	$S\backslash S$	$\Rightarrow$	$NP_o\backslash S$

Table 2: Argument structure-changing predicates

Argument-reducing morphemes	
Category : $(PP_{o1}\backslash PP_s\backslash S)\backslash PP_s\backslash S$	etc.
Passive	*(ら)れ (ra)re*
Tough Pred.	がたい *gata-i* 'hard', にくい *niku-i* 'hard'
	づらい *zura-i* 'hard', やすい *yasu-i* 'easy'
Perfect	*(て)ある (te) ar-u* 'have already been'
Argument-increasing morphemes	
Category : $(PP_s\backslash S)\backslash(PP_{o1}\backslash PP_s\backslash S)$	
Causative	*(さ)せ (sa)se*
Benefactive	*(て)もらう (te) mora-u* 'get sth. done'
	(て)いただく (te) itadak-u
	'get sth. done' (referent honorific)

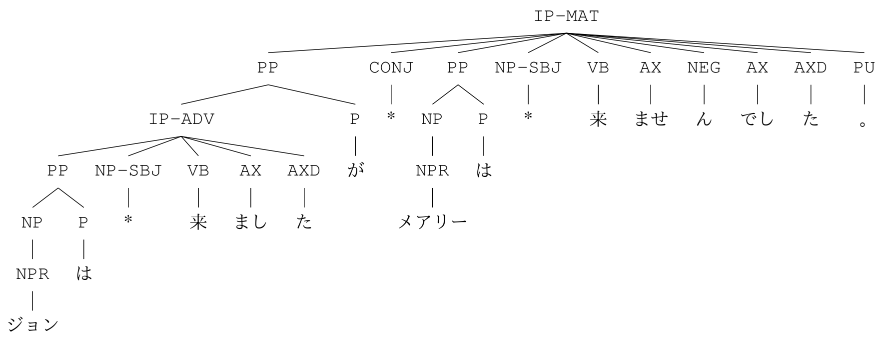
The Keyaki Treebank makes available information that is necessary to assign the right syntactic categories to these argument structure-changing predicates, since it explicitly represents all the obligatory arguments of a predicate with possible use of empty categories. We exploit this information in the source treebank in the conversion process. After the conversion into the ABC Grammar, the passive morpheme *-(ra)re* and the causative morpheme *-(sa)se* are assigned plausible categories: The former reduces the valence of predicates by one and the latter increases it by one. The following (3) and (4) show the categories[7] for both with their semantic representations in the framework of the standard event semantics (Davidson, 1967; Parsons, 1990).[8]

(3) The passive morpheme *-rare*



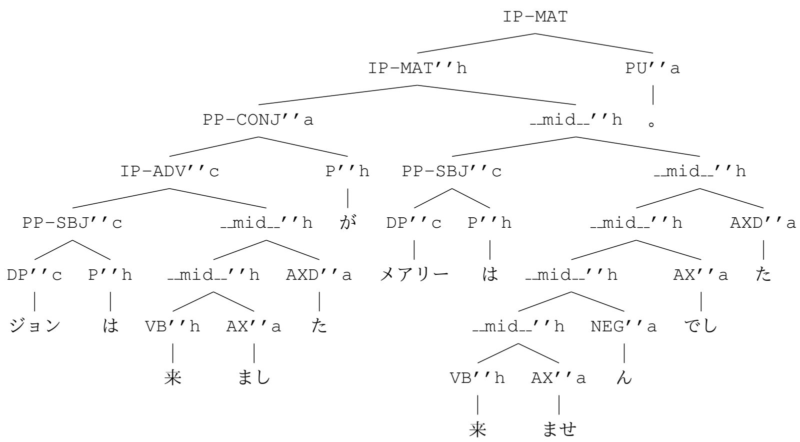
(4) The causative morpheme *-sase*

[2] http://www.compling.jp/keyaki/
[3] http://npcmj.ninjal.ac.jp/?lang=en
[4] Here we use the notion for categories in the Japanese CCGBank: NP_{ga} for NP with nominative case and NP_o for NP with accusative case.
[5] CCGBank ID: devel-3649
[6] CCGBank ID: devel-4462
[7] Following the Keyaki Treebank, we use the category PP_s for subject postpositional phrase (PP) and the category PP_{o1} for primary object PP. See also Figure 3 for a description of categories.
[8] The semantic translation here is given for illustration purposes only; the syntactic analysis is agnostic about the specific choice of the underlying semantic theory, in particular the treatment of event variables in compositional semantics



①
- Unnecessary annotations in the original tree are deleted.
- Grammatical roles are added to nodes. `''h`: head, `''c`: complement, `''a`: adjunct/auxiliary, `''ac`: auxiliary – control predicate

②
- The whole tree is binarized.
  - `__mid__` fills intermediate projections which are eventually replaced with appropriate AB Grammar categories.



③, ④
- The tree is transformed to an AB Grammar tree following the general algorithm of (Hockenmaier and Steedman, 2007). `PPs`: subject postpositional phrase (PP), `PPo1`: primary object PP, `PPo2`: secondary object PP, `PP-LGS`: logical subject in passive sentences, `Sm`: matrix clause (sentence), `Sa`: adverbial clause, `Ssmc`: small clause
- At the same time, the categories of sentence-final expressions are contracted so as to make them as simple as possible (see Section 4.2), invoking function composition.

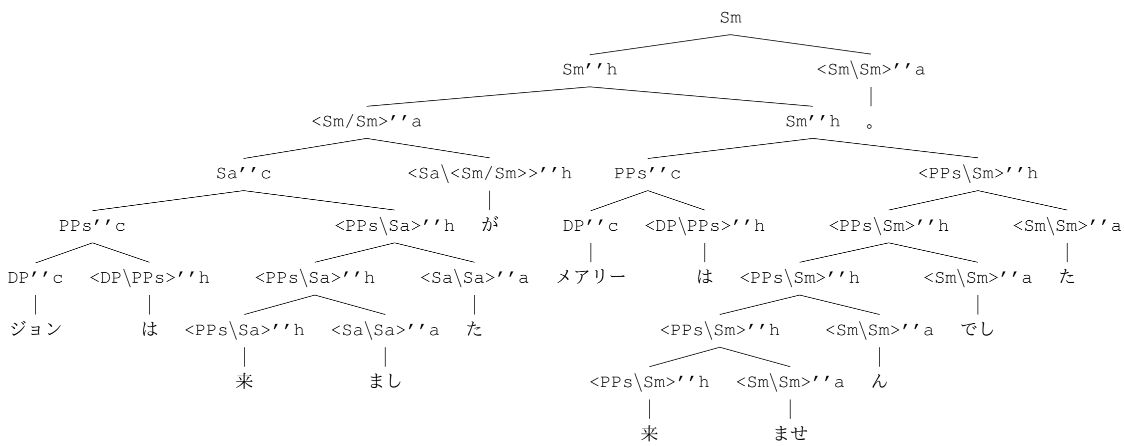


Figure 3: Sample sentence (Keyaki ID: `43_textbook_djg_basic`) going through the conversion process

Table 3: Raising and control predicates in Japanese

Category：S\S	
Raising Predicates	
Aspect	終わる *owar-u* ‘finish’、かける *kake-ru* ‘be about to’
Copula	だ *da*, で *de*, *(で)* ある *(de) ar-u*
Addressee Honorifics	です *des-u*, ます *mas-u*
Focus	の *(だ) no (da)*
Negation	ない *nai*, ず *zu*
Speaker-oriented Modals	かもしれない *kamoshirenai* ‘might’
Sentence-final Particles	よ *yo*, ね *ne*
Category：$(PP_s\backslash S)\backslash(PP_s\backslash S)$	
Control Predicates	
Referent Honorifics	*(て)* いらっしゃる *(te) irasshar-u* (marker for respectful language)
	申し上げる *mooshiage-ru* (marker for modest language)
Subject-oriented Modals, Attitudes	つもり *tsumori* ‘intend’, *(て)* みる *(te) mi-ru* ‘try’
Benefactives	*(て)* あげる *(te) age-ru*, *(て)* くれる *(te) kure-ru*

$PP_{o1}\backslash PP_s\backslash S$
$\lambda y\lambda x\lambda e.\exists e'(\textbf{cause}(e,y,e') \wedge \textbf{leave}(e',x))$

$PP_s\backslash S$	$(PP_s\backslash S)\backslash(PP_{o1}\backslash PP_s\backslash S)$
引き揚げ	させ
hikiage (leave)	*sase* (CAUS)
$\lambda x\lambda e.\textbf{leave}(e,x)$	$\lambda V\lambda y\lambda x\lambda e.\exists e'(\textbf{cause}(e,y,e') \wedge V(e',x))$

Table 2 summarizes these and other expressions which change the argument structure. (8) in Figure 4 at the end of the paper is an example tree obtained by the conversion.

4.2. Control and raising predicates

From a linguistic point of view, predicates that take other predicates as arguments can be classified into two types: raising and control predicates. One major difference between them is that the latter impose a selectional restriction on the subject while the former do not. For example, the raising predicate *seem* does not impose a selectional restriction on the subject, so that the expletive *it* can appear as its subject, as in (5a) below. This is not the case with control predicates like *try*, which requires that its subject be volitional, as the ungrammaticality of (5b) shows.

(5) a. It seems to be raining.
b. * It tries to be raining.

A similar distinction also exists in Japanese. Many predicate-embedding predicates can be regarded as raising[9] predicates. Some examples include verbs such as *das-u* ‘start’ and *owar-u* ‘finish’, which constitute compound verbs, and auxiliaries such as the addressee honorific marker *mas-u* and the past tense marker *ta*. There are also verb-embedding predicates which are classified as control predicates due to their selectional restriction on the subject, including the subject honorific verb *nasar-u* and the adjectival suffix *ta-i* ‘want’. Furthermore, aspectual verbs such as *hajime-ru* ‘begin’ and *tsuzuke-ru* ‘continue’ are often regarded as ambiguous between raising and control predicates in the linguistic literature (Kageyama, 1993; Matsumoto, 1996).

In Japanese CCGbank, both raising and control predicates are given the category S\S. (6) is an example of the control predicate *mi-ru* ‘try’.

(6) a.

この	荒地	に	立っ	て	みる
kono	arechi	ni	**tat**	**te**	**miru**
this	barren	LOC	stand	CONT	try
と *[...]*					
to					
CONJ					

‘When one tries standing on this barren [...]’[10]

b.

立っ	て	みる		
tat (stand)	**te** (CONT)	**miru**		
$NP_{ga}\backslash S$	$S\backslash S$	$S\backslash S$	$\Rightarrow$	S

In the ABC treebank, in contrast, control predicates are assigned the category $(PP_s\backslash S)\backslash(PP_s\backslash S)$, taking their selectional restriction on the subject into consideration, while raising predicates are assigned S\S.[11] This distinction makes it possible to derive a plausible semantic representation for control predicates, capturing their selectional restriction on the subject, as is illustrated in (7).

(7) Semantic representation for a control structure

$PP_s\backslash S$
$\lambda x\lambda e.\exists e'(\textbf{try}(e,x,e') \wedge \textbf{stand}(e',x))$

$PP_s\backslash S$ — $\lambda x\lambda e.\textbf{stand}(e,x)$; $(PP_s\backslash S)\backslash(PP_s\backslash S)$

$PP_s\backslash S$	$S\backslash S$	$(PP_s\backslash S)\backslash(PP_s\backslash S)$
立っ	て	みる
tat (stand)	*te* (CONT)	*miru* (try)
$\lambda x\lambda e.\textbf{stand}(e,x)$	*id*	$\lambda V\lambda x\lambda e.\exists e'(\textbf{try}(e,x,e') \wedge V(e',x))$

Table 3 classifies predicates and sentence-final expressions in terms of the raising/control distinction. (9) and (10) in Figure 4 show trees which are actually obtained by the conversion.

[9] Although the term *raising* has an implication that an argument of an embedded predicate is moved to a higher position to be realized as an argument of the embedding predicate, there is no consensus as to whether such a movement exists in Japanese, even among researchers who endorse a movement theory in syntax (see (Kishimoto, 2008) for an overview). In this paper, predicates are classified as raising and control based on selectional restriction on the subject, without any commitment to a particular type of syntactic analysis.

[10] CCGBank ID: `test-2156`

[11] See Jacobson (1990) for a similar analysis for English.

5. Generating semantic representations

Semantic representations based on event semantics, as are shown above, can be automatically generated by using ccg2lambda[12], a semantic parser based on CCG (Mineshima et al., 2016; Martínez-Gómez et al., 2016). Since the ABC grammar is a subset of CCG, the compositional mapping of semantic representations for the ABC grammar can be implemented in the same manner as in CCG. We use annotated trees in the ABC treebank as input and manually design templates to map them to formal semantic representations in event semantics. Figure 5 at the end of the paper shows the derivation trees with semantic representations automatically generated for the three sentences in Figure 4.[13]

The quality and consistency of conversion in the process of constructing the treebank will be examined by checking whether appropriate representations are obtained from derivation trees of ABC grammar. The templates used for compositional semantics will be improved and expanded through this process. This will allow us to annotate part of the ABC treebank with gold-standard formal semantic representations that can be used in a variety of NLP applications.

6. Further Issues

We have so far focused on sophisticating the grammatical design of argument structures of complex predicates, which is crucial for the ABC Grammar. The next goal is to extend the coverage of various linguistic phenomena. Specifically, we are currently planning on addressing the following issues:

- scrambling
- headless relative clauses
- pro-drop
- refined argument/adjunct distinction reflecting verb semantics properly

In ongoing work, we are also developing a syntactic parser trained on our treebank. Since the standard rules in the ABC grammar (function application and function composition) are also rules in CCG, we can make use of an off-the-shelf CCG parser (Yoshikawa et al., 2017) for our purpose.

Acknowledgements Our appreciation goes to Masashi Yoshikawa (NAIST) for his help and constructive suggestions. This work is supported by JSPS KAKENHI GRANTs 18K00523 and 15H03210, and the NINJAL collaborative research project 'Cross-linguistic Studies of Japanese Prosody and Grammar'.

7. Bibliographical References

Ajdukiewicz, K. (1935). Die syntaktische Konnexität. In Storrs McCall, editor, *Polish Logic 1920–1939*, pages 207–231. Oxford University Press, Oxford. Translated from *Studia Philosophica*, 1, 1–27.

Bar-Hillel, Y. (1953). A quasi-arithmetic notation for syntactic descriptions. *Language*, 29:47–58.

Bresnan, J. (1982). The passive in lexical theory. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 3–86. MIT Press, Cambridge, Mass.

[12] https://github.com/mynlp/ccg2lambda

[13] The current system outputs predicates in Japanese words. In Figure 4, we replace these predicates with the corresponding English words for readability. We abbreviate $\lambda X_1, \ldots, \lambda X_n.M$ as $\lambda X_1, \ldots X_n.M$. See Mineshima et al. (2016) for more detail on compositional event semantics for Japanese in categorial grammar.

Carpenter, B. (1997). *Type-Logical Semantics*. MIT Press, Cambridge, Mass.

Clark, S. and Curran, J. R. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Davidson, D. (1967). The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

Hockenmaier, J. and Steedman, M. (2007). CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Honnibal, M., Curran, J. R., and Bos, J. (2010). Rebanking CCGbank for improved NP interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 207–215.

Jacobson, P. (1990). Raising as functional composition. *Linguistics and Philosophy*, 13(4):423–475.

Kageyama, T. (1993). *Bumpoo to gokeesee [Grammar and Word Formation]*. Hitsuji Shoboo.

Kishimoto, H. (2008). On Verb Raising. In Shigeru Miyagawa et al., editors, *The Oxford Handbook of Japanese Linguistics*, pages 107–140. Oxford University Press.

Kubota, Y. and Levine, R. (2015). Against ellipsis: Arguments for the direct licensing of 'non-canonical' coordinations. *Linguistics and Philosophy*, 38(6):521–576.

Kubota, Y. (2014). The logic of complex predicates: A deductive synthesis of 'argument sharing' and 'verb raising'. *Natural Language and Linguistic Theory*, 32(4):1145–1204.

Lambek, J. (1958). The mathematics of sentence structure. *The American Mathematical Monthly*, 65(3):154–170.

Manning, C. D., Sag, I. A., and Iida, M. (1999). The lexical integrity of Japanese causatives. In Robert Levine et al., editors, *Studies in Contemporary Phrase Structure Grammar*, pages 39–79. Cambridge University Press.

Martínez-Gómez, P., Mineshima, K., Miyao, Y., and Bekki, D. (2016). ccg2lambda: A compositional semantics system. In *Proceedings of ACL-2016 System Demonstrations*, pages 85–90.

Matsumoto, Y. (1996). *Complex Predicates in Japanese*. CSLI/Kurosio, Stanford/Tokyo.

Mineshima, K., Tanaka, R., Martínez-Gómez, P., Miyao, Y., and Bekki, D. (2016). Building compositional semantics and higher-order inference system for a wide-coverage Japanese CCG parser. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242.

Moortgat, M. (1997). Categorial Type Logics. In Johan van Benthem et al., editors, *Handbook of Logic and Language*, pages 93–177. Elsevier, Amsterdam.

Moot, R. (2015). A type-logical treebank for French. *Journal of Language Modelling*, 3(1):229–264.

Morrill, G. (1994). *Type Logical Grammar: Categorial Logic of Signs*. Kluwer, Dordrecht.

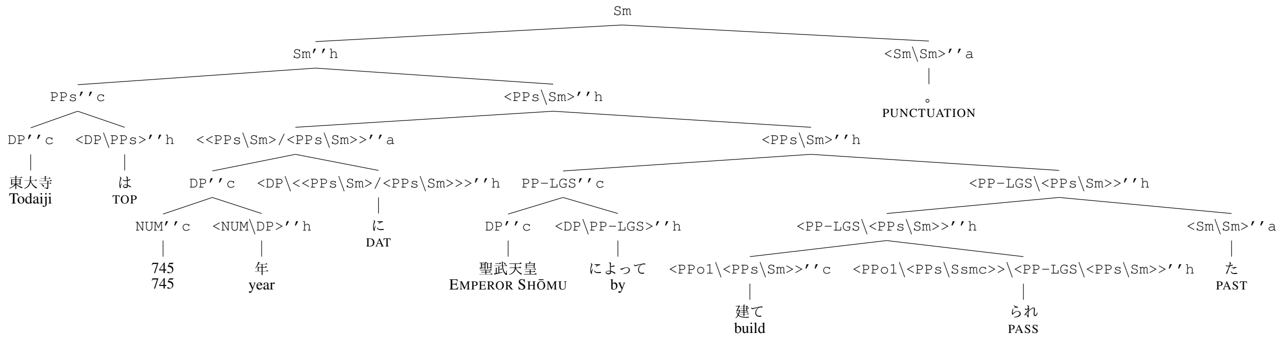
Parsons, T. (1990). *Events in the Semantics of English*. MIT Press, Cambridge, Mass.

Steedman, M. (2000). *The Syntactic Process*. MIT Press, Cambridge, Mass.

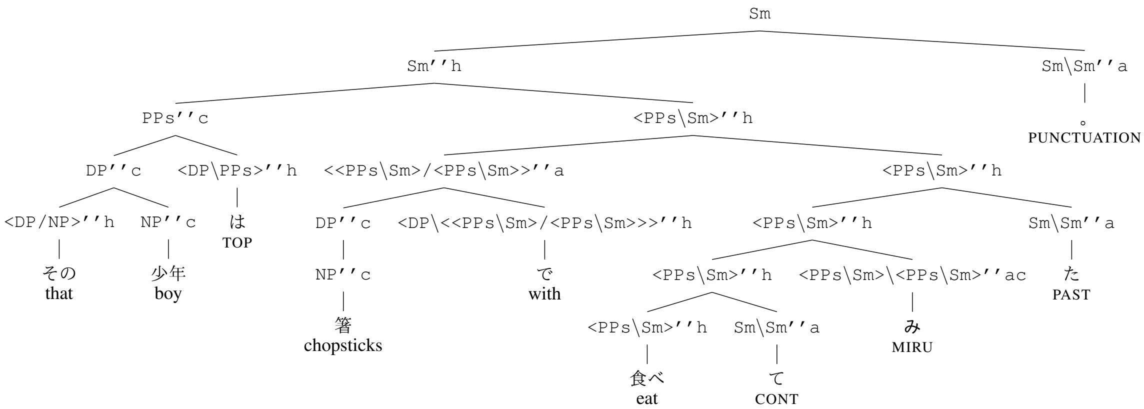
Uematsu, S., Matsuzaki, T., Hanaoka, H., Miyao, Y., and Mima, H. (2015). Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 14(1):1–24.

Yoshikawa, M., Noji, H., and Matsumoto, Y. (2017). A* CCG parsing with a supertag and dependency factored model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 277–287.

(8) Passive sentence (Keyaki ID: 138_textbook_purple_intermediate)



(9) The control predicate *mi-ru* 'try' (Keyaki ID: 321_textbook_TANAKA)



(10) The raising predicate *owar-u* 'finish' (Keyaki ID: 236_textbook_djg_basic)

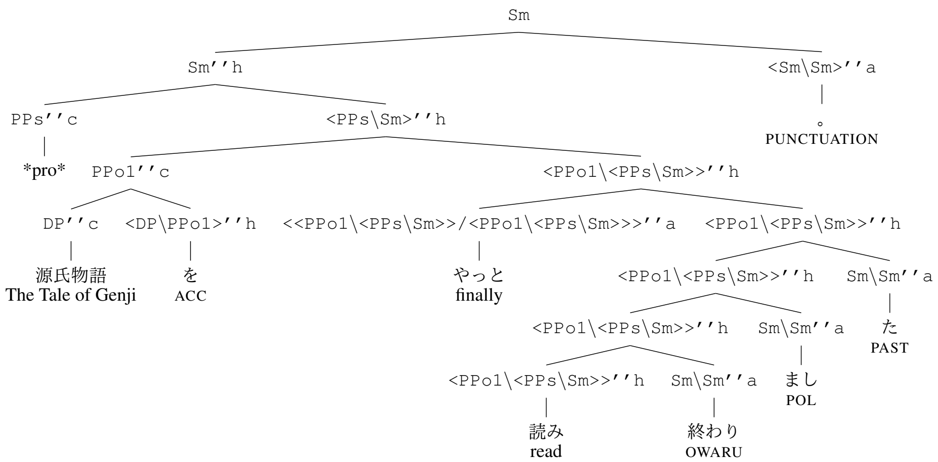
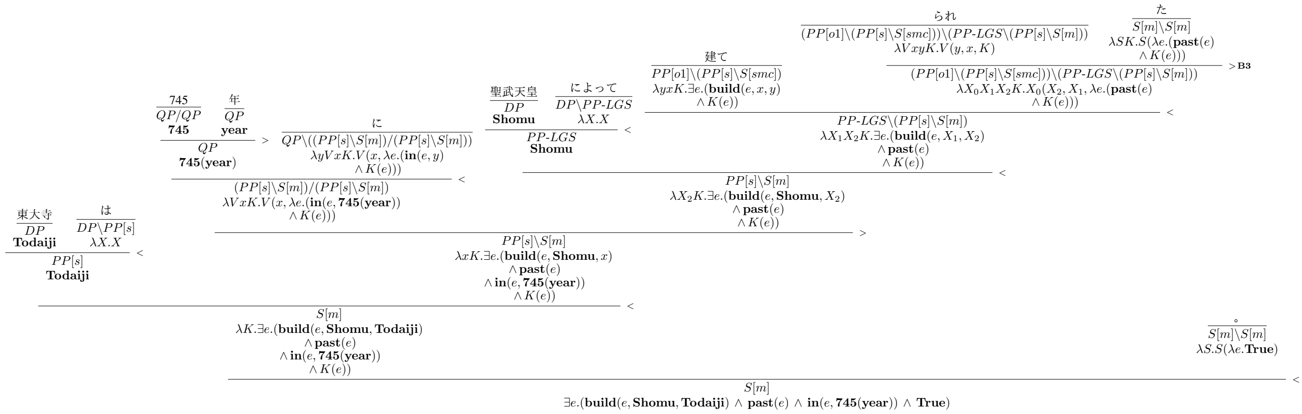
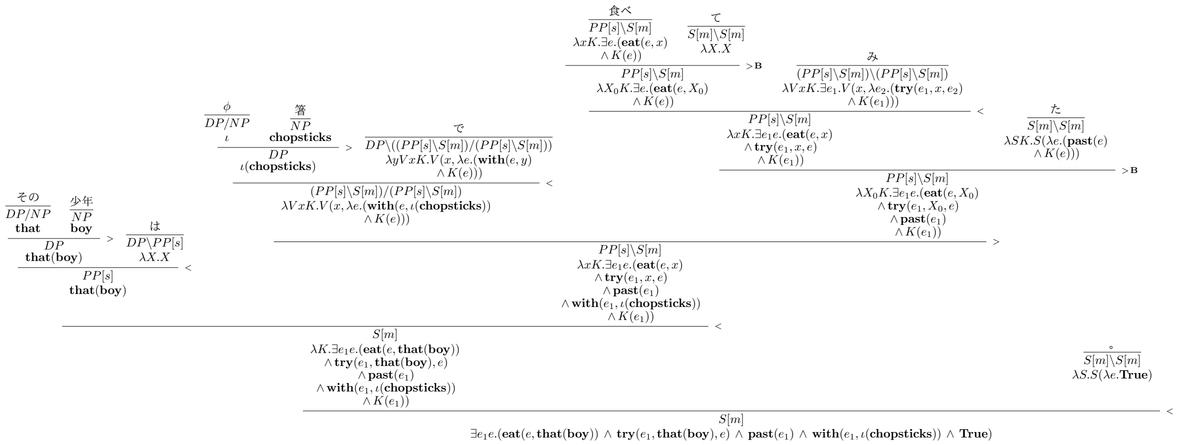


Figure 4: Trees of various constructions attested in the ABC Treebank

(8′) Passive sentence (Keyaki ID: `138_textbook_purple_intermediate`)



(9′) The control predicate *mi-ru* ‘try’ (Keyaki ID: `321_textbook_TANAKA`)



(10′) The raising predicate *owar-u* ‘finish’ (Keyaki ID: `236_textbook_djg_basic`)

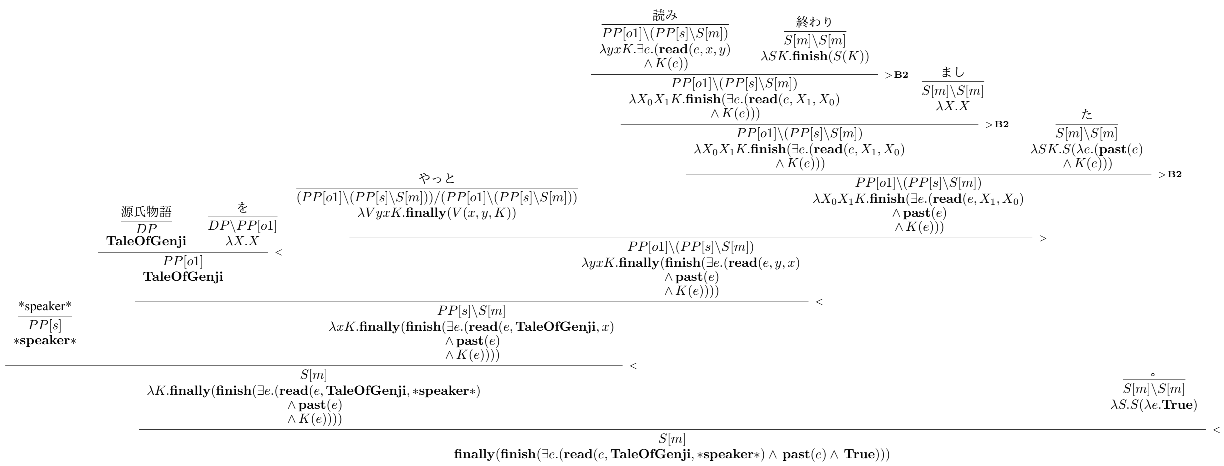


Figure 5: Derivation trees with semantic representations for the examples in Figure 4