ABCツリーバンク: 学際的な言語研究のための基盤資源

窪田悠介

国立国語研究所

概要: ABCツリーバンク (Kubota et al., 2020)

- 汎用的なカテゴリ文法 (categorial grammar) の日本語ツリーバンク
- 理論言語学と自然言語処理の両分野での研究資源
- https://github.com/ABCTreebank/

カテゴリ文法の特徴

- 明示的な文法理論
- 統語構造と意味解釈が一対一に対応
- 理論言語学研究と自然言語処理研究を橋渡しする言語理論として有望
  - 意味解析まで行うパージング
  - 分布意味論と接合させることができる記号処理的な文法体系

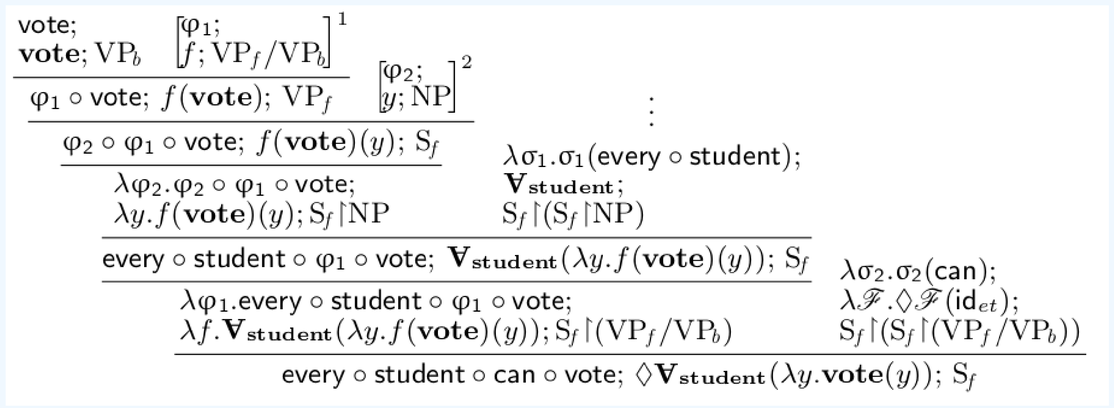


Figure: 分析図の例 (英語の助動詞と量化詞のスコープ)

いろいろなカテゴリ文法

カテゴリ文法は論理学に基づく文法理論だが、生成文法に様々な亜種があるように、大きく二つのタイプに分類される様々な亜種がある。

- **CCG (組み合わせカテゴリ文法)** (Steedman, 1996; Bekki, 2010):
  - 自然言語処理で幅広く使われている
  - 非変換統語論 (nontransformational syntax) の一種
- **TLG (タイプ論理文法)** (Lambek, 1958; Morrill, 1994; Moortgat, 1997; Kubota and Levine, 2020):
  - 論理学的な特徴の研究がCCGより進んでいる
  - 「統語変換」の概念を厳密に定式化できる (前節「分析図の例」参照)

既存のカテゴリ文法ツリーバンクとの比較

- 既存のカテゴリ文法のツリーバンクはすべてCCG、TLGのどちらかに依拠
- 問題点:
  - 汎用性に欠ける
  - CCGとTLGの比較が困難

	元コーパス	変換後
Hockenmaier and Steedman (2007)	Penn Treebank	英語CCG
Uematsu et al. (2013)	京大コーパス(係受け)	日本語CCG
Moot (2013)	French PSG Bank	仏語TLG
本研究	けやきツリーバンク (= 句構造+α)	日本語ABC文法 (= 汎用CG)

Table: 本研究と他の範疇文法ツリーバンク

元コーパス: けやきツリーバンク (Butler et al., 2017)

- 句構造文法のツリーバンク
- 文法役割やゼロ要素のアノテーションあり

提案手法:「中間言語」としてのABC文法

- CCGとTLGは骨格部分を共有している

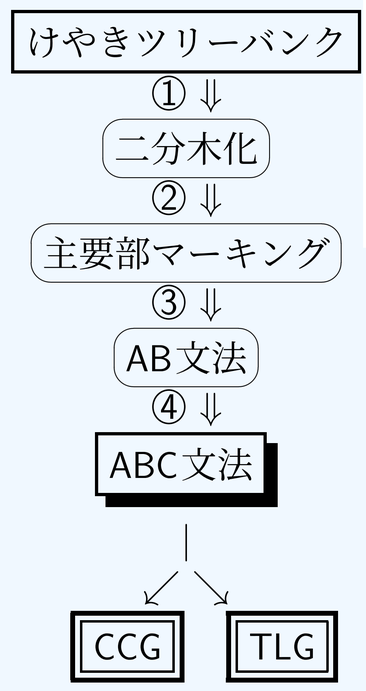
関数適用 (function application)	関数合成 (function composition)
A/B B ⇒ A	A/B (B/C)/$ ⇒ (A/C)/$
B B\A ⇒ A	$\(C\B) B\A ⇒ $\(C\A)

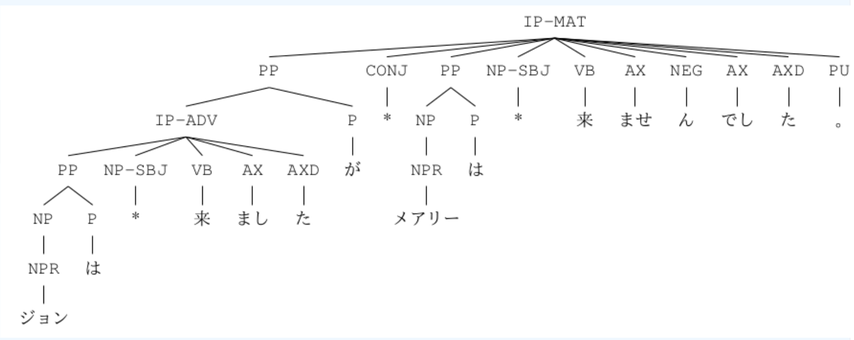
Table: ABC文法の規則

- この骨格部分だけからなる文法を便宜的に**ABC文法**と呼ぶ (AB文法 (Ajdukiewicz, 1935; Bar-Hillel, 1953) + Function Composition)
- ABC文法のツリーバンクを作ることで、CCGとTLGの両方に簡単に変換できるツリーバンクが得られる

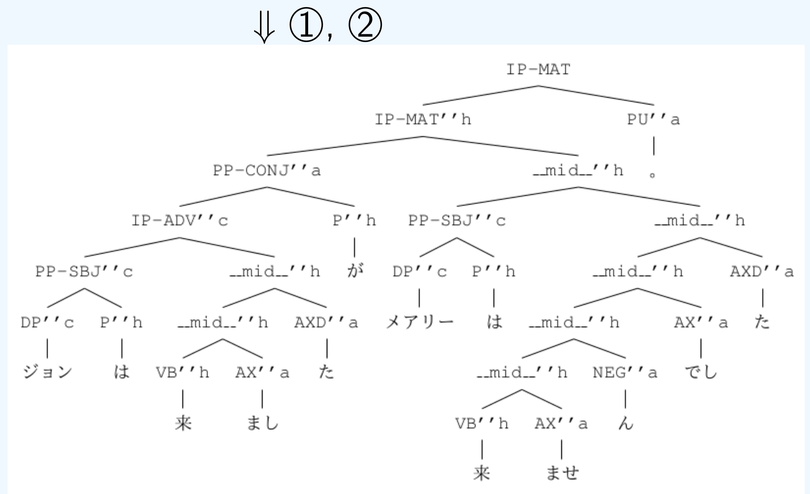
変換工程

大まかな変換工程

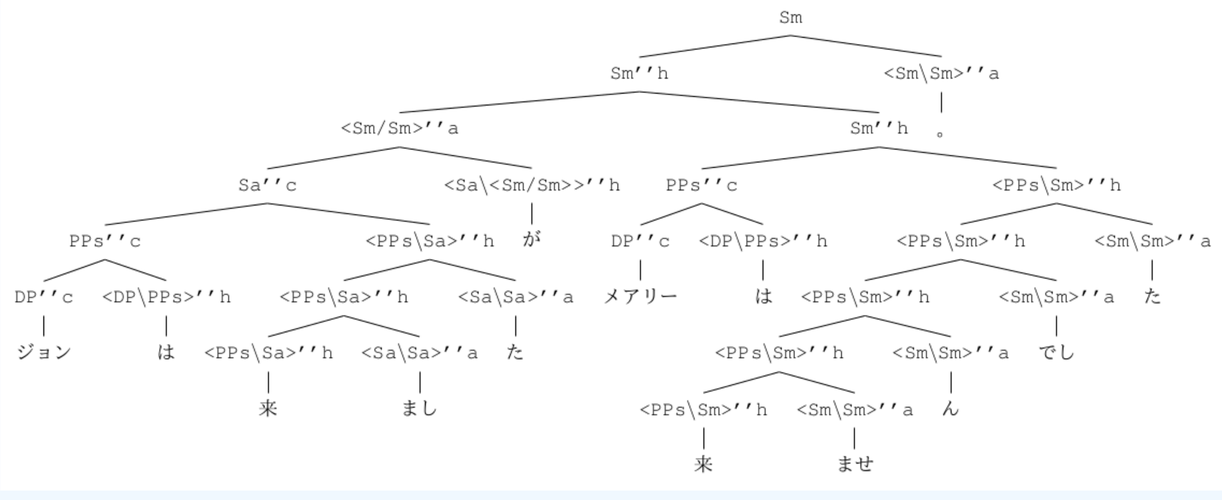




⇓ ①, ②

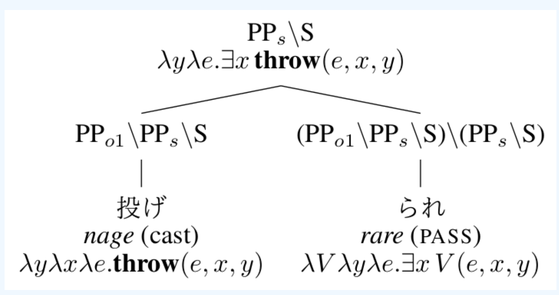


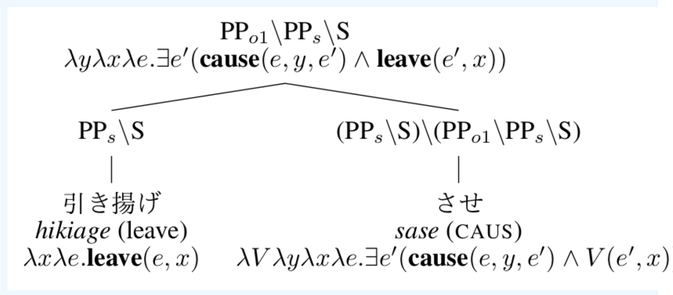
⇓ ③, ④

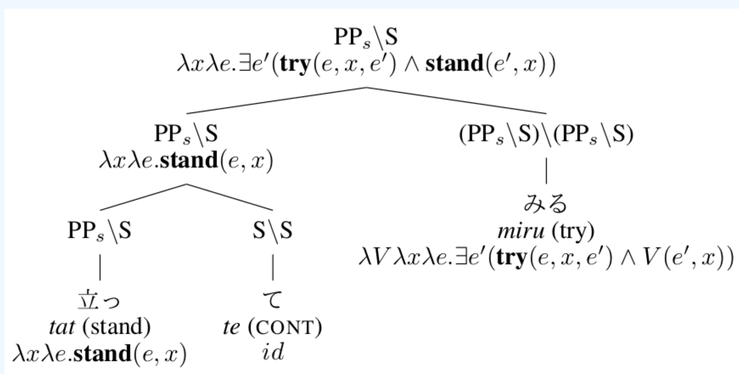


日本語CCGBank (Uematsu et al., 2013) との比較

- 受身、使役、コントロール述語は動詞の項構造を参照したカテゴリを付与 (日本語CCGBankでは文末述語はすべて S\S) ⇒ 意味解釈の導出が容易







今後の課題

- 意味解析を伴った大規模なパージングによる性能評価
- TLGパーザ構築 ⇒ 2021年度より研究に着手 (科研21K00541)

謝辞: 本研究は国立国語研究所共同研究プロジェクト「対照言語学の観点から見た日本語の音声と文法」、JSPS科研費18K00523, 15H03210の研究成果の一部である。

参考文献:

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

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

Bekki, D. (2010). *Nihon-Go Bunpoo-no Keishiki-Riron (Formal Theory of Japanese Grammar)*. Kurosio, Tokyo.

Butler, A., Yoshimoto, K., Hiyama, S., Horn, S. W., Nagasaki, I., and Kubota, A. (2017). The Keyaki Treebank Parsed Corpus. http://www.compling.jp/Keyaki/.

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.

Kubota, Y. and Levine, R. (2020). *Type-Logical Syntax*. MIT Press, Cambridge, MA. Available Open Access at https://direct.mit.edu/books/book/4931/Type-Logical-Syntax.

Kubota, Y., Mineshima, K., Hayashi, N., and Okano, S. (2020). Development of a general-purpose categorial grammar treebank. In *LREC2020*, pages 5195–5201.

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

Moortgat, M. (1997). Categorial type logics. In van Benthem, J. and ter Meulen, A., editors, *Handbook of Logic and Language*, pages 93–177. Elsevier, Amsterdam.

Moot, R. (2013). A type-logical treebank for French. In Duchier, D. and Parmentier, Y., editors, *High-Level Methods for Grammar Engineering, Workshop at ESSLLI 2013*, Düsseldorf, Germany.

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

Steedman, M. (1996). *Surface Structure and Interpretation*. MIT Press, Cambridge, MA.

Uematsu, S., Matsuzaki, T., Hanaoka, H., Miyao, Y., and Mima, H. (2013). Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1042–1051.