極小主義に基づく並列ツリーバンクの構築

野元 裕樹
東京外国語大学
nomoto@tufs.ac.jp

概要

本研究では，TUFS Asian Language Parallel Corpus のマレー語とインドネシア語のデータに対して極小主義統語論に基づく構成素構造の統語アノテーションを行い，並列ツリーバンクを構築した．それぞれ 1,386 文，1,385 文から成る．極小主義に基づくアノテーションは Penn Treebank 式句構造文法よりも詳細で，言語の性質を正確に捉えられる．特徴としては，厳格な二分股枝分かれ，内心構造の明示的表示，無形要素の多用，述語項構造の句構造への直接的反映，項と付加詞の明確な区別が挙げられる．

1 はじめに

一連の文に対して統語的アノテーションを付したツリーバンクは，単一の語やその周辺の数語だけからでは不可能な言語の特徴の把握を可能にするため，自然言語処理や言語学にとって重要な言語資源である．しかし，形態情報などのアノテーションに比べ，統語構造のアノテーションは難しいため，本格的なツリーバンクが存在する言語は限られている．本研究では，筆者らが構築したマレー語とインドネシア語のツリーバンクについて論じる[1)]．このツリーバンクは，TUFS Asian Language Parallel Corpus (TALPCo) [1] の両言語のデータに構成素構造のアノテーションを行ったものである．その結果，並列ツリーバンクとなっている．また，アノテーションの基礎となる文法理論として極小主義 (minimalism) の枠組み [2] を採用した．極小主義は，四半世紀以上に渡り，マレー語・インドネシア語を含む多くの言語の言語学的研究で最も広く採用されている，構成素構造を中核とする統語理論である [3].

1） マレー語 (ISO 639-3: zsm) とインドネシア語 (ISO 639-3: ind) は，広義のマレー語 (ISO 639-3: msa) の 2 つの地域変種である．前者はマレーシア，シンガポール，ブルネイの公用語であり，後者はインドネシアの公用語である．両者の違いは語彙や音韻が中心だが，文法においても細かな違いが存在する．

2 関連研究

筆者の知る限り，マレー語のツリーバンクは存在しない．インドネシア語には少なくとも 5 つのツリーバンクが存在する．まず，Penn Treebank（以下，PTB）式句構造文法 [4, 5] に依拠するものとして，Kethu: An Indonesian Constituency Treebank in the Penn Treebank Format [6] がある．このツリーバンクは Indonesian Treebank [7] の改訂版で，1,030 文のニュース文から成る．次に，主辞駆動句構造文法 (HPSG) に依拠するものとして，JATI [8] と Cendana [9] がある．前者は 1,253 の辞書の定義文（ほぼすべて名詞句），後者は旅行会社のオペレーターと顧客のチャット 715 文から成る．4 つ目は，Universal Dependencies [10, 11, 12] で，規模が最も大きい．フォーマルな会話 (GSD; 5,598 文)，ニュース・Wikipedia (PUD; 1,000 文)，ニュース (CSUI; Kethu からの変換 1,030 文) から成る．最後に，語彙機能文法 (LFG) に基づく ParGram Parallel Treebank (ParGramBank) [13] があるが，かなり小さい（79 文）．

3 TALPCo ツリーバンクの概要

本研究では，極小主義の枠組みに従い，TALPCo のデータに構成素構造のアノテーションを付与したが，このデータを便宜的に「TALPCo ツリーバンク」と呼ぶことにする．TALPCo ツリーバンクは，マレー語 1,386 文，インドネシア語 1,385 文から成る．

TALPCo のデータは，日本語文からの翻訳文である．日本語文は，日本語能力試験 N5 レベルの基礎語彙の学習のための例文で，フォーマルな会話で用いられる比較的短いものである（例：「帰る電車がなかったので，友達の家に泊まりました.」）．翻訳文も同様にフォーマルな口語体の文となっている．

アノテーションは筆者と東京外国語大学言語文化学部の学部生 4 名の計 5 名で行った．この学生達は，マレー語またはインドネシア語を専攻言語として学び，いずれも統語論の授業を履修済みである．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

学生が行ったアノテーションはすべて筆者が確認・修正を行った．作業開始前に筆者が基本的な構文をカバーしたアノテーションマニュアル（執筆言語は日本語）を作成した．マニュアルは随時更新し，2022 年 1 月時点で 109 ページに及ぶ．このマニュアルは，Indonesian Treebank の手引き（54 ページ，執筆言語はインドネシア語）[2] よりも充実している．

アノテーションのツールとして Syntax Tree Generator[3] を用いた．このツールでは，ラベル付き括弧表示を入力すると構文木が表示される（図 1）．図の上部のボックスにあるようなラベル付き括弧表示を作成し，アノテーション結果として表計算ファイルに貼り付けていった．すべての結果を TALPCo のホームページ[4] で公開する予定である．

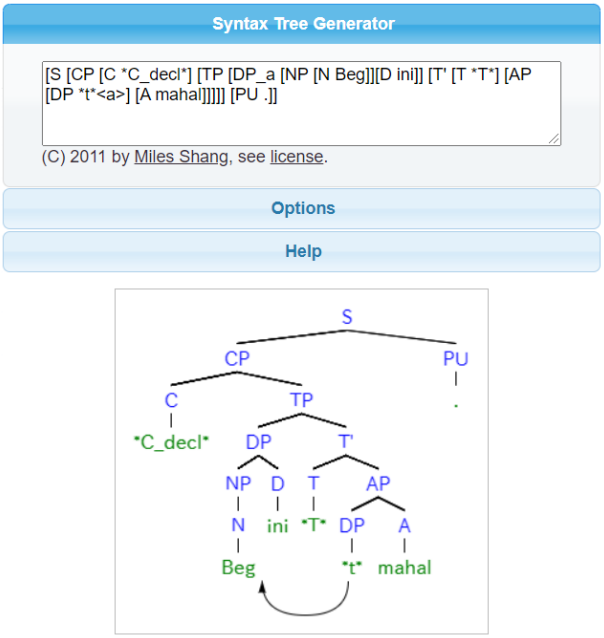

図 1　アノテーションの際に用いた Syntax Tree Generator

極小主義に基づく構文木は他の文法理論に比べて大きくなる．図 1 は，Beg ini mahal.（このかばんは高かったです）というマレー語文の統語構造を示したもので，非終端節点は 14 ある．Indonesian Treebank の手引きに従って，PTB 式に同じ文を分析すると図 2 のようになり，非終端節点の数は半分の 7 となる．表 1 は TALPCo ツリーバンクの非終端節点の数をまとめたものである．参考のために Kethu の情報も加えた．PTB 式のアノテーションにもかかわらず，Kethu の数値が大きくなるのは，平均文長が TALPCo よりもはるかに長いためである．

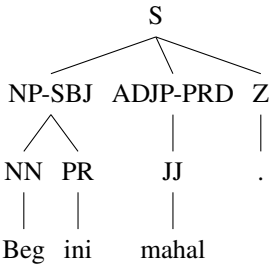

図 2　図 1 と同じ文を Penn Treebank 式に分析した構文木

表 1　TALPCo ツリーバンクと Kethu の非終端節点の数

コーパス	全体	文あたり平均
TALPCo マレー語	47,102	34.0
TALPCo インドネシア語	45,180	32.6
Kethu	58,023	56.3

4　アノテーションの特徴

本節では，TALPCo ツリーバンクのアノテーションの枠組みの特徴について論じる．その際，必要に応じて PTB 式との違いも指摘する．アノテーションにおいては，できる限り言語学における標準的分析を反映するようにした．しかし同時に，不必要に細部にこだわり過ぎないようにもした．これは，アノテーション作業を十分遂行可能なものにするためと，結果として得られるアノテーションが過度に複雑にならないようにするためである．このような実際的な妥協により，本来なら可能な分析が不可能になることもあった．その場合は，統語範疇に-PostV（vP 指定部が例外的に右側に出る）のようなフラグを付すなどして対応した．

4.1　二分肢枝分かれ

句構造の枝分かれは二分肢のみを原則とする．ただし，文末以外の句読点 (XP, YP; “XP,”) は例外として三分肢以上を認める．これは文法ではなく，正書法上の理由から生じる例外である．

二分肢枝分かれの原則は，極小主義における句構造生成のメカニズムである併合 (Merge) が 2 つの統語的構成物に対する操作であることによる．一方，PTB 式句構造文法では枝分かれ数に制限が存在しない．実際に Kethu では四分肢や五分肢の枝分かれ構造が見られる．そのようなアノテーションは容易だろうが，構成素構造を誤って表示しているため，アノテーションのためのアノテーションでしかない．

2）https://github.com/famrashel/idn-treebank/blob/master/BracketingGuidelines.pdf
3）http://mshang.ca/syntree/
4）https://github.com/matbahasa/TALPCo

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

4.2 内心構造

自然言語の句構造は通常，内心構造を示す．すなわち，XP は内部に必ず主要部／主辞 X を持つ．アノテーションではこのことが明示的に分かるようにする．ただし，アノテーション対象全体に対して付す S は例外である．この例外は，二分肢枝分かれの原則を守りつつ句読点をアノテーションに含めるために生じるもので，やはり正書法上の理由による．

PTB 式アノテーションでは自然言語の重要な特性である内心構造が不明瞭になることが多い．例えば，JJ（形容詞）が ADJP（形容詞句）の主要部であることはタグの形式だけからは分からない（図 2）．

4.3 無形要素

実際には発音されない無形要素を多用する．無形要素は極小主義の統語分析から来るものと，統語構造を基に行われる意味解釈を無理なく行うために仮定したものがある．例えば，後者の一つである無形の前置詞*selama*は名詞句 satu jam（一時間）を前置詞句（英 for an hour）として解釈できるようにする．

1. 統語分析に基づく無形要素
   (a) 空代名詞：*PRO*, *pro* (b) 空演算子：*Op* (c) 痕跡：*t* (d) いわゆるØ：*C*, *C_cont*, *C_decl*, *C_excl*, *C_imp*, *C_int*, *Top*, *Foc*, *T*, *v_tr*, *v_act*, *v_pass*, *v_intr*, *v_unerg*, *v_unacc*, *v_cop*, *v_eq*, *Appl*, *D*, *D_def*, *D_indef*, *exp*, *Poss*, *Num*, *PL*, *N*, *N_nmlz*, *Conj*
2. 意味解釈のための無形要素
   ada, *atau*, *dan*, *dari*, *dengan*, *di*, *hari*, *kalau*, *ke*, *pada*, *per*, *sebanyak*, *selama*, *untuk*, *yang*, *0*

4.4 移動（内的併合）

極小主義に基づく統語分析では，句構造は併合の操作により，ボトムアップで派生される．すでに構造上に存在する要素を再び併合することも可能で（内的併合），その場合，要素が「移動」する．移動元の要素を <> で示すと以下のようになる．

$$\text{Merge}(\gamma, \{\alpha, \{\beta, \gamma\}\}) \rightarrow \{\gamma, \{\alpha, \{\beta, <\gamma>\}\}\}$$

<> 内の要素は，移動の痕跡 (trace) と呼ばれ，発音はされない．TALPCo ツリーバンクでは痕跡は*t*で示し，対応する要素に添え字を付ける（図 1 参照）．後者はアノテーションツールの仕様によるものである．移動の考えは PTB 式でも疑問詞疑問文や受動文などで採用されている．TALPCo ツリーバンクではさらに多くの移動を取り扱う．特に重要なのが述語内主語仮説の採用に伴う主語の移動である．

述語内主語仮説とは，ある述語の項はすべてその述語が投射する句の中に生起するという仮説である．これにより，句構造から述語項構造を読み取ることが可能になる．例えば，図 1 では，形容詞 mahal（高価な）の投射する句 AP の中に DP の痕跡が存在し，mahal は DP 項を一つ要求することが分かる．そして，添え字 a により当該項は移動して表層の主語として生起することが分かる．また，主語は TP 指定部という構造上の位置として定義できるため，PTB 式アノテーションの-SBJ のような特別なフラグで主語にアノテーションをする必要はない．

さらに，述語内主語仮説によりマレー語・インドネシア語などに特有の裸受動構文 [14] がうまく分析できる．図 3 は，Gambar itu saya ambil pada bulan lepas（あの写真は先月撮りました）というマレー語の裸受動文の構造を示したものである．被動作主 gambar itu（あの写真）は日英語の受動文と同様に表層の主語の位置に移動するが，動作主 saya（私）は日英語のように付加詞 PP となはならず，名詞句の形で基底の位置にとどまる．

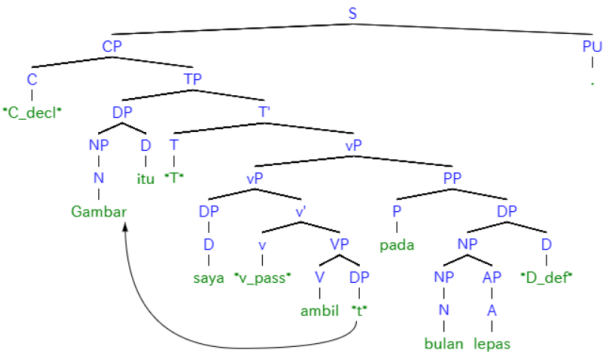

図 3 裸受動文の構造

4.5 項と付加詞の区別

述語内主語仮説により，述語の項はその述語が投射する句の中に生起する．一方，付加詞はその外側に位置し，同じ句ラベルを繰り返す．この違いにより述語の項と付加詞の区別が句構造から読み取れる．例えば，図 3 では，動詞 ambil（撮る）の項である動作主 saya（私）と被動者 gambar itu（あの写真）（の痕跡）は動詞句 vP 内に位置する．一方，付加詞

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

pada bulan lepas（先月に）はその vP の外側に現れる．PTB 式だと，すべて VP の直下に現れて，項と付加詞の区別は付かない（図 4）[5]．

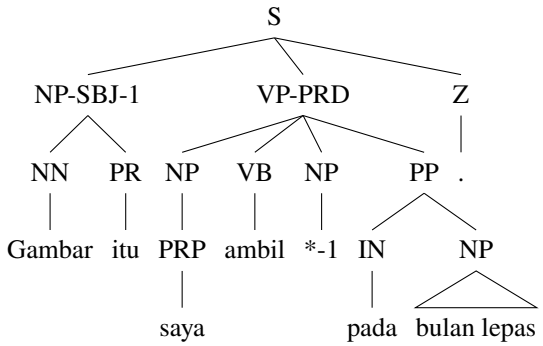

図 4　図 3 と同じ文を Penn Treebank 式に分析した構文木（紙幅の都合上 bulan lepas の内部構造は省略）

4.6　トークン化と POS タグ

マレー語・インドネシア語のトークン化では普通，語と接語の分割を行うが，極小主義統語論に基づく TALPCo ツリーバンクではより小さな単位へのトークン化が必要になる．具体的には，以下の接辞が新たにトークン化の対象となる．

1. 態を表す接辞：meN-（能動態），di-（受動態），-kan（受益者適用態），-i（場所適用態）
2. 名詞句に付いて「〜を持つ，伴う」の意味を持つ動詞接頭辞 ber-

併合は 2 つの統語的構成物を対象とするため，極小主義の句構造には非枝分かれ節点は本来的に生じない．つまり，[XP [X α]] はあり得ず，それは [XP α] となる．そのような句構造を裸句構造 (bare phrase structure) という．しかし，TALPCo ツリーバンクでは裸句構造は採用せず，終端節点すなわち構文木の葉の部分は必ず XP でなく X の形になるようにした．これは葉の部分から簡単に POS タグの情報を得られるようにするという言語資源としての有用性を考慮してのことである．TALPCo ツリーバンクに用いられている POS タグを表 2 にまとめる．

5　おわりに

TALPCo ツリーバンクは，マレー語では初のツリーバンクであり，インドネシア語では既存の構成素構造に基づくツリーバンクと肩を並べるものと言える．アノテーションの形式は基本的に PTB と同じであるため，[ ] を ( ) に変換することで Tregex[15] などの PTB 式句構造文法に従ったツリーバンク用に開発されたツールを利用できる．言語分析は極小主義統語論に基づいており，PTB 式句構造文法よりも言語の性質をうまく捉えることができている．そのため，言語学の研究や教育においての利用が期待できる．一方，自然言語処理の分野では PTB 式句構造文法や依存文法が主流であり，英語の大規模言語資源が登場した段階で分野全体の言語分析が固定化する傾向が強いため，マレー語・インドネシア語の自然言語処理で TALPCo ツリーバンクが直接利用されることは考えにくい．そこで，PTB 式や組合せ範疇文法 (CCG) への変換を通しての利用が考えられる．二分股枝分かれを基本とし，意味解釈を考慮に入れている TALPCo ツリーバンクは CCG との相性がよいと言えよう．

TALPCo ツリーバンクはサイズは決して大きくないものの，より大規模なツリーバンクの構築への足掛かりとなる．極小主義は言語学の統語論で最も広く採用されている理論であり，統語論の授業を履修した学部上級生から大学院生であれば，アノテーションガイドを参照しつつアノテーション作業を行うことができる．さらに，今後アノテーションマニュアルを翻訳することで，母語話者によるアノテーションにもつなげたい．

TALPCo には他に日本語，朝鮮語，タイ語，ベトナム語，ビルマ語，英語が含まれる．これらの言語についても同様のアノテーションを付すことで，より有益な並列ツリーバンクが構築できる．すでに英語のデータの一部についてアノテーションを行った．

表 2　TALPCo ツリーバンクにおける POS タグの分布

タグ	zsm	ind	タグ	zsm	ind
A	578	553	Nume	299	292
Adv	298	309	P	1,214	1,167
Appl	12	22	Part	19	8
C	1,875	1,797	Poss	243	256
Conj	208	178	PU	1,592	1,596
D	4,787	4,645	Q	70	59
Foc	42	6	T	1,885	1,787
Int	29	28	TITLE	197	207
Mod	221	156	Top	44	67
N	2,715	2,606	v	1,638	1,570
Neg	123	134	V	1,509	1,435
Num	162	150			

5）Indonesian Treebank の手引きには裸受動文のアノテーション法に関する記載がない．また，Kethu のアノテーションを用いて裸受動文を抽出する方法もなさそうである．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JSPS 科研費 JP18K00568 および JP20H01255 の助成を受けた.

参考文献

[1] Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. TUFS Asian Language Parallel Corpus (TALPCo). In **Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing**, pp. 436–439, 2018.

[2] Noam Chomsky. **The Minimalist Program**. MIT Press, Cambridge, MA, 1995.

[3] Andrew Carnie. **Syntax: A Generative Introduction**. Wiley-Blackwell, Oxford, 4th edition, 2021.

[4] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn Treebank: Annotating predicate argument structure. In **Proceedings of the Workshop on Human Language Technology**, pp. 114–119, 1994.

[5] Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Marcinkiewicz, and Britta Schasberger. Bracketing guidelines for Treebank II style Penn Treebank project, 1995.

[6] Jessica Naraiswari Arwidarasti, Ika Alfina, and Adila Alfa Krisnadhi. Converting an Indonesian constituency treebank to the Penn Treebank format. In **2019 International Conference on Asian Language Processing (IALP)**, pp. 331–336, 2019.

[7] Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In **The Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014**, pp. 66–69, 2014.

[8] David Moeljadi. **An Indonesian Resource Grammar (INDRA) and Its Application to a Treebank (JATI)**. PhD thesis, Nanyang Technological University, 2017.

[9] David Moeljadi. Building Cendana: A treebank for informal Indonesian. In **Proceedings of 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33)**, pp. 156–164, 2019.

[10] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tackstrom, Claudia Bedini, Nuria Bertomeu Castello, and Jungmee Lee. Universal Dependency annotation for multilingual parsing. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**, pp. 92–97, 2013.

[11] Ika Alfina, Arawinda Dinakaramani, Mohamad Ivan Fanany, and Heru Suhartanto. A gold standard dependency treebank for Indonesian. In **Proceedings of 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33)**, pp. 1–9, 2019.

[12] Ika Alfina, Indra Budi, and Heru Suhartanto. Tree rotations for dependency trees: Converting the head-directionality of noun phrases. **Journal of Computer Science**, Vol. 16, No. 11, pp. 1585–1597, 2020.

[13] Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, Gyöorgy Rákosi, Cheikh Bamba Dione, Helge Dyvik, Victoria Rosen, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoğlu, I Wayan Arka, and Meladel Mistica. ParGramBank: The ParGram parallel treebank. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**, pp. 550–560, 2013.

[14] Hiroki Nomoto. Bare passive agent hierarchy. In Henrison Hsieh and Keely New, editors, **Proceedings of the Twenty-Seventh Meeting of the Austronesian Formal Linguistics Association**, pp. 57–70, Ontario, 2021. University of Western Ontario.

[15] Roger Levy and Galen Andrew. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In **5th International Conference on Language Resources and Evaluation (LREC 2006)**, 2006.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).