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）

2022/3/15 @言語処理学会第28回年次大会
(NLP2022)

概要

- TUFS Asian Language Parallel Corpus (TALPCo) [1]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 (minimalism) 統語論** [2] 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
- 便宜上、「TALPCoツリーバンク」と呼ぶ

[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. https://github.com/matbahasa/TALPCo
[2] Noam Chomsky. *The Minimalist Program*. MIT Press, Cambridge, MA, 1995.

TALPCo

- TUFS Asian Language Parallel Corpus (TALPCo) [1]
- https://github.com/matbahasa/TALPCo
- **日本語文から**の翻訳文
- 日本語文
  - 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
  - **フォーマルな会話**で用いられる比較的短い文

[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. https://github.com/matbahasa/TALPCo

TALPCo並列データの例 (#3756)

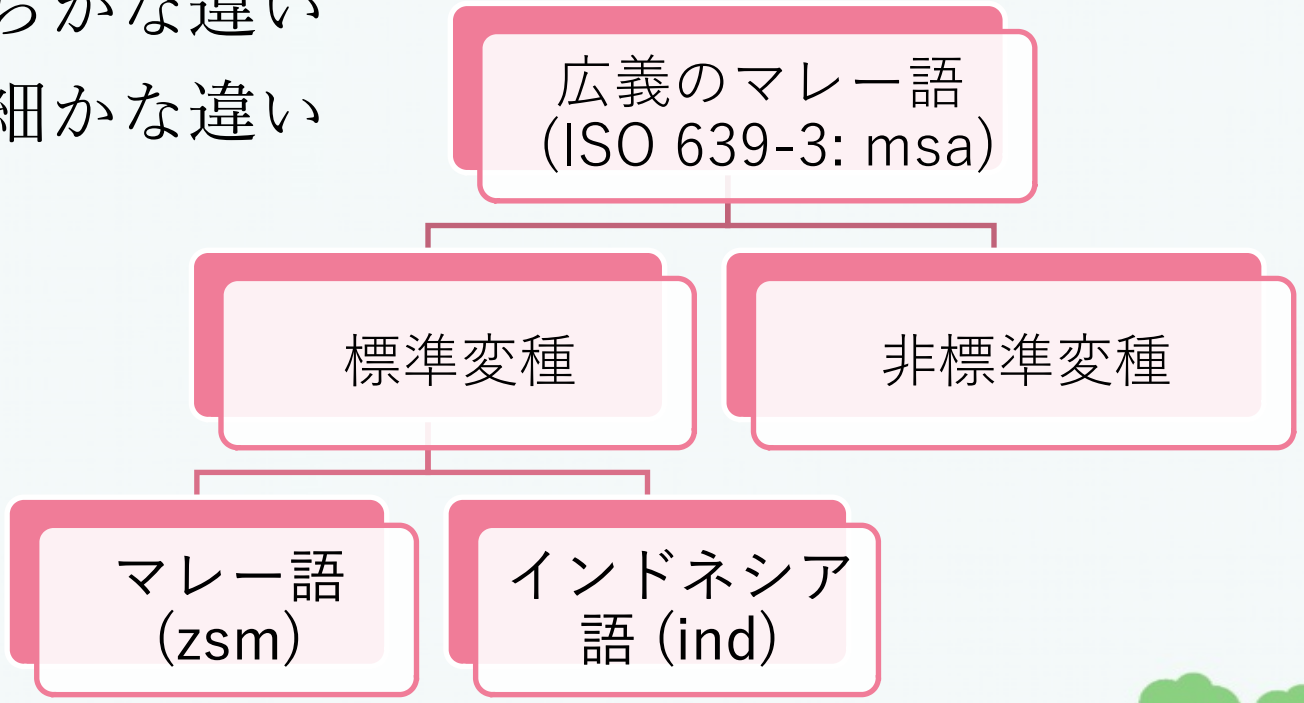
［イさんに対しての発話］**代名詞代用表現 (pronoun substitute)**の使用
（cf. D1-2 岡野+「アジア三言語における代名詞代用・呼びかけ語の共通項目調査」）

- 日本語：　イさんは どっちの 飲み物が いいです
- 朝鮮語：　이 씨는 어느 쪽 음료수가 좋습니까?
- マレー語：　Encik Lee nak minuman yang mana?
- インドネシア語：　Pak Lee mau minuman yang mana?
- タイ語：　คุณอิรับเครื่องดื่มอันไหนดีคะ
- ベトナム語：　Anh Lee thích loại đồ uống nào?
- ビルマ語：　မစ္စတာလီ ဘယ်သောက်စရာ သောက်မလဲ။
- 英語：　Which drink would you like?　（×Mr. Lee）

英語・中国語データには
（ほぼ）存在しない、
人称の曖昧性

マレー語とインドネシア語

- 語彙や音韻に明らかな違い
- 文法においても細かな違い



マレーシア、シンガポール、ブルネイ　　インドネシア

マレー語のツリーバンク

- 筆者の知る限り、TALPCoツリーバンクが初

インドネシア語のツリーバンク

	ツリーバンク	文法理論	ドメイン	サイズ
①	Kethu: An Indonesian Constituency Treebank in the Penn Treebank Format [6]	Penn Treebank (PTB) 式句構造文法	ニュース	1,030文
②	JATI [8]	HPSG	辞書の定義文（ほぼすべて名詞句）	1,253文
③	Cendana [9]	HPSG	旅行会社のオペレーターと顧客のチャット	715文
④	Universal Dependencies [10, 11, 12]	依存文法	フォーマルな会話 ニュース・Wikipedia ニュース (①から変換)	5,598文 1,000文 1,030文
⑤	ParGram Parallel Treebank (ParGramBank) [13]	LFG	言語学者の作例？	79文

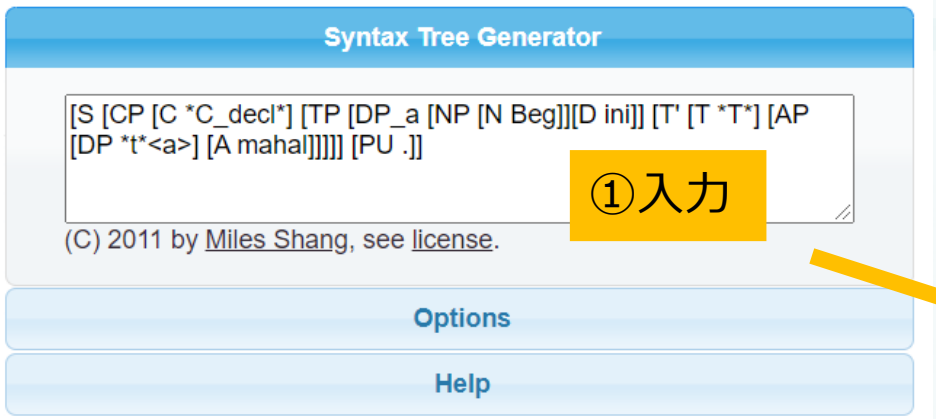
TALPCoツリーバンクの概要

- 文法理論：極小主義　←言語学で最も広く採用
- ドメイン：フォーマルな口語体
- サイズ：
  - マレー語：1,386文
  - インドネシア語：1,385文　cf. Kethuは1,030文
- アノテーター：発表者＋東京外大学部生4名
  （マレー語・インドネシア語専攻、統語論履修済み）
- マニュアル：日本語で109ページ

  cf. Kethuの基になったIndonesian Treebank [7] の手引き：インドネシア語で54頁

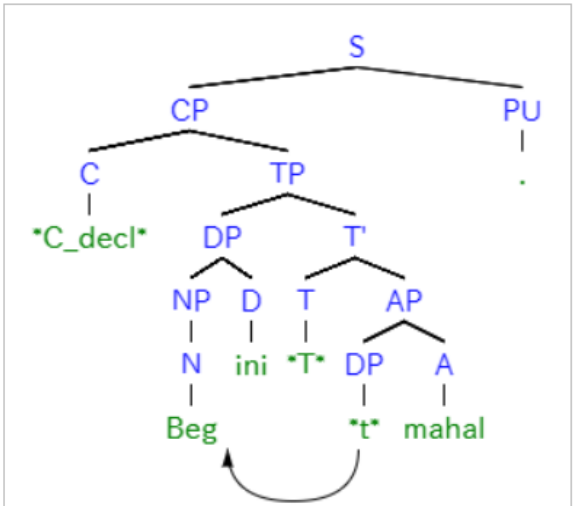
[7] Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In *The Proceedings of the International Conference on AsianLanguage Processing 2014, IALP 2014*, pp. 66–69, 2014. https://github.com/famrashel/idn-treebank

アノテーションの方法



1. Syntax Tree Generator
   http://mshang.ca/syntree/
2. Googleスプレッドシート

②貼り付け



	A	B	C	D	
1	ID	文	括弧表示	リンク	担当
2	1176	Encik Tanaka bukan pelajar.	[S [CP[C *C_dec	http://mshang.ca	松浦
3	1178	Ayah saya guru.	[S [CP [C *C_d	http://mshang.ca	太田
4	1180	Sekolah cuti.	[S [CP [C *C_de	http://mshang.ca	松浦
5	1194	Cuaca di Tokyo cerah.	[S [CP [C *C_de	http://mshang.ca	太田
6	1222	Ada pokok di taman.	[S [CP[C *C_dec	http://mshang.ca	松浦
7	1229	Encik Tanaka ada di mana?	[S [CP [C *C_int	http://mshang.ca	太田
8	1233	Saya tidak ada wang.	[S [CP [C *C_de	http://mshang.ca	Nomoto

サイズ：非終端節点の数

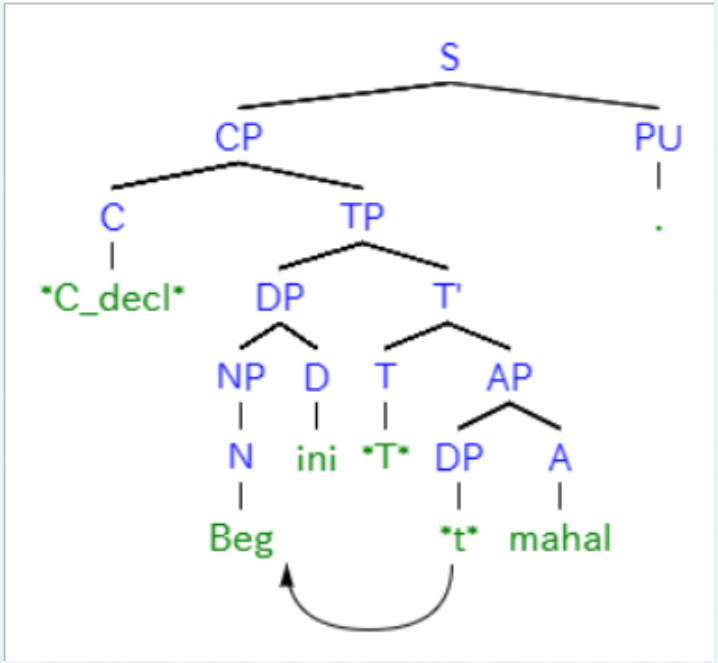
コーパス（サイズ）	全体	文あたり平均
TALPCo マレー語 (1,386文)	47,102	34.0
TALPCo インドネシア語 (1,385文)	45,180	32.6
Kethu [6] (1,030文)	58,023	56.3

Kethuはニュース文から成るため、平均文長がかなり長い。

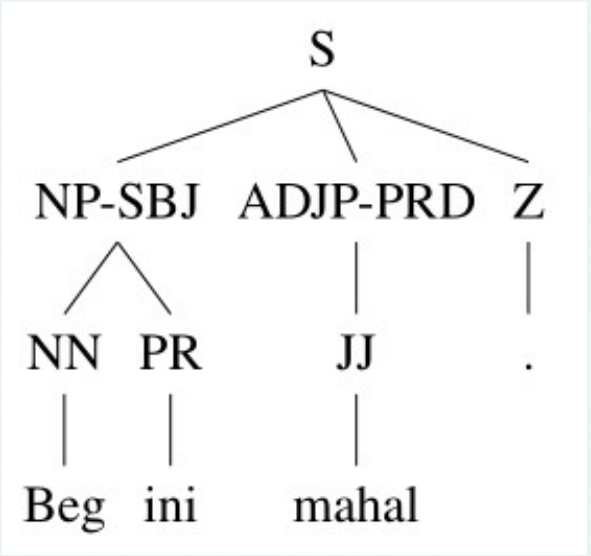
[6] Jessica Naraiswari Arwidarasti, Ika Alfina, and Adila Alfa Krisnadhi. Converting an Indonesian constituency treebank to the Penn Treebank format. In *2019 International Conference on Asian Language Processing (IALP)*, pp. 331–336, 2019. https://github.com/ialfina/kethu

Beg ini mahal.（このかばんは高かったです）

極小主義：非終端節点7



PTB式句構造文法：非終端節点4



アノテーションの基本的指針

- できる限り**言語学における標準的分析**を反映する。
- しかし同時に、不必要に**細部にこだわり過ぎない**。
  - アノテーション作業を十分遂行可能なものにするため
  - 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
- 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
  →統語範疇に**フラグ**を付すなどして対応
  例：-PostV（vP 指定部が例外的に右側に出る）

アノテーションの特徴

1.二分肢枝分かれ
2.内心構造
3.無形要素
4.移動（内的併合）
5.項と付加詞の区別
6.トークン化と POS タグ

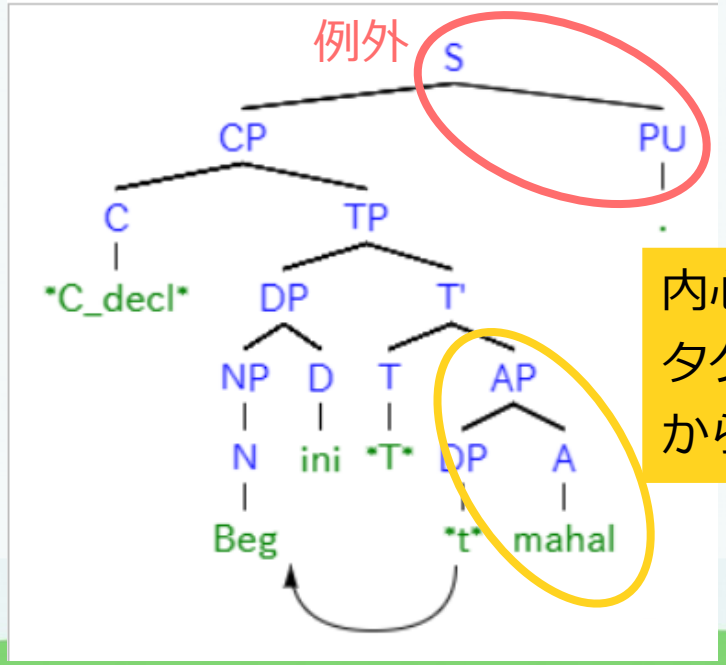
①二分肢枝分かれ

- 例外：文末以外の句読点
  - XP , YP
  - “ XP . ”
- 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。
- PTB 式句構造文法には枝分かれ数に制限なし
- Kethu には四分肢や五分肢の枝分かれ構造が存在
  - 誤った構成素構造
  - 言語の性質を軽視した、アノテーションのためのアノテーション

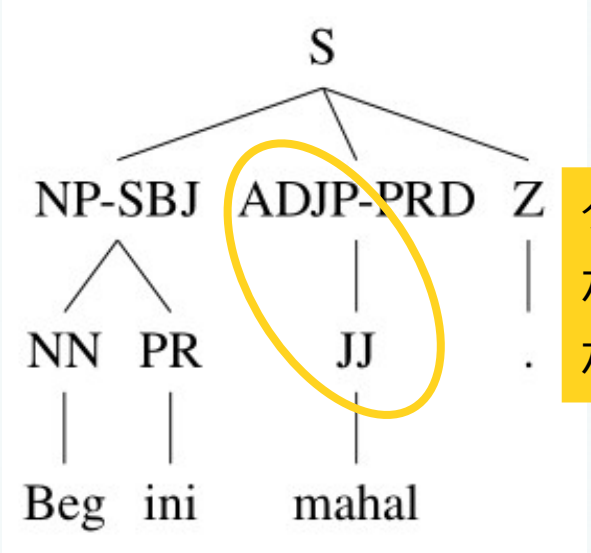
②内心構造

内心構造：XP は内部に必ず主要部／主辞 X を持つ

例：AP（形容詞句）の主要部＝A（形容詞）



内心構造がタグの形式から分かる



タグの形式だけからは内心構造が分からない

③無形要素

- 実際には発音されない無形要素を多用

1.極小主義の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*

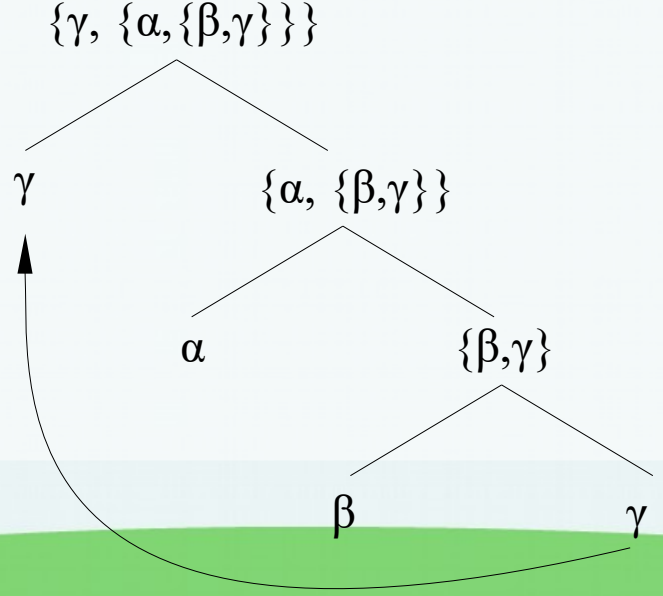
2.意味解釈のための無形要素

ada, *atau*, *dan*, *dari*, *dengan*, *di*, *hari*, *kalau*, *ke*, *pada*, *per*, *sebanyak*, *selama*, *untuk*, *yang*, *0*

例：*selama* satu jam [*for* one hour]

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
- 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（内的併合）＝要素が「移動」



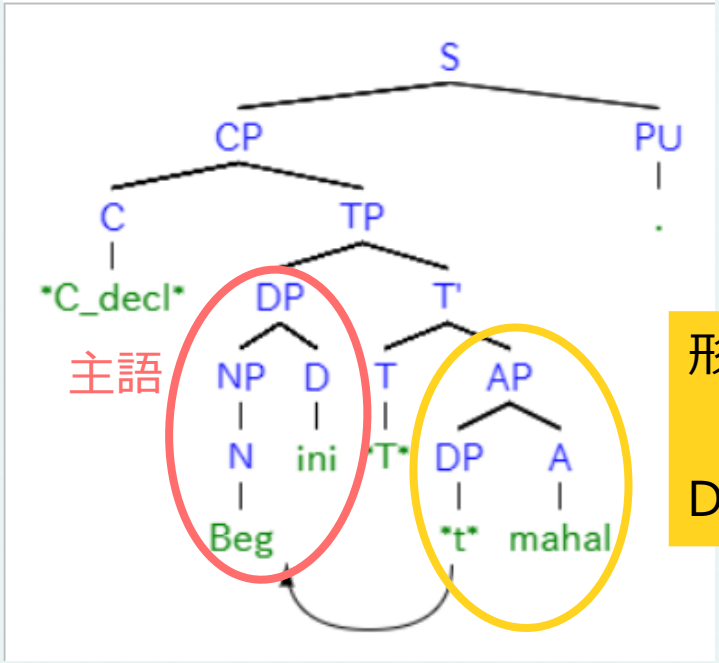
Merge(γ, {α, {β,γ}})　type? → internal
which node? → γ

Merge(α, {β,γ})　type? → external [default]

Merge(β,γ)　type? → external [default]

述語内主語仮説

- 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
- 句構造から述語項構造を読み取ることが可能に



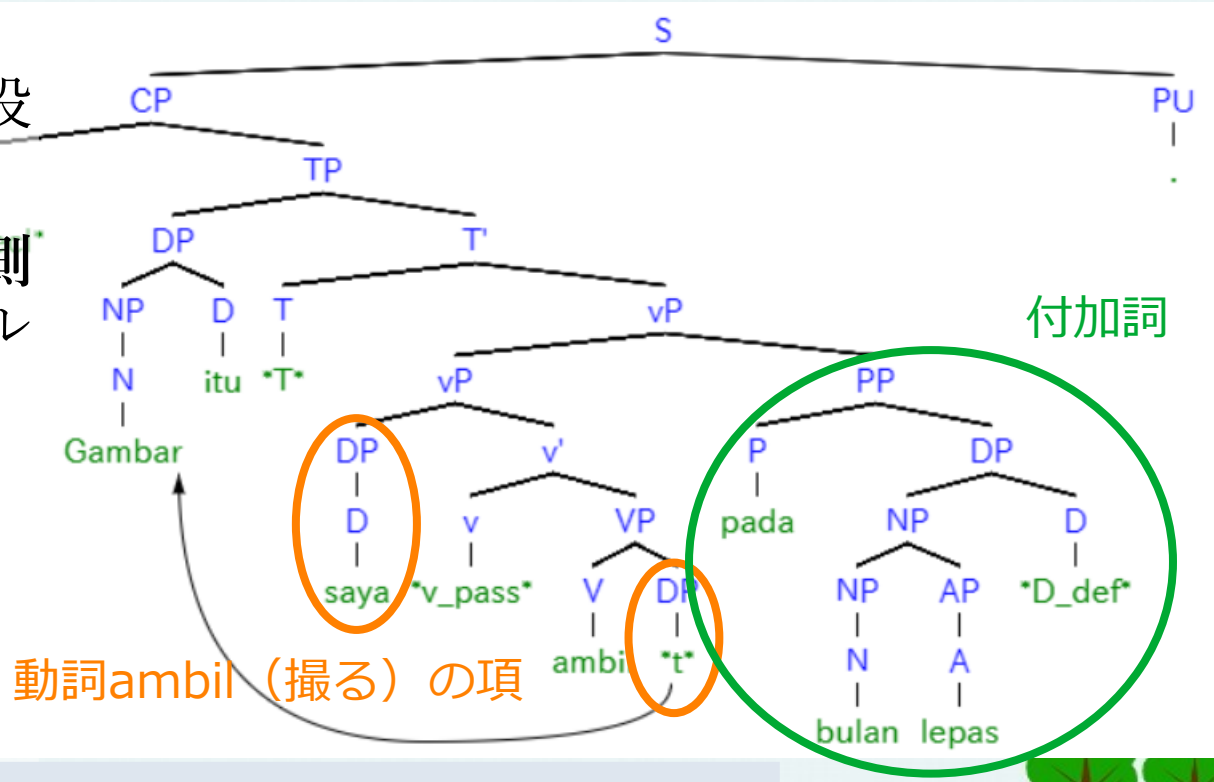
形容詞mahal（高価な）はDP項を1つ要求

[TP [DP_a [NP [N Beg]] [D ini]] …
[AP [DP *t*<a>] [A mahal]]]]

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**句の内部**に生起。
- 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。

→項と付加詞の区別が構造木から読み取れる

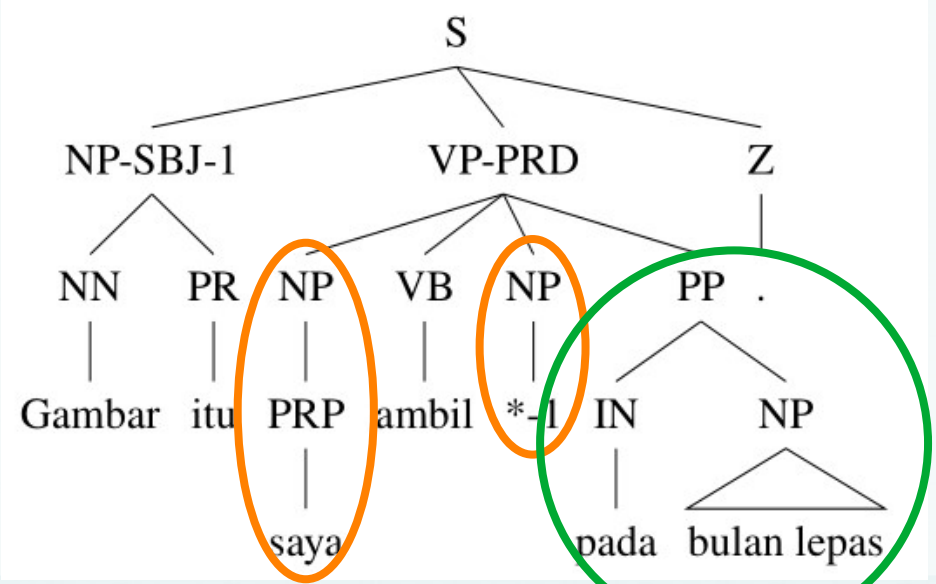


Gambar itu saya ambil pada bulan lepas.
（あの写真は先月撮りました）

PTB式句構造文法の場合

すべて VP の直下

→項と付加詞の区別が付かない



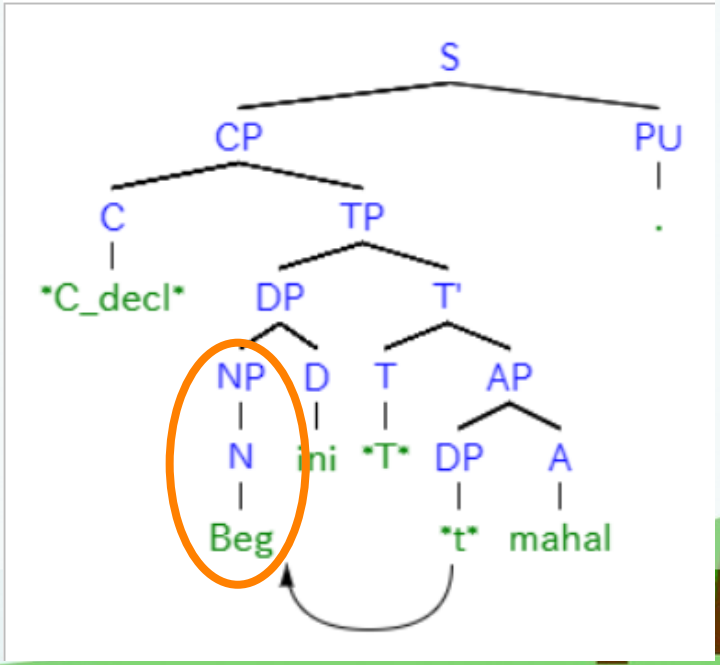
⑥トークン化とPOSタグ

マレー語・インドネシア語NLPのトークン化では普通、語と接語の分割を行うが、 言語学における標準的分析に合わせ、より小さな単位へのトークン化が必要

1. 態を表す接辞
    a) meN-（能動態）
    b) di-（受動態）
    c) -kan（受益者適用態）
    d) -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-

POSタグ情報の取得

- 終端節点（＝構文木の葉）は必ずXPでなくXの形
- 極小主義における句構造生成のメカニズムから逸脱
  2つの要素の併合
  Merge(α, β)
- 言語資源としての有用性を重視



まとめ

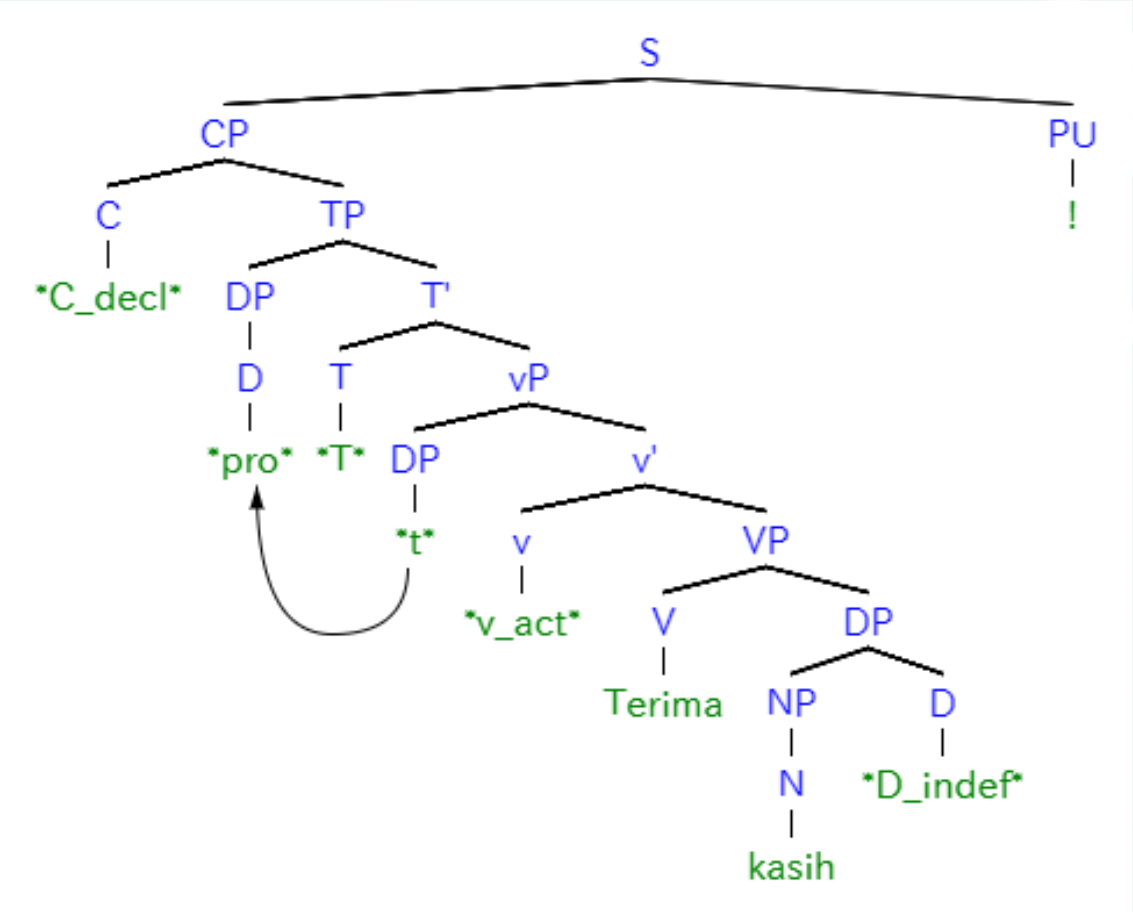
- TALPCo ツリーバンクは、マレー語では初のツリーバンクであり、インドネシア語では既存の構成素構造に基づくツリーバンクと肩を並べるものと言える。
- サイズは小さいものの、より大規模なツリーバンクの構築への足掛かりとなる。
- アノテーションの形式は基本的に PTB と同じ
  → PTB 式句構造文法に従ったツリーバンク用のツールを利用可能
  （例：Tregex [15] ）
- 極小主義統語論に基づくため、
  - PTB 式句構造文法よりも言語の性質をうまく捉えることができている。
  - 統語論の授業を履修した学部上級生〜大学院生であれば、アノテーションガイドを参照しつつアノテーション作業を行うことができる。

[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.

利用法

- 言語学

  研究・教育において、そのままの形での利用が期待できる
- 自然言語処理
  - PTB 式句構造文法や依存文法が主流
  - 英語の大規模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
  - TALPCo ツリーバンクの直接利用は考えにくい
  - 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



ご清聴ありがとうございます！

謝辞

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