統語的語の認定問題

京都大学
村脇 有吾

Universal Dependenciesシンポジウム@国語研 (2019年9月4日)

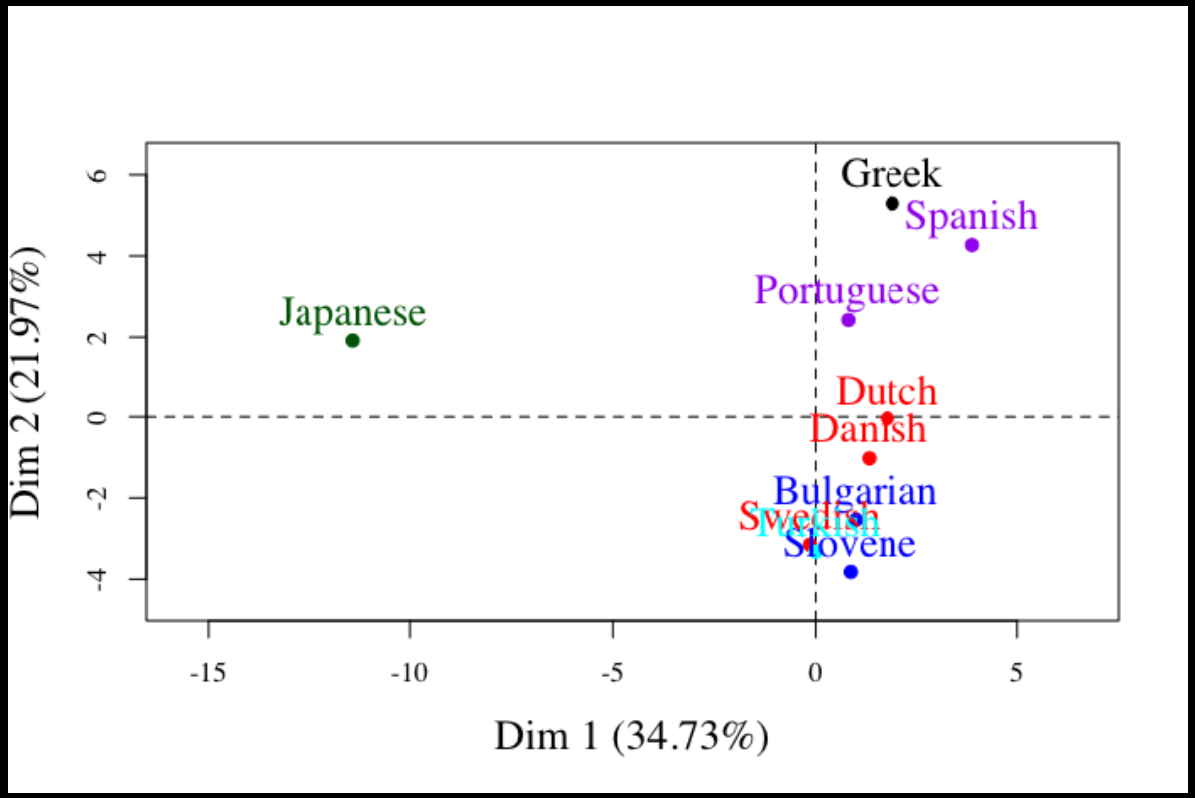
自己紹介: 村脇 有吾

- 京都大学大学院情報学研究科知能情報学専攻 黒橋・河原研究室 助教 (2016-)
- 研究的背景は計算機科学
- 主な研究課題
  - 表: 常識的知識を用いた言語解析とその応用
  - 裏: 言語類型論の特徴の統計的分析 (2014-2019)
- UDプロジェクトではアームチェアXXX

UDへの個人的な期待

- ビッグウェーブに乗って日本語処理の中長期的な生き残りを
  - 当面の言語解析の研究はKyoto *X* Corpusで間に合っている
- データ駆動型の多言語比較の実現
  - これまでの定量的比較には、言語学者が作った言語類型論の特徴 (e.g. 基本語順) を利用
  - 同一パラメータ空間上でのパーサの学習や教師なしモデルの評価

言語の特性 vs アノテーション基準の恣意性



[Cohen+, EMNLP2011]　　※文脈自由文法のモデルで、依存文法ではない

今日のお話:
日本語の統語的語 (syntactic word)

- 係り受けの単位を他の言語と揃えたい
- 現行UD Japaneseが採用する国語研短単位は統語的語ではないし、統語的語が日本語処理で採用されたこともない
- 統語的語に基づくアノテーション方針と解析パイプラインの提案
- しかし費用・便益比は微妙では?

Computer Science > Computation and Language

On the Definition of Japanese Word

Yugo Murawaki

(Submitted on 24 Jun 2019)

The annotation guidelines for Universal Dependencies (UD) stipulate that the basic units of dependency annotation are syntactic words, but it is not clear what are syntactic words in Japanese. Departing from the long tradition of using phrasal units called bunsetsu for dependency parsing, the current UD Japanese treebanks adopt the Short Unit Words. However, we argue that they are not syntactic word as specified by the annotation guidelines. Although we find non-mainstream attempts to linguistically define Japanese words, such definitions have never been applied to corpus annotation. We discuss the costs and benefits of adopting the rather unfamiliar criteria.

Subjects: **Computation and Language (cs.CL)**

Cite as: **arXiv:1906.09719 [cs.CL]**

(or **arXiv:1906.09719v1 [cs.CL]** for this version)

Bibliographic data

[Enable Bibex (What is Bibex?)]

cjvlang

Thoughts on the Universal Dependencies proposal for Japanese

The problem of the word as a linguistic unit

18 September 2016 (with some later modifications)

The following is a rather long and rambling rumination on the current implementation of Universal Dependencies for Japanese from a general linguistic point of view.

http://www.cjvlang.com/Spicks/udjapanese.html

UDにおける統語的語

The UD annotation is based on a lexicalist view of syntax, which means that dependency relations hold between words. Hence, morphological features are encoded as properties of words and there is no attempt at segmenting words into morphemes. However, it is important to note that the basic units of annotation are syntactic words (not phonological or orthographic words), which means that we systematically want to split off clitics, as in Spanish dámelo = da me lo, and undo contractions, as in French au = à le. We refer to such cases as multiword tokens because a single orthographic token corresponds to multiple (syntactic) words. In exceptional cases, it may be necessary to go in the other direction, and combine several orthographic tokens into a single syntactic word.
…後略…

http://universaldependencies.org/u/overview/tokenization.html

3種類の語

- orthographic word: 分かち書きの単位
  - 日本語には参考にならない
- phonological word: 音韻的まとまり
  - 音声合成等を行わずテキスト処理に留まる限り無視できる
- syntactic word: 統語的まとまり

語・接語・接辞

- Word (by itself) (語)
- Clitic (接語): 他の語に寄りかかるが語の一種
  - Proclitic (後接語): 直後の語によりかかる (語の直前にくる)
    - 仏: je t'aime
  - Enclitic (前接語): 直前の語によりかかる (語の直後にくる)
- Affix (接辞): 単独では語をなさない
  - Prefix (接頭辞)
  - Suffix (接尾辞)

接語・接辞の認定基準例

[服部, 1950]

服部原則

1. 職能や語形変化の異なる色々の自立形式につくものは自由形式 (すなわち「附属語」(=接語)) である
2. 二つの形式の間に別の単語が自由に現れる場合には、その各々は自由形式である。従って、問題の形式は附属語(=接語)である。
3. 結びついた二つの形式が互に位置を取りかえて現れ得る場合には、両方ともに自由形式である

附属語(=接語)と認定されなかった形式が消極的に附属形式(=接辞)と認定される

接語・接辞の具体例 1/2

[宮岡, 2015]

- =が: 前接語
  - 名詞との間に =こそ, =だけ を挿入できる
- =だ: 前接語
  - ゴミ=だらけ=だ, 静か=だ, 書く=だけ=だ
- -た/-だ: 屈折接尾辞
  - 動詞語基 (派生をおえたもの) に後続しパラダイムをなす閉じた集合の1要素
  - 食べ-た, 食べ-られ-た, 食べ-させ-た, 食べ-させ-られ-た

接語・接辞の具体例 2/2

[宮岡, 2015]

- -ます: 派生接尾辞 (VV型)
  - 食べ-ます, 食べ-まし-た
- -さ: 派生接尾辞 (VN型)
  - 美し-さ, 食べ-にく-さ
- -らしい: 派生接尾辞 (NV型)
  - 男-らしい, 男-らし-そう=だっ-た
  - cf (女ではなく) 男=らしい, 男=だけ=らしい
- -はじめる: 派生接尾辞
  - 泳ぎ-はじめ-た, 泳ぎ-づかれ-はじめ-た

学校文法の分類 (国語研体系のベース)

- 自立語: 単独で文節を作れるもの
- 付属語: 単独で文節を作れないもの
  - 助動詞: 活用するもの
  - 助詞: 活用しないもの

接語・接辞の区別と直交的

国語研の諸単位は
いずれも統語的語と一致しない

短単位: ❙彼女❙の❙美し❙さ❙が❙失わ❙れ❙ない❙

統語的語: ❙彼女❙の❙美し　さ❙が❙失わ　れ　ない❙

長単位: ❙彼女❙の❙美し❙さ❙が❙失わ❙れ❙ない❙

文節: ❙彼女　の❙美し　さ　が❙失わ　れ　ない❙

従来の日本語処理

- 形態素への分割 → 文節へのチャンキング → 文節単位の係り受け
  - JUMAN/ChaSen/MeCab, YamCha, KNP/CaboCha
- 形態素に明確な基準はなく、工学的妥協の産物
- 辞書に載っていることが事実上の基準

従来の日本語コーパス研究

- (人力の) 語彙調査のための規定
- 操作主義的な単位設計
  - これこれこういうものを「〜 単位」とする，という規定をするだけで， その「〜 単位」が言語学的にどのようなものなのか， 単語なのか， 単語でないとすれば，どこが単語とちがうのか， といった問題には， まったくふれない (国立国語研究所1987:11)

統語的語に基づく アノテーション・解析の提案

- 短単位を (近似的に) 形態素とみなし、短単位の1個以上の結合を統語的語とする
- いわゆる付属語を接語・接辞に分類
  - ほとんどの場合は辞書項目レベルで対応可能なはず
  - 例外: =らしい/-らしい の識別はコーパス中の各出現に対して行わなければならない
- 統語的語の一部となった接辞の情報は素性 (feature) を使って表現
- 解析パイプライン: (1) 短単位への分割、(2) 統語的語へのチャンキング、(3) 係り受け

素性: UD Turkishの例

6 yapamazlar yap VERB Verb Aspect=Imp|Mood=Pot|Number=Plur|Person=3|Polarity=Neg|Tense=Aor 0 root

FORM: yapamazlar
((彼らが) できない)
yapmak する
-a-maz (不可能)
-lar (3人称複数)

LEMMA: yap
yapmakの語幹

FEATS
接辞の役割は素性列で表現

6 geliştirilmesi geliş VERB Verb Aspect=Perf|Case=Nom|Mood=Ind|Number[psor]=Sing|Person[psor]=3|Polarity=Pos|T

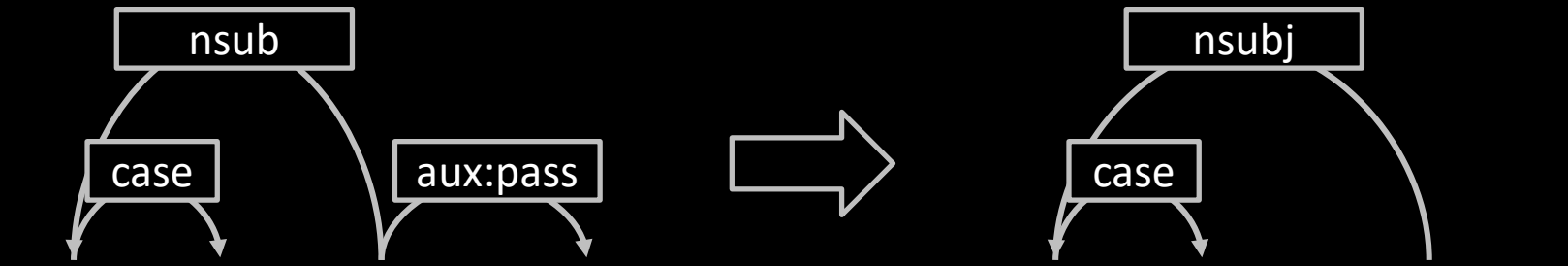
FORM: geliştirilmesi
((彼の) 開発させられる)
gelişmek 開発する
-tir (使役)
-il (受身)
-me (不定形)
-si (所有3人称単数)

LEMMA: geliş
gelişmekの語幹
(geliştirilではない?)

FEATS
接辞の役割は素性列で表現

※トルコ語知識は付け焼き刃なので間違っていたらすみません
雰囲気は伝わると思います

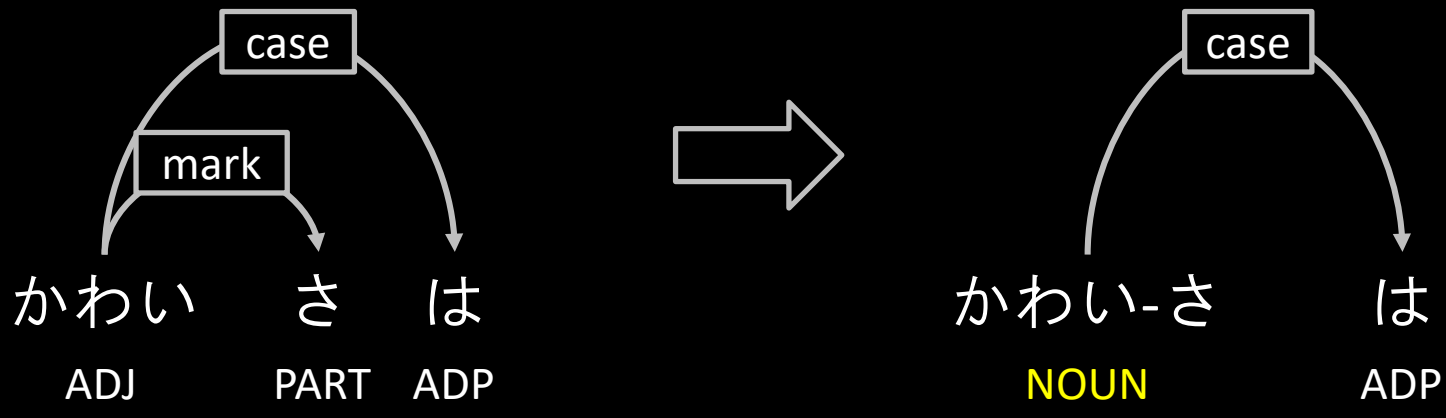
係り受け木の変更 1/2



Mood=Ind
|Tense=Pres
|Aspect=Imp
|Voice=Pass
|Polality=Pos
|Aspect=Imp

※雰囲気だけ

係り受け木の変更 2/2



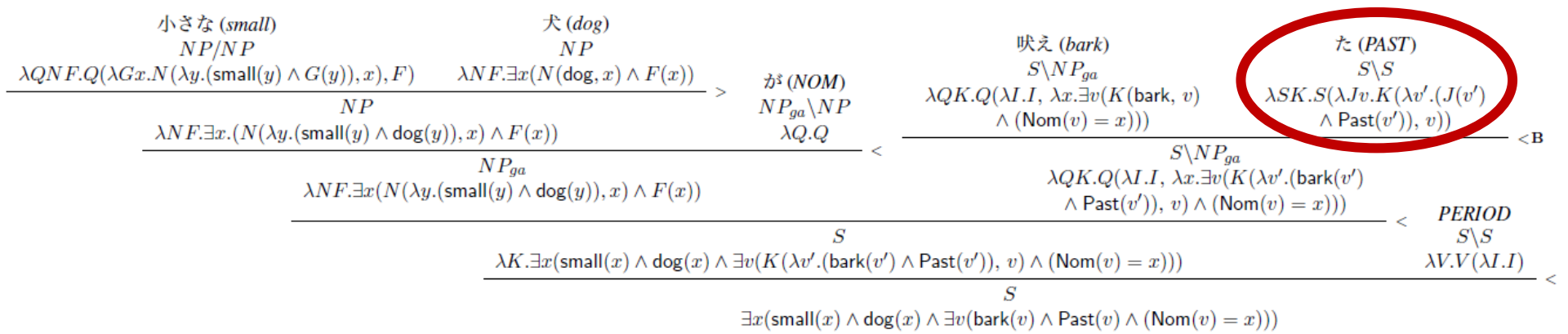
※雰囲気だけ

統語的語の採用のPros/Cons

- Pros
  - アノテーションの恣意性を軽減し、多言語比較の信頼性が向上
- Cons
  - 基準作成が大変
    - 宮岡伯人『「語」とはなにか・再考』(2015) が接語・接辞の詳細な認定例を示すが、もちろんアノテーションとしては不充分
  - 短単位への分割 → 統語的語へのチャンキングという処理は日本語独自となり、単一モデルによる多言語処理というUDの利点が失われる
    - かといって、いきなり統語的語へ分割すると解析精度が下がりそう
  - そもそも応用処理は統語的語を必要としないのでは?

形態素単位の (接辞を分離した) 方が応用上好都合

[Mineshima+, EMNLP2016]



- 応用では、ある機能がどのような形態統語的手段で実現されていたかは重要でないことが多い
- 日本語では接辞でも形/機能の透過性が高い
  - cf. ドイツ語 gute (Mutter)
- 接語と接辞にまったく別の表現 (独立の語 vs 素性) を与えるのが良くないのでは?

まとめ

- 日本語における統語的語の認定方針 (=接語・接辞の識別) を提案
- しかし、この作業の費用・便益比は微妙
- arXiv論文の改稿を計画していますので、フィードバックをお願いします