

語の認定と それに付随する提案

京都大学
村脇 有吾



自己紹介

- 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-)
- 出身背景は計算機科学
- 主な研究課題
 - － 表: 特に知識を用いた言語解析 (最近は応用も)
 - 博士論文 (2011): Automatic Acquisition of Japanese Unknown **Morphemes**
 - － 裏: 言語の系統、変化の統計的分析 (特に**言語類型論**の特徴を使った) (2014-)

UDへの期待

- 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
- **一貫した基準**によるタグ付けで多言語比較を可能に
 - これまで言語学者が作った類型論の特徴 (e.g. 基本語順) を分析に使ってきた
 - 統語的特性を分析するにはタグ付けされたテキストがほしい

今日のお話

- 現在のUD日本語はUDの語の認定方針に従っていない
- 採るべき道は二つに一つ
 1. UD日本語をUDの方針にあわせる
 2. UDの方針を変更させる
 - cf [Croft+ \(2017\)](#)
- 1を採る場合、宮岡 (2015) がたたき台として使えそう

提案の概要

- 短単位を語として扱うのはやめる
- 1個以上の短単位のまとまりを語とする
 - 短単位 $\leq$ 語 $\leq$ 文節
 - 大雑把には:
 - 名詞に後続する助詞等はわけたまま
 - 動詞に後続する屈折辞はくっつける
 - 派生接辞はくっつける
- 語の一部となった形態素の形態統語的役割はuniversal featuresで表現する

UD トルコ語の例

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

geliştirilmesi gelişt VERB Verb Aspect=Perf|Case=Nom|Mood=Ind|Number[psor]=Sing|Person[psor]=3|Polarity=Pos|Tense=Aor 0 root

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: gelişt
gelişmekの語幹
(geliştirilにすべき?)

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
雰囲気は伝わるとと思います

UDにおける語

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.

出典: <http://universaldependencies.org/u/overview/tokenization.html>

3種類の語

- orthographic word: 分かち書きの単位
 - UD文書が検討しているのはorthographic wordとsyntactic wordの関係だけ
 - 日本語には参考にならない
- phonological word: 音韻的まとまり
 - 現代日本語では多くの助詞は固有のアクセントを持たず、先行する名詞に支配されるので、名詞 + 助詞 はphonological wordを構成
- 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)

接語と接辞の研究史

- Humboldt (1836): encliticsへの言及あり
- 20世紀の (構造主義的) 研究
 - Zwicky and Pulum (1983), Matthews (1974) 等
 - 服部 (1950) 「附属語と附属形式」
 - 附属語 = 接語, 附属形式 = 接辞
- [Halpelmath \(2011\)](#): 類型論的興味からの通言語的な再検討
- 宮岡 (2015): 日本語に対する具体的な認定

語の認定基準の例

[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」(注: =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

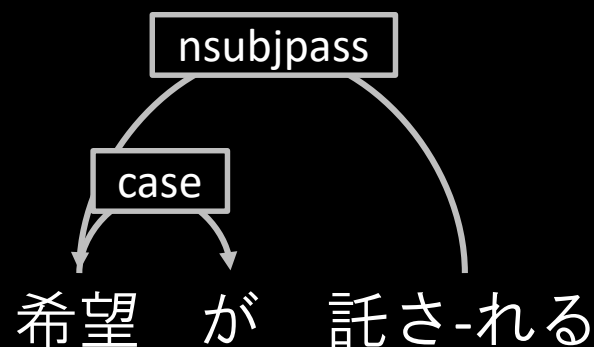
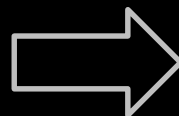
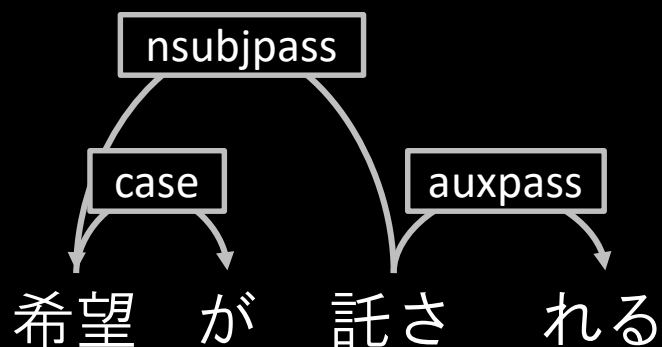
語の認定の具体例 1/2

- =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

語の認定の具体例 2/2

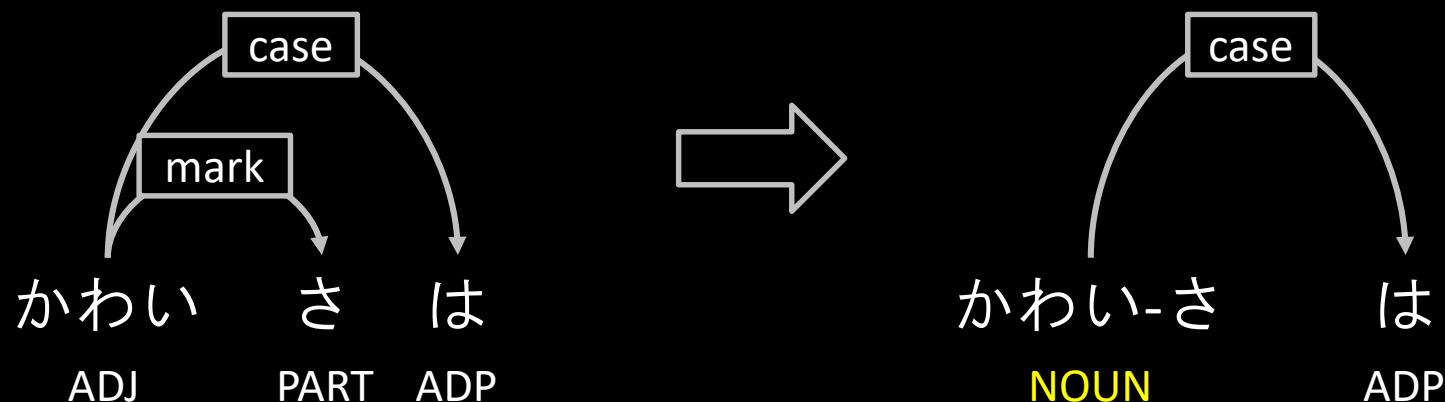
- -ます: 派生接尾辞 (VV型)
 - 食べ-ます, 食べ-まし-た
- -さ: 派生接尾辞 (VN型)
 - 美し-さ, 食べ-にく-さ
- -らしい: 派生接尾辞 (NV型)
 - 男-らしい, 男-らし-そう=だっ-た
 - cf (女ではなく) 男=らしい, 男=だけ=らしい
- -はじめる: 派生接尾辞
 - 泳ぎ-はじめ-た, 泳ぎ-づかれ-はじめ-た

係り受け木の例 1/2



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

係り受け木の例 2/2



- 接辞のチャンキングに付随する処理
 - FORM, LEMMAの修正
 - POSの修正 (派生辞の場合)
 - FEATSの生成 (用言の場合)
- 長単位認定で同様の手法を開発済み?

宮岡 伯人.2015.「語」とはなにか・再考: 日本語文法と「文字の陥穽」

- 宮岡はユピック語を専門とする言語学者
 - ユピック語は複統合的
qaya -pi -li -vkar -ni -anga
kayak -real -make -A'.CAU -A''.say -IND.3sg.1sg
he [A''] says that (someone [A']) let me make a genuine kayak
- 『「語」とはなにか: エスキモー語から日本語をみる』(2002) でも日本語の語の認定基準について議論
- 2015年版は日本語研究史を整理するとともに、具体的な認定結果を詳細に議論
- UD日本語のたたき台として使えるのでは

【表 2】接辞と接語の対照表 (i-xiv の分類については、4.2.3 の (13) を参照)

接 尾 辞 i-v		前 接 語 vi-viii
i., v. 用言性接尾辞		vi. 用言性前接語
i. 派生接尾辞(-VV-) ふたつ以上の連続可能 (相互承接)	v. 屈折接尾辞(-VV) 範列からの選択 (ふたつ以上の連続なし)	=らしい- [複合前接語] =らしい=です- =の=だ-(=です- =である-) =の=は≠ない-(≠あり-ませ-ん) =よう・そう・みたい=だ-(=です- =である-) [不変化型] =[だ]ろう(=でしょう) =[な]ら[ば] (完了「-た -だ」の後での[だ・な]の脱落)
未然[a] - (さ)せる- -しめる- - (ら)れる- - ない- -ぬ[ん]- 未然[o] 連用[i] -音便 -ます- -たい- -がる- -たがる- [2次の接尾辞] -だす- -かける- -はじめる- -おわる おえる- -あげる- -はたす- -きる- -しまう- -つづける- -なおす- -なれる- -すぎる- -あう- -ちがう- -やすい- -にくい- -がたい- -づらい- -ぐるしい- [複合接尾辞] -そう=だ- (vs. vi. =そう=だ-) 終止連体[(r)u] -べし- 仮定[e] 命令[e]	[終止法] -(よ)う[意志・推量] -た -だ[過去] -Ø[連用形・中断] -て -で[中断] -たり -だり[並列] -に[目的] -つつ [同時] -Ø[非過去] -な[禁止] -まい[否定意志] (ruの任意脱落) -と -に[前件] -ば[条件] -ど(も)[反対] -Ø -よ* -ろ[命令]	
		vii. 用言化前接語 =だ- (=です- =である-) viii. 体言性・体言化前接語 =が =の =に =を =へ =と =で =の[関係詞節化] =と[名詞節化] =し[同時] =と(も) [反対] =も =は =か =ぞ =ね =わ =よ* =さ =さえ =すら =ずつ =だけ =きり =こそ =でも =まで =から =より =のに =かな =ごと =ほど =など =なり =とか =のみ =しか =やら =ずつ =ばかり =だらけ =だてら =ぐらい =かしら =ながら =がてら =ついで(≠ついで) (～-ながら -がてら -ついで)

ii. 用言化接尾辞(-NV-)

-めく- -めかす- -ばむ- -ばる- -ぶる- -だつ-
 -づく- -づける- -がかる- -じみる-
 -らしい-(vs. vi. =らしい) -たらしい- -くさい- -つばい-
 -がましい-

iii. 体言性接尾辞(-NN)

-ども -さん -がた
 -ずき -ぎらい -ばなれ [転成名詞的]

iv. 体言化接尾辞(-VN)

-さ -がち
 -すぎ -ぎらい -ばなれ -づかれ [転成名詞的]

準 前 接 語 ix-x

ix. 用言性・用言化準前接語

≠する- ≠なさる- ≠ある- [再立ち上げ]
 ≠あらず-
 ≠なる- ≠ない-
 ≠なら-ない- (-ね・-なければ とともに複合接尾辞)
 ≠できる- (≠でき-ます-)

[不変化型]

≠けれど(も)

[動詞幹+「-て|-で」に前接：4.5.2.1]

≠いる・いらっしやる- ≠ある- ≠おる- ≠おく- ≠みる-
 ≠いく- ≠くる- ≠やる-
 ≠しまう- ≠すます- ≠みせる-
 ≠もらう- ≠いただく- ≠やがる- ≠くさる-
 ≠あげる- ≠あがる- ≠くれる- ≠いただく-
 ≠ほしい-

[後接語「お|ご=」と連動]

≠なさる- ≠いたす- ≠くださる- ≠いただく-
 =に≠なる-

x. 体言性・体言化準前接語

≠こと[名詞節化] ≠はず ≠かた ≠もの
 ≠ところ ≠つもり ≠だから ≠しだい ≠ついで

複 合 前 接 語 xi-xii

xi. 用言性・用言化複合前接語[たぶん非閉鎖的なリスト]

= (な=) の=だ・です・である-[用言化前接語=だ]

= (の=) で=は≠ない-

= だ・です=が・から・=ので

= と≠する-・≠なる-・≠みる-

= と≠いう- = と=もうし≠あげる- = と≠いう≠こと=だ

=に#する-・#なる-・#ちがい#ない-
 =(の)かも#しれ#ない- =(の)=よう=だ|です-
 =(の)=ほう=が(=いい-)
 =[な]ら(ば)≠いい- (完了「-た|-だ」の後での「な」の脱落)
 =そう=だ・です- =みたい=だ・です- ≠ほど=だ・です-
 ≠はず=だ・です- ≠わけ=だ・です- ≠ばかり=だ・です-
 =かも#しれ#ない-
 ≠はず・わけ=が≠ない-
 ≠こと=だ(=から) ≠こと=で≠ある(=から)
 ≠こと=が≠ある・できる- =こと=は≠ない-
 ≠こと・ほど=に≠なる・する-

≠ところ=が・で ≠ところ=だ・です-
 =どころ=か どころ=で=ない-
 ≠つもり=が ≠つもり=だ・です-

xii. 体言性・体言化複合前接語

=で=も =で≠もつ-て =(の)=に・で・も
 =に (=も) ≠かかわら-ず =に≠たい≠し-て
 =(の)=とき・ころ・ため・なか・ほう・ところ・あいだ(=
 =から=に=は
 =たる≠もの
 ≠もの=の

接頭辞 xiii

用言性・体言性接頭辞 (VV- NN-)

さ- ま- か- こ- み- さ- す-

[2次的接頭辞]

うち(うっ)- おち(ぶっ)- ひき(ひっ)-

なま- もの- かた-

後接語 xiv

お=
 な=

この= あの= …
 どんな= おおきな= …
 きたる= いわゆる= …

Greg Pringle. Thoughts on the Universal Dependencies proposal for Japanese

- 本人の卒論 (1977) をもとにした (らしい) 日本語 UD批判
- UD主要開発者がUD日本語の問題に気づいたきっかけ
- Pringle本人は文節推し
- Pringleの代替案は服部・宮岡路線
 - particles following nouns as clitics
 - 'agglutinative morphemes' following verbs as suffixes or word endings
 - the copula as a separate word
- 品詞割り当ての問題も論じるが、大半は語の認定に関連

UD Issue #377: Definition of word

- Nivre: Following up on the discussion about UD Japanese, it seems clear that the segmentation of derivational morphology is not consistent with general UD principles, nor with practice in other UD treebanks, so I definitely think that should be changed. Similarly, when it comes to verbal morphology, all the descriptions so far point to this being **a regular and systematic case of agglutination**, similar to the situation in Turkish and Finnish, so it seems that an analysis using morphological features is most compelling there as well.
- Nivre: The main point is that "changing a noun into an adjective" is not a syntactic relation, so there is no appropriate UD relation to put on an arc between these two elements. This is a clear indication that this **should be treated as word formation**, not syntax, and it is exactly parallel to the argument against the old "DERIV" links in the Turkish treebank.

日本語とトルコ語の違い

1 Kahveleri kahve NOUN Noun Case=Acc|Number=Plur|Person=3 3 obj _ _

FORM: Kahveleri
(コーヒーたちを)
kahve コーヒー
-ler (複数)
-i (対格)

LEMMA: kahve
kahveの語幹

FEATS
名詞にも素性列

- 日本語は文法的にアルタイ的としばしばいわれるが、名詞に付属する要素の自立性の高さは非アルタイ的では？

チベット語

ཁ་ས་མི་ཕྱག་པོ་འདིའི་ཁང་པ་གོང་ཆེན་པོ་ཞིག་གཞིགས་སོང་།

ཁ་ས་ མི་ ཕྱག་པོ་ འདི འི་ ཁང་པ་ གོང་ཆེན་པོ་ ཞིག་ གཞིགས་ སོང་ །

昨日 男 金持ちの この [能格] 家 高い 1 買う[過去] [過去] 。

昨日この金持ちの男は高い家を買った。

- 日本語と同様、比較的膠着的だが分かち書きしない言語
 - 点 (tseg) は音節境界
 - 形態素への分割: 基本的には音節をチャンキング [Liu+ 2011]
 - 一部例外的に音節内を分割: འདིའི་ → འདི་ + འི་
- アクセントの関係からphonological wordを認定する例はあるが、syntactic wordの認定する研究は未見
 - そもそも依存構造コーパス自体がまだない...
 - 現代語の動詞は付属要素を引き連れる傾向にあるので、日本語と同じく形態素のチャンキングが必要か気になる