Universal Dependenciesと Syntactic Wordの闇

京都大学
村脇 有吾

第9回入力メソッドワークショップ LT@京都大学 (2017年12月29日)

にゃかがわりょうじ
@nyakagawa_r

Follow

「ネットで見つけた格好いい船。なんて名前なんだろう？」

×「この船の名前を教えて下さい」
→「ググレカス」

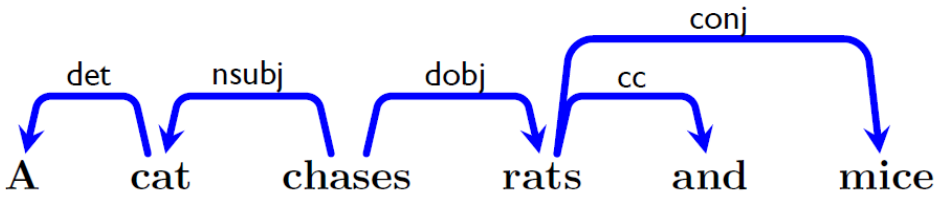
○「この"戦艦"超かっこいい！」
→「FF外から失礼します。戦艦ではなく駆逐艦です」「ズムウォルト級です」「アーレイ・バーク級に代わる最新鋭次世代駆逐艦として建造が予定されていま

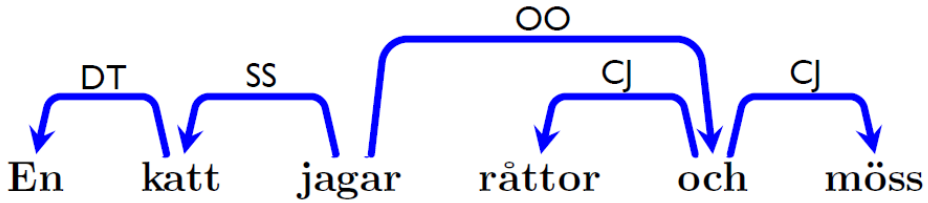
Translate from Japanese

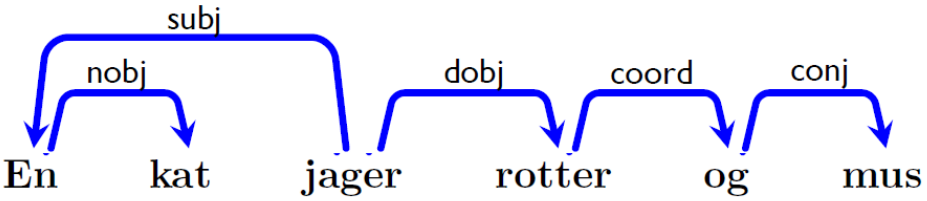
https://twitter.com/nyakagawa_r/status/945933804182560769

Universal Dependencies (UD)

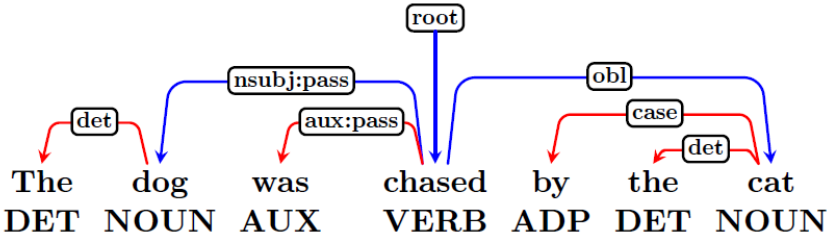
UD以前

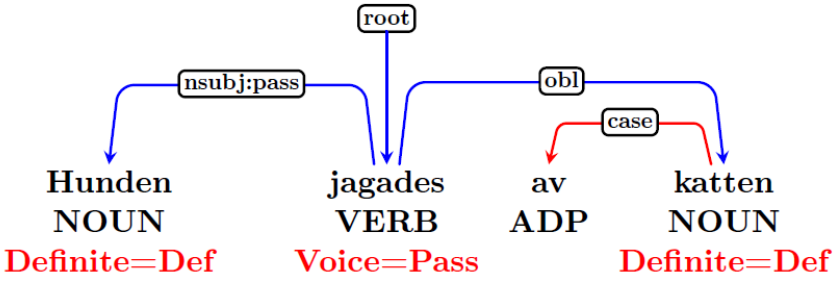


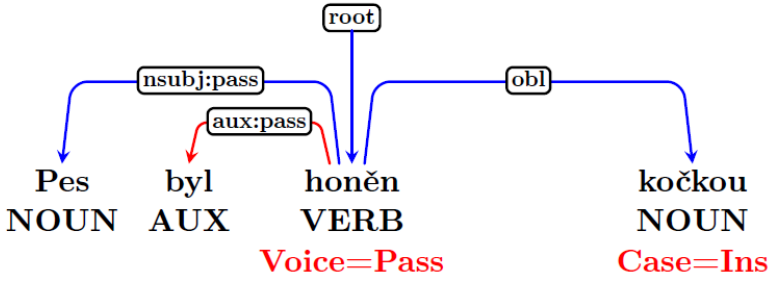




UD







言語学でもいまだ作られたことのない言語アノテーションを工学系の我々が設計しようとしているし、その難しさをコア開発者たちは当初認識していなかったように見える

UDが開いたパンドラの箱

- 伝統文法は言語ごとにばらばら
- 言語類型論の沼
  - 形容詞という範疇は通言語的に存在するか?
  - 比較のための一部の言語現象を取り出すのと、各言語に一貫した体系を与えるのは別
  - どうしても広く浅い比較になりがち
- 語彙、形態論と統語論の切り分けの妥当性?
- 応用にとって本当に便利な表現?

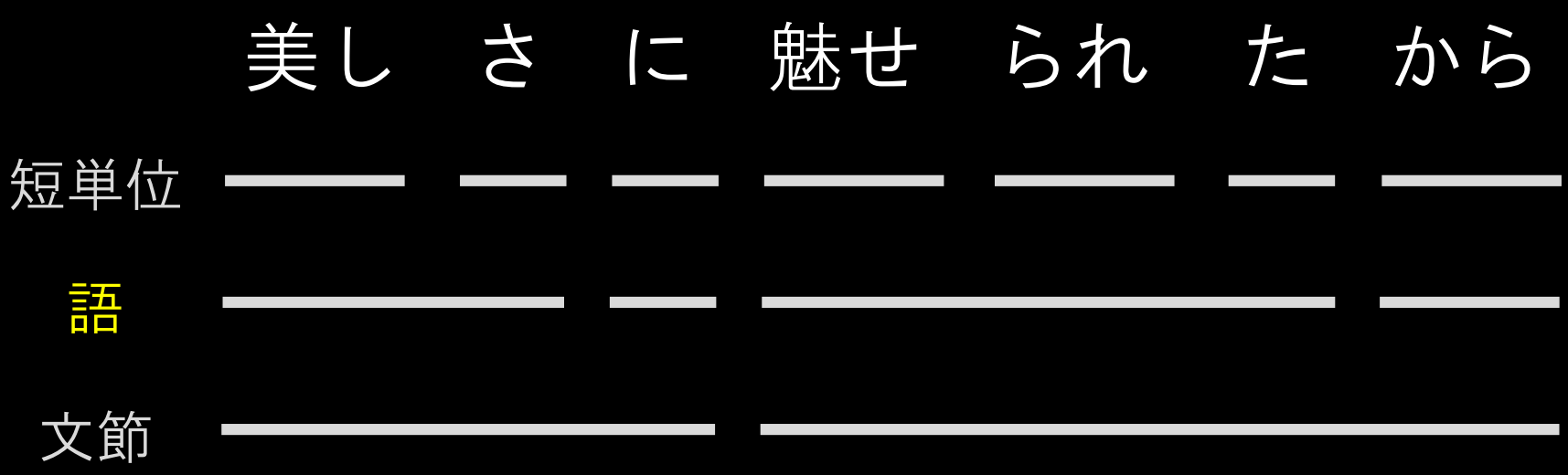
UDにおける語: Syntactic Word

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

Syntactic wordを採用すると日本語ではどうなるか?

- 既存の文節、BCCWJ短単位、中単位、長単位のいずれでもない、新しい単位を導入せざるを得ない [村脇, UD Japanese mtg. 2017-06]
- おおよそ: 短単位 ⊆ 語 ⊆ 文節



そもそもsyntactic wordを採用すべきか?

- 構造主義言語学以来の古い単位 [服部, 1950]
- 類型論の立場から見直すと、通言語的に一貫した基準を見いだせない [Haspelmath, 2011]
- 意味的スコープとの不整合
  - [大阪に寄って、神戸に帰っ] た
- 複合語の扱いの問題

素人談義には限界があるので、みなさんの協力をお待ちしております