UD Japanese-KTC: 京大コーパス句構造版からの Universal Dependencies化

2018-06-16 第1回 UD研究会

NTT コミュニケーション科学基礎研究所
田中貴秋

発表の概要

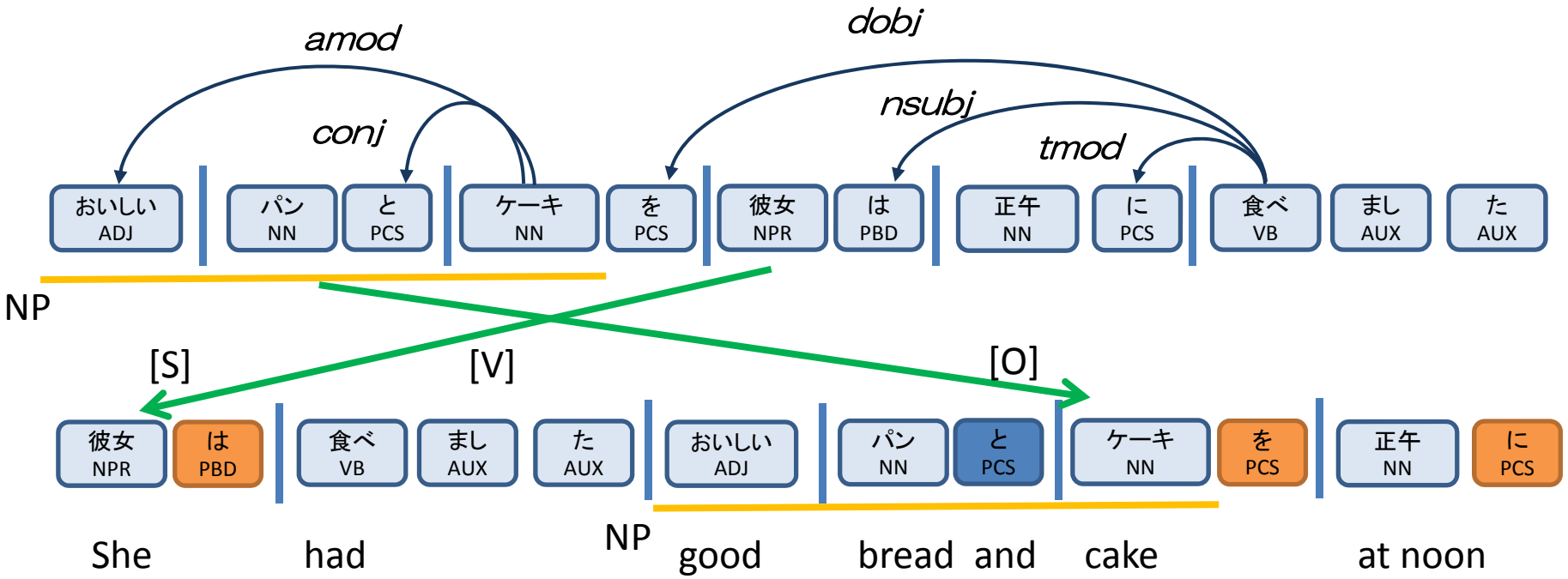
UD以前に構築していた句構造コーパス・単語依存構造コーパスと
UDとの関係についてお話します

- 日本語の文法機能ラベル付き構文木
  - 日本語句構造ツリーバンク「楓」
    - 京都大学テキストコーパスの1万文
    - 2分木
    - 文法機能ラベル
  - 句構造から単語依存構造へ
    - UDとその他の単語依存構造
- Universal Dependencies への変換
  - 変換に必要な情報
    - 格関係，節の機能，並列構造，複単語表現
  - 変換方法の実際
    - 主辞規則（左右の子，親）
    - 句ラベル => 依存構造ラベル

日本語の文法機能ラベル付き構文木

構築の背景

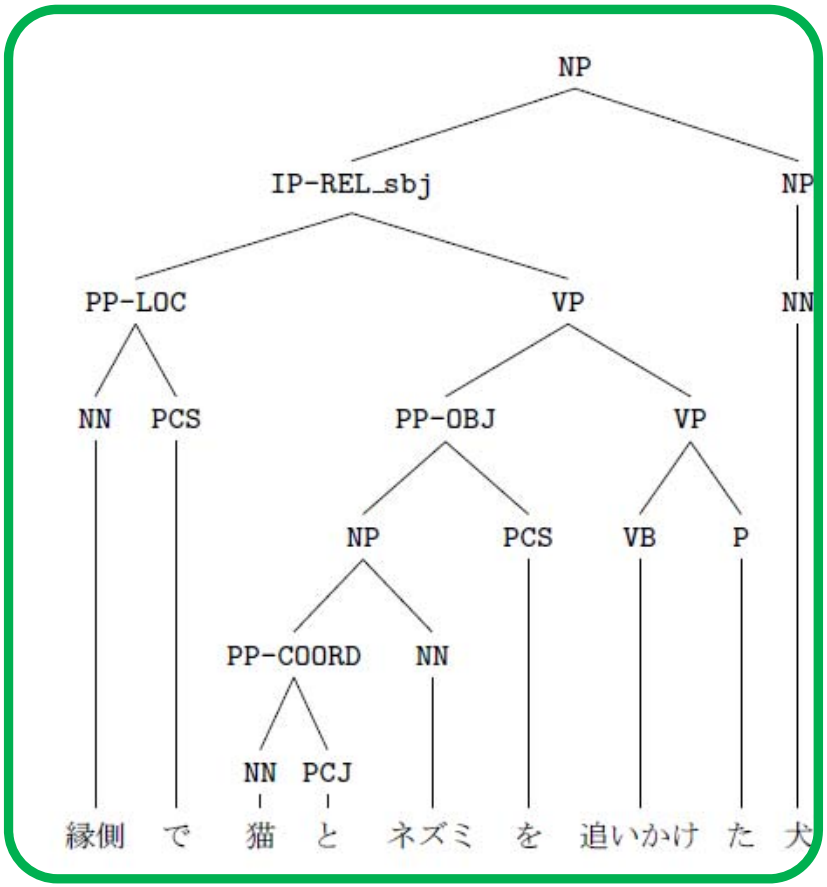
- 係り受けに機能ラベルが欲しい
  - 例えば，日英のSMTの事前並び替え（SOV->SVO）
    - どれがSで，どれがOかわからない
    - 文節が移動単位と合わない
  - 機能ラベルが文節間の依存関係に付与しにくい：名詞句，並列構造



句構造情報と述語項構造情報

- 句構造ツリーバンクを構築

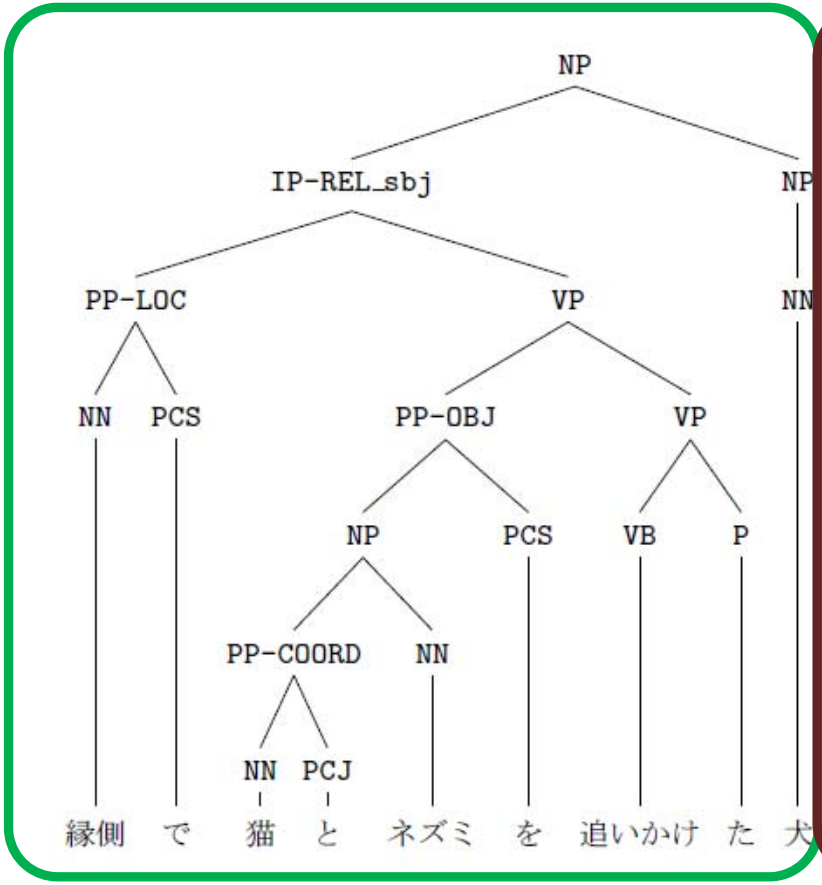
句構造



句構造情報と述語項構造情報

- 述語項構造情報を追加

句構造



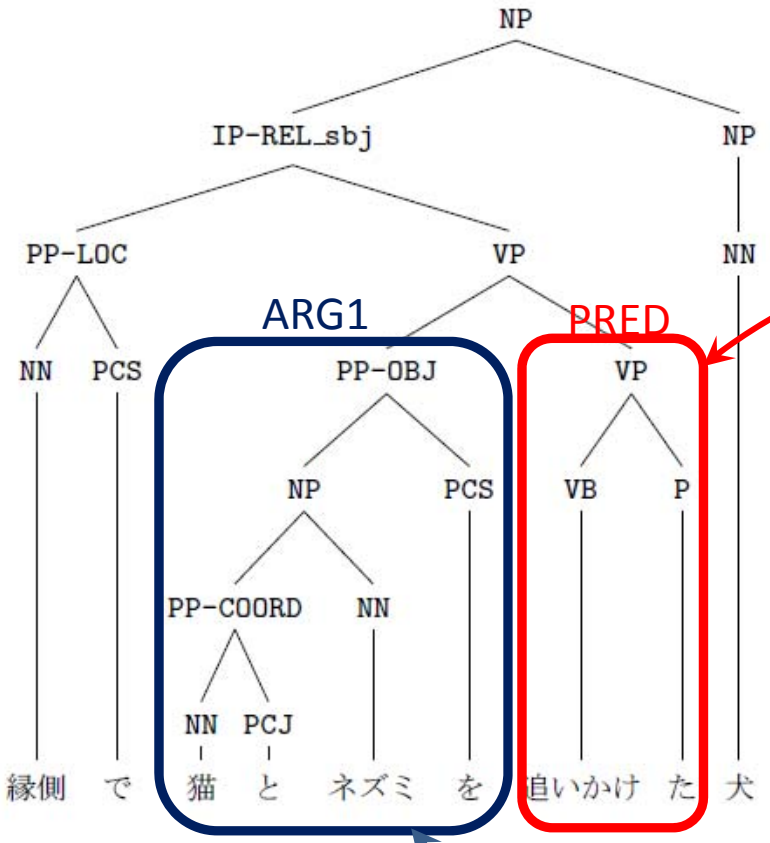
述語項構造

Voice:	ACT
Case frame ID:	023553 (追いかける)

	Phrase	Phrase ID	marker
PRED	追いかけた		-
ARG0	犬		adnom
ARG1	猫とネズミ/を		を
ARG1_P	猫		-
ARG1_P	ネズミ		-
LOC	縁側/で		で

句構造情報と述語項構造情報

- 句構造と対応する述語項構造



Voice:	ACT
Case frame ID:	023553 (追いかける)

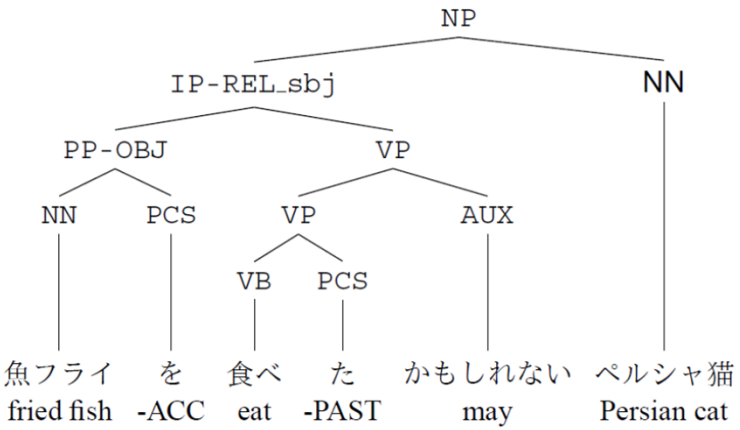
	Phrase	Phrase ID	marker
PRED	追いかけた		-
ARG0	犬		adnom
ARG1	猫とネズミ/を		を
ARG1_P	猫		-
ARG1_P	ネズミ		-
LOC	縁側/で		で

句構造ツリーバンク「楓」

京大コーパス[Kurohashi+2013]から構築した句構造ツリーバンク[田中+2014]

※ 元々は CCG への変換を目的として NII 宮尾さん，植松さんらと構築

- 2分木
- 文法機能タグ
  - 項の情報（-SBJ, -OBJ, -OB2）
  - 節の情報：関係節（IP-REL），補足節（CP-THT）など
  - 並列構造の情報: 並列(-COORD), 同格(-APPOS)
- 単語：国語研長単位
- 1万文



句構造情報と述語項構造情報の統合

句構造に対応する述語項情報を付与
同じ句IDを共有することにより，句を対応付け

		句構造情報	述語項構造情報
構文情報	木構造	○	―
	句カテゴリ	○（句ラベル）	―
	節カテゴリ	○（節ラベル）	―
格情報	必須格	○（表層格出現形）	○（表層格基本形，ゼロ代名詞含む）
	随意格	△（時間格，場所格）	○
	態	―	○
	格フレーム	―	○

句構造のアノテーション

2分木

文法機能

格情報

- 必須格
  - 文法役割(GR)ラベルセット：主格（-SBJ），対格（-OBJ），与格（-OB2）
- 随意格
  - 時間格（-TMP），場所格（-LOC）

節

- 主節 (-MAT)
- 従属節
  - Adverbial clause (-ADV), Adnominal clause,…

連体修飾節のタイプ

- Gapping relative clause (-REL_sbj, -REL_obj, -REL_ob2,)
- Non-gapping rel. clause (-ADN)

並列構造

- 並列句 (-COORD)
- 同格句 (-APPOS)

句構造のアノテーション

2分木
文法機能
　格情報
- 必須格
  - 文法役割(GR)ラベルセット：主格（-SBJ），対格（-OBJ），与格（-OB2）
- 随意格
  - 時間格（-TMP），場所格（-LOC）

　節

Active

犬-が　猫-を　追った　　((PP-SBJ 犬-が)((PP-OBJ 猫-を)(VP 追っ た)))
dog-NOM cat-ACC chased
The dog chased the cat

Passive

犬-に　猫-が　追われた　　((PP-OB2 犬-に)((PP-SBJ 猫-が)(VP 追わ れた)))
dog-DAT cat-NOM chased-PASS
The cat was chased by the dog

句構造のアノテーション

従属節

副詞節 (IP-ADV)

連体修飾節

- 関係節 (IP-REL)：-REL_sbj, -REL_obj, -REL_ob2,

- 内容節，補充節 (IP-ADN)
  - 「さんまを焼くにおい」
  - 「家に帰る途中」

補文 (CP-NNF)

- 「会場に着くのは明日だ」

引用節 (CP-THT)

- 「早く帰りたいと思った」

疑問節 (CP-QUE)

- 「いつ来るかわからない」

句構造のアノテーション

従属節

副詞節 (IP-ADV)

連体修飾節

副詞節

猫-が エサ-を 取った-ので，犬-は　追いかけ た

(IP-MAT (IP-ADV ((PP-SBJ 猫-が) ((PP-OBJ エサ-を) (VP 取っ た)) (PP ので))

(IP-MAT (PP-SBJ 犬-が) (VP 追いかけ た)))

– 「家に帰る途中」

補文 (CP-NNF)

- 「会場に着くのは明日だ」

引用節 (CP-THT)

- 「早く帰りたいと思った」

疑問節 (CP-QUE)

- 「いつ来るかわからない」

句構造のアノテーション

従属節

副詞節 (IP-ADV)

連体修飾節

- 関係節 (IP-REL)：-REL_sbj, -REL_obj, -REL_ob2,

- 内容節，補充節 (IP-ADN)
  - 「さんまを焼くにおい」
  -

補文 (CP-N

- 「会場

引用節 (CP-

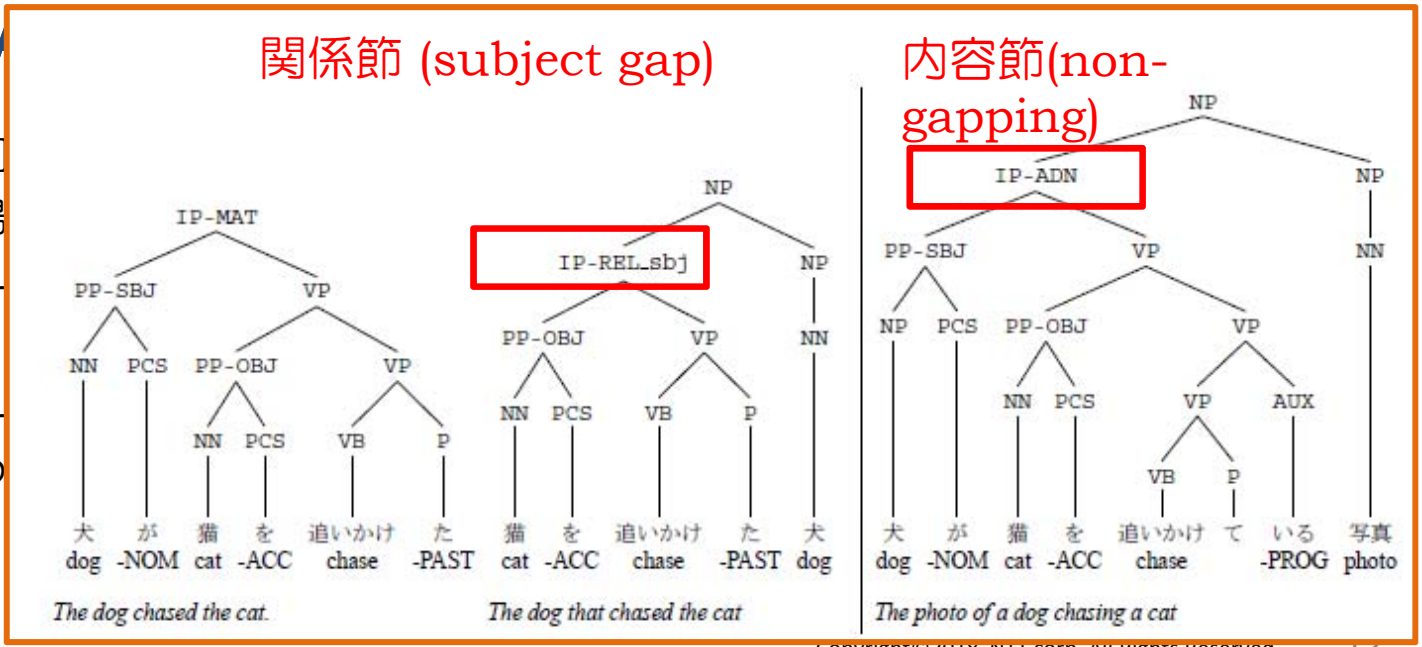
- 「早く

疑問節 (CP-

- 「いつ

関係節 (subject gap)

内容節(non-gapping)



句構造のアノテーション

従属節

副詞節 (IP-ADV)

補文

猫-が エサ-を 取った-の　は，意外だった

(IP-MAT　(CP-NNF ((PP-SBJ 猫-が) ((PP-OBJ エサ-を) (VP 取っ た)) (PPの))　(PP は))

(((NADJ 意外) (PP だっ)) た))

引用節

犬-は，猫-が エサ-を 取った と，知った

(IP-MAT (PP-SBJ 犬-は)

(PP-OBJ (CP-THT (PP-SBJ 猫-が) ((PP-OBJ エサ-を) (VP 取っ た))) (PPと)) (VP知っ た)))

補文 (CP-NNF)

- 「会場に着くのは明日だ」

引用節 (CP-THT)

- 「早く帰りたいと思った」

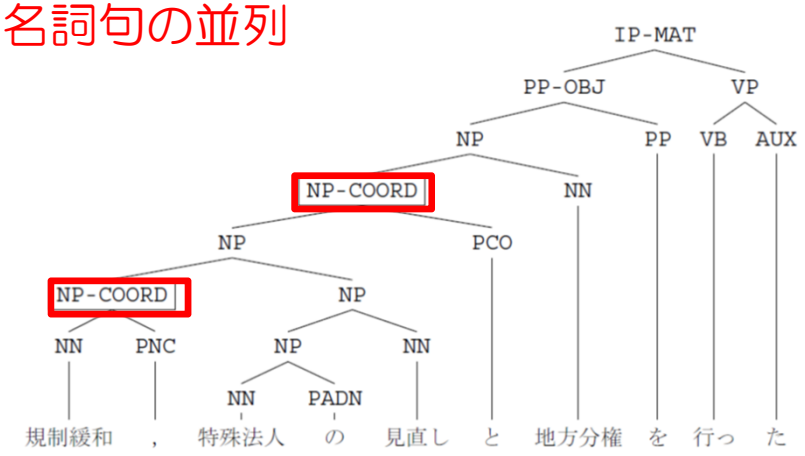
疑問節 (CP-QUE)

- 「いつ来るかわからない」

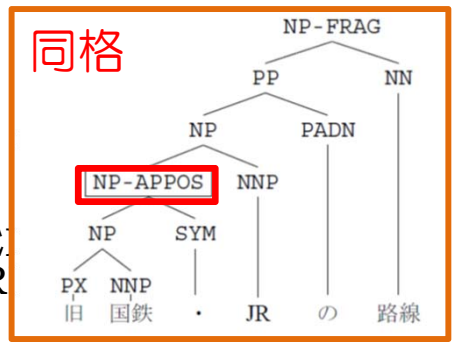
句構造のアノテーション

2分木
文法機能
　格情報
- 必須格
  - 文法役割(GR)ラベルセット：主格（-SBJ），対格（-OBJ），与格（-OB2）
- 随意格

名詞句の並列



同格



- 並列句 (-COORD)
- 同格句 (-APPOS)

句構造から単語依存構造へ

- 単語単位の係り受け
  - Stanford Dependencies (SD) [de Marneffe+ 2006]
    - 英語の単語間
    - 4-50程度の依存関係ラベル

    => SD風の日本語単語依存構造[Tanaka+ 2015]

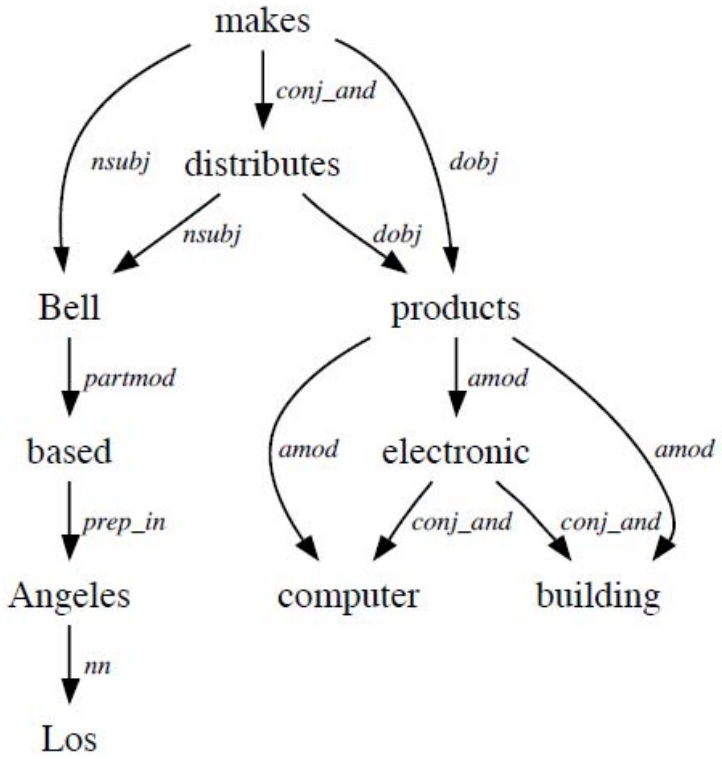
  - 日本語単語係り受け[Mori+ 2014, 2015]
    - 国語研 短単位間（語尾細分割）
    - 依存関係ラベルなし

  - Universal Dependencies (UD)
    - 多言語横断のためのSD拡張
    - 日本語版 [金山+ 2015]

SD風依存構造とUDの比較について簡単に述べる

Stanford typed dependencies (SD)

- 文法的な関係を簡潔に記述 [de Marneffe+ 2006]
  - *relation* (head, dependent)
- 単語間の文法的な関係を
  - 格関係(*nsubj*, *dobj*, *iobj*)
  - 名詞句内の関係(*nn*, *amod*, *num*, *prerp*, …)
  - 関係節 (*rcmod*)
  - 並列, 同格 (*conj*, *appos*)
  - …



- 多言語拡張
  - Universal Stanford Dependencies [de Marneffe+ 2013]
  - Universal Dependencies

依存構造の設計に必要な要素

- 依存関係を定義する単位
- 主辞の決定方法
- 依存関係ラベル

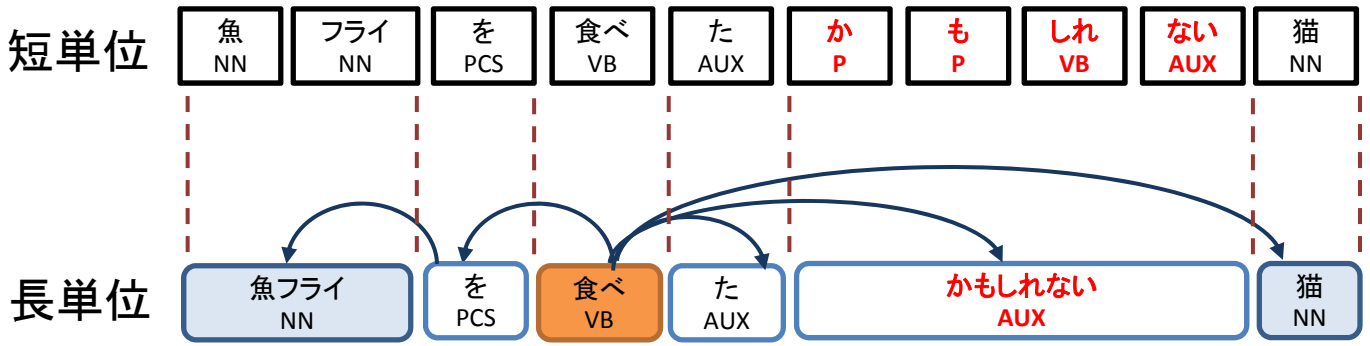
依存構造の設計に必要な要素

SD風単語依存構造 [Tanaka+ 2015]

- 依存関係を定義する単位
  - 国語研 長単位
- 主辞の決定方法
- 依存関係ラベル

依存関係を定義する単位

- BCCWJ [Maekawa+ 2014]の2階層の単位を採用
  - 短単位：情報の最小単位　<= 現**UD**はこちらを採用
    - 分割単位，品詞の揺れが少ない
    - 依存関係の単位としては，細かい
      e.g. 「だ->が」「に->つい->て」「か->も->しれ->ない」
  - 長単位：依存関係の単位
    - 粗い定義：文節内を内容語と機能語に分割した単位
    - 機能語間等の冗長な依存関係を無視する
    - ○機能語（複合辞）が1単語として扱える
      e.g. 「について(格助詞)」，「かもしれない(助動詞)」



依存構造の設計に必要な要素

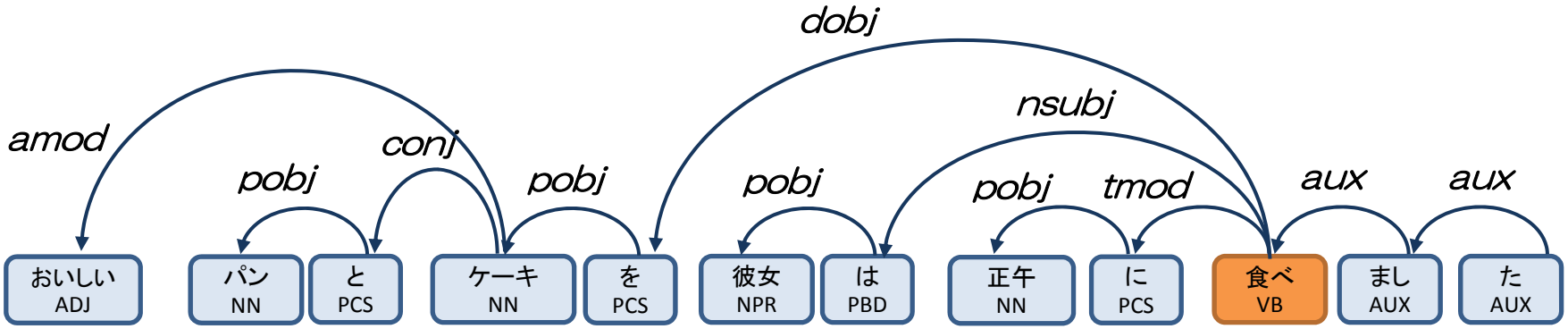
SD風単語依存構造 [Tanaka+ 2015]

- 依存関係を定義する単位
  - 国語研 長単位
- 主辞の決定方法
  - 名詞句 <- 格助詞類
  - 述語句内はいくつかのバリエーション
- 依存関係ラベル

主辞の決定方法

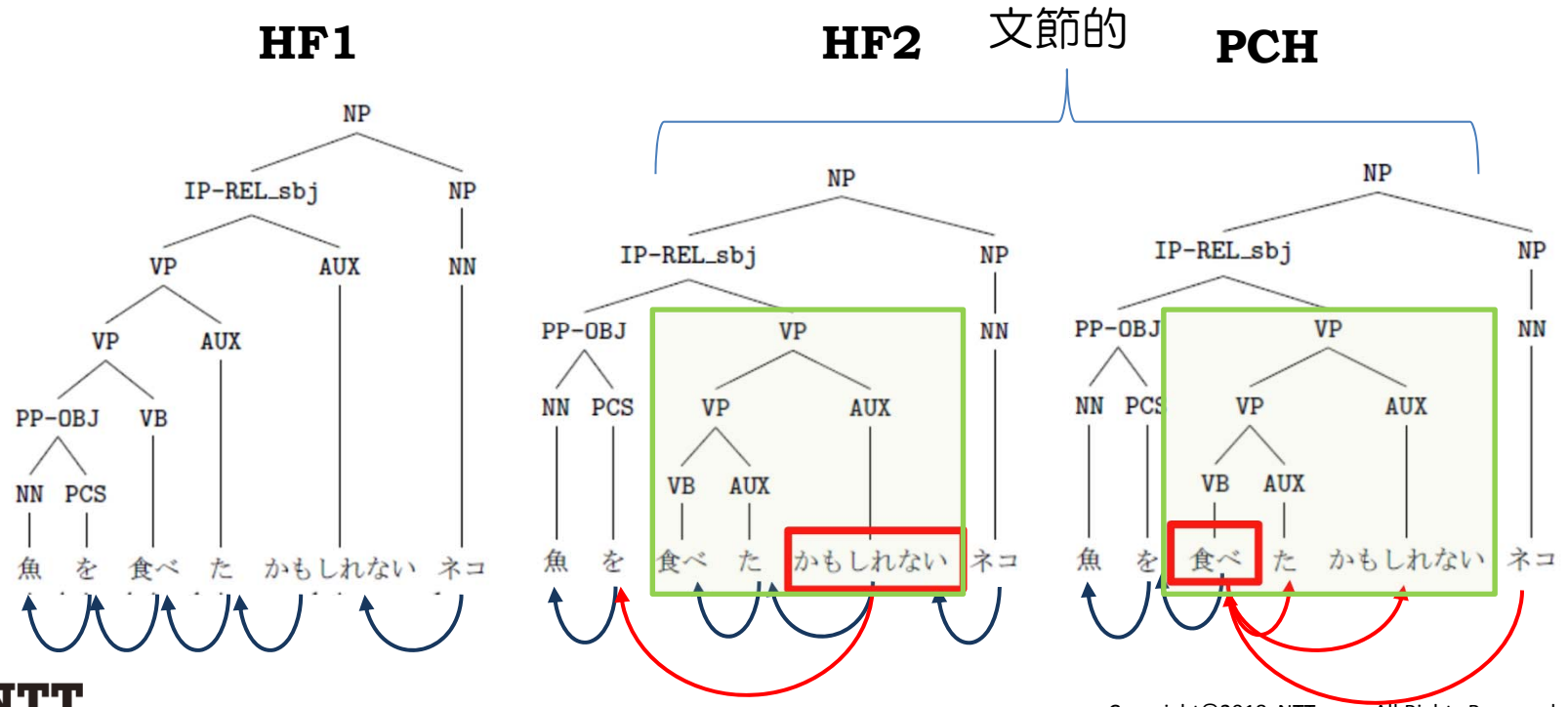
- 基本原則
  - 依存関係のある語のうち右側の語が主辞となる
    - 後置詞句（主に名詞+格助詞）は，助詞を主辞とする
  - ただし，副助詞，句読点，閉じ括弧類は，左側の要素を主辞とする

※述語句の構造については，3パタンを考える



述語句内の主辞決定

- 依存構造のタイプ：述語のまとめ方
  - **HF1**：主辞後置型１, Head Final type 1（右側主辞）
  - **HF2**: 主辞後置型２, Head Final type 2（述語句最右主辞）
  - **PCH**: 述語内容語主辞型, Predicate Content word Head type（述語句最左主辞）



依存関係ラベルの定義

- SDをベースに日本語化
  - 連体修飾に関して拡張（関係節における空所の区別）
  - 35のラベル(下は一部)

格関係	必須格	*nsubj*, *dobj*, *iobj*	述語項構造
	随意格	*tmod*(時間), *lmod*(場所), *arg*	
節	関係節	*rcmod_nsubj*, *rcmod_dobj*, *rcmod_iobj*	述語項構造
	補充節(外の関係)	*ncmod*	
	補足節	*ccomp*	
	副詞節	*advcl*	
内容語間の修飾		*amod*, *advmod*, *nmod*, *num*	
機能語関連		*aux*, *pobj*, *post*	
並列，同格		*conj*, *appos*	並列構造

依存構造の設計に必要な要素

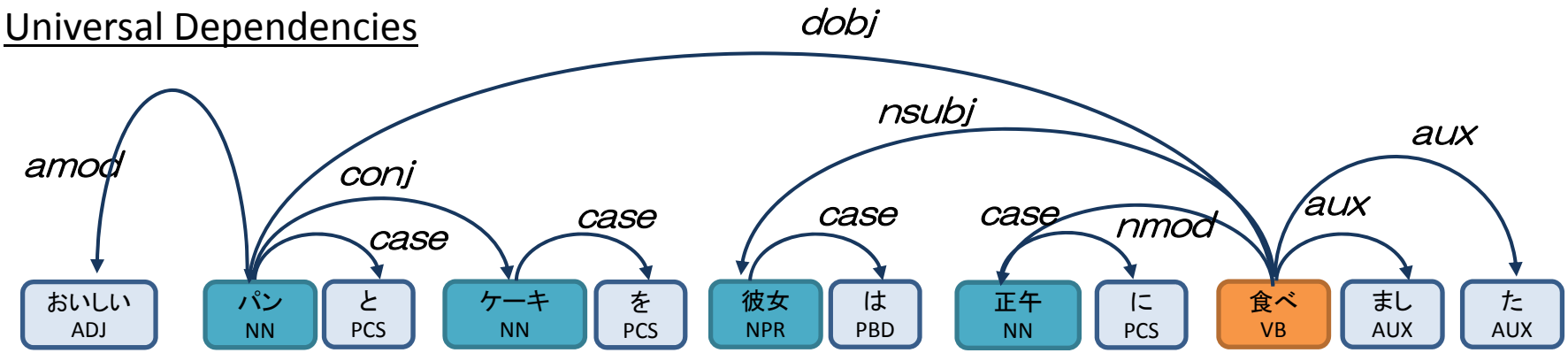
Universal Dependencies

- 依存関係を定義する単位
  - 国語研 短単位
  - 国語研 長単位
- 主辞の決定方法
  - 内容語 -> 付属語 , 付属語 <- 内容語
  - 内容語 <- 内容語 （主辞後置）
  - 名詞句 <- 格助詞類
  - 述語句内はいくつかのバリエーション
- 依存関係ラベル

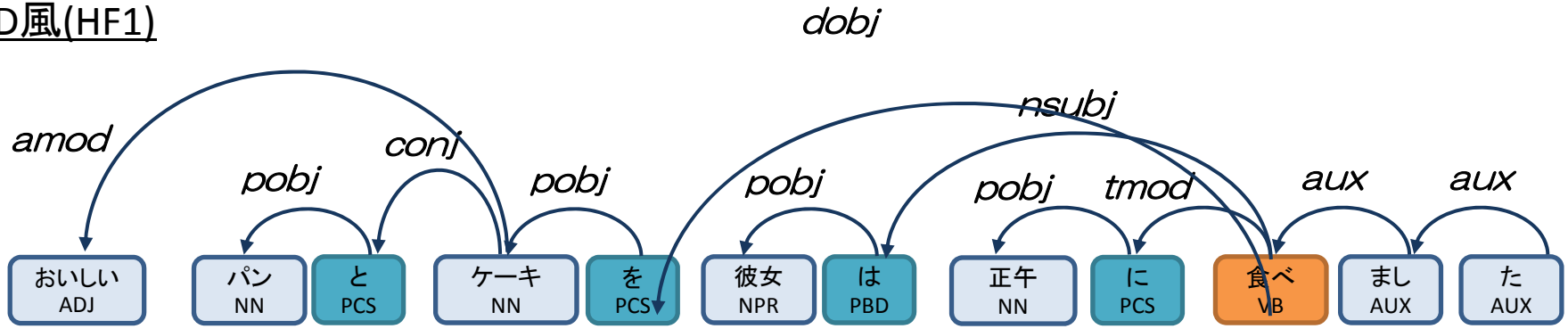
日本語UDと他の単語依存構造との比較

- UDは多言語横断が主目的（内容語間の依存関係中心）[金山+ 2015]

Universal Dependencies



SD風(HF1)



依存関係ラベルの定義

- UDはSDをベースに簡略化

格関係	必須格	*nsubj, dobj, iobj* => *nsubj, obj, iobj*
	随意格	*tmod*(時間), *lmod*(場所), *arg* => *obl*
節	関係節	*rcmod_nsubj, rcmod_dobj, rcmod_iobj* => *acl*
	補充節(外の関係)	*ncmod* => *acl*
	補足節	*ccomp* => *ccomp*
	副詞節	*advcl* => *advcl*
内容語間の修飾		*amod, advmod, nmod, num* => *acl, advmod*
機能語関連		*aux, pobj, post* => *aux, case*
並列，同格		*conj, appos* => *conj, appos*

句ラベルと依存構造ラベル

- UDv1 -> UDv2 と簡略化の方向

	句構造	UDv1	UDv2
格	出現形の表層格 (-SBJ, -OBJ, -OB2)	出現形の表層格 (nsubj,dobj,iobj)	出現形の表層格 (nsubj,obj,iobj)
随意格	時間格(-TMP)，場所格(-LOC)	名詞による修飾 (nmod)	随意格 (obl)
関係節	空所の区別あり (-REL_sbj, …)	連体修飾節 (acl)	連体修飾節 (acl)
補充節	区別あり(-ADN)	連体修飾節 (acl)	連体修飾節 (acl)
補足節	区別あり(CP-THT)	区別あり (ccomp)	区別あり (ccomp)
連体修飾	節の区別あり (ADJ , -REL)	節の区別あり (amod, acl)	節の区別なし (acl)
連用修飾	節の区別あり (-ADV, ADV)	節の区別あり (advmod, advcl)	節の区別あり (advmod, advcl)
複単語表現 (機能語)	AUXCOMP,..	mwe	fixed

Universal Dependencies への変換

構築の方針:既存コーパス情報を活用

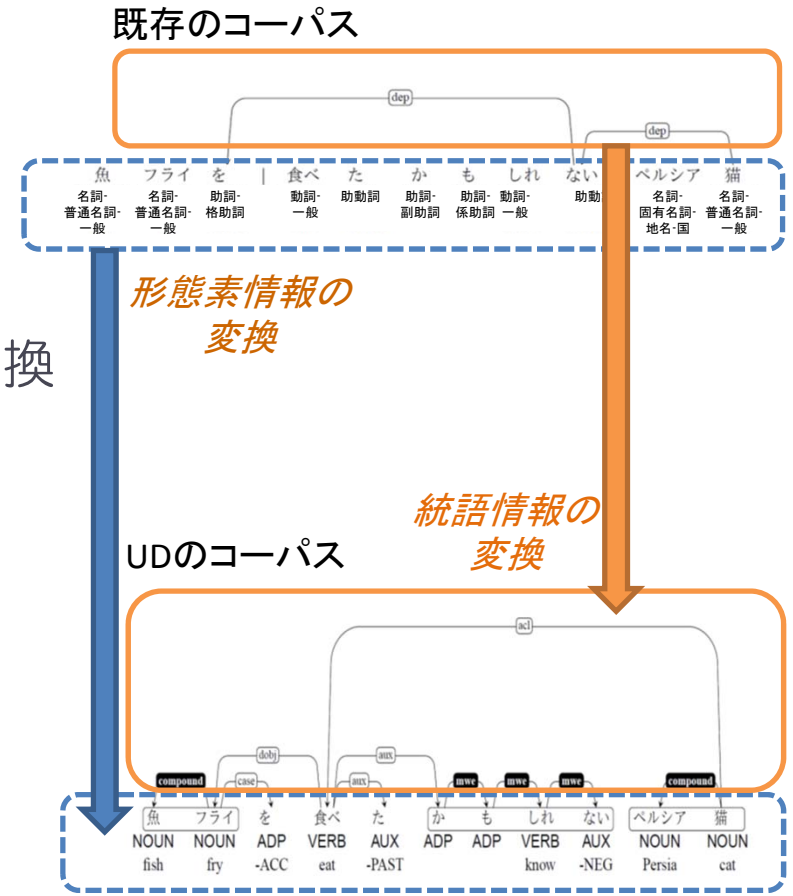
直接 UD のコーパスを構築しない

- 形態素情報が粗い（品詞17種類）
- 構造になじみがなく直接構築しづらい

=> 既存のコーパスから変換

- 形態素情報： BCCWJ の体系から変換
- 統語情報：文節依存構造，句構造から変換

ただし，他の構造からの変換は単純ではない



変換に必要な情報

(A) 品詞

- 品詞ラベルの対応情報

(B) 単語依存構造

- 主辞決定に関する情報
- 複単語表現に関する情報（特に機能表現）

(C) 依存関係ラベル

- 格関係(nsubj, obj, iobj, obl)
- 節の機能(acl, advcl, csubj, ccomp)
- 並列構造(conj)
- 複単語表現に関する情報（特に機能表現）

従来の係り受けコーパスのみでは不足する情報

日本語UDの表示例（CoNLL-U）

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	魚	魚	NOUN			2	compound	-	-
2	フライ	フライ	NOUN			4	obj	-	-
3	を	を	ADP			2	case	-	-
4	食べ	食べる	VERB	-	-	10	acl	-	-
5	た	た	AUX			4	aux	-	-
6	か	か	ADP			4	aux	-	-
7	も	も	ADP			6	fixed		
8	しれ	知る	VERB	-	-	6	fixed		
9	ない	ない	AUX	-	-	6	fixed	-	-
10	ネコ	猫	NOUN	-	-	0	root	-	-

(A) 品詞

Universal POS

(B) 単語依存構造

主辞のID
（矢印の根元）

(C) 依存関係
ラベル

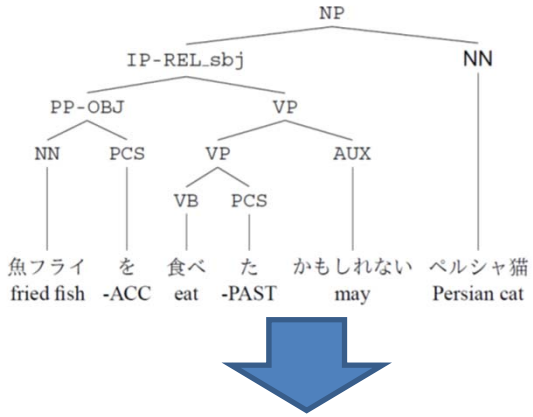
句構造からの変換

句構造ツリーバンク「楓」
- 1万文
- 文法機能タグ
  - 項の情報（-SBJ, -OBJ, -OB2）
  - 節の情報：関係節（IP-REL），補足節（CP-THT）など
  - 並列構造の情報: 並列(-COORD), 同格(-APPOS)

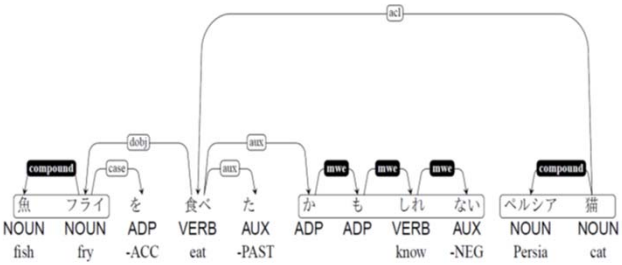
- (A) 単語の変換
  - 非終端記号からUPOSへの変換
  - 長単位から短単位への変換
- (B) 単語依存構造への変換
  - 部分木ごとの主辞決定規則
- (C) 依存関係ラベルの同定
  - 部分木からの依存関係ラベル変換規則

植松さんのスクリプトにより変換
※現在公開版は，UDv1版．
UDv2版は近日公開予定．

句構造

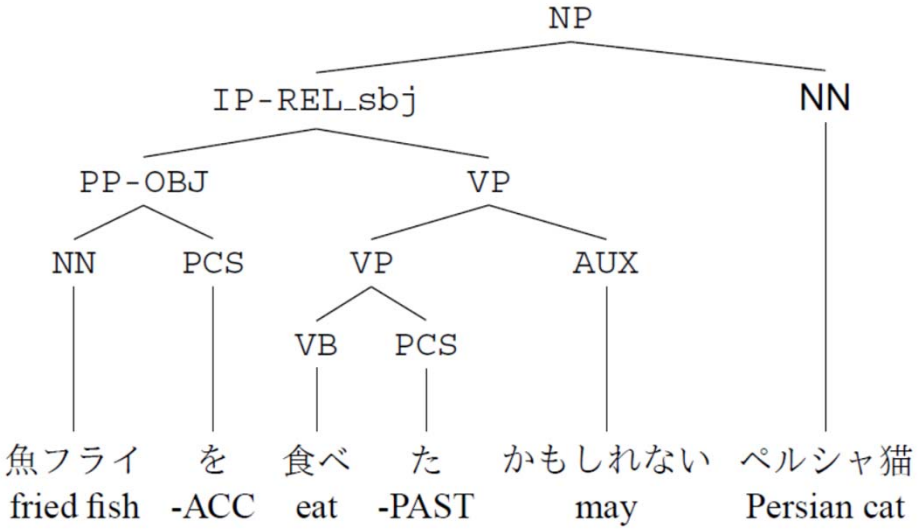


単語依存構造

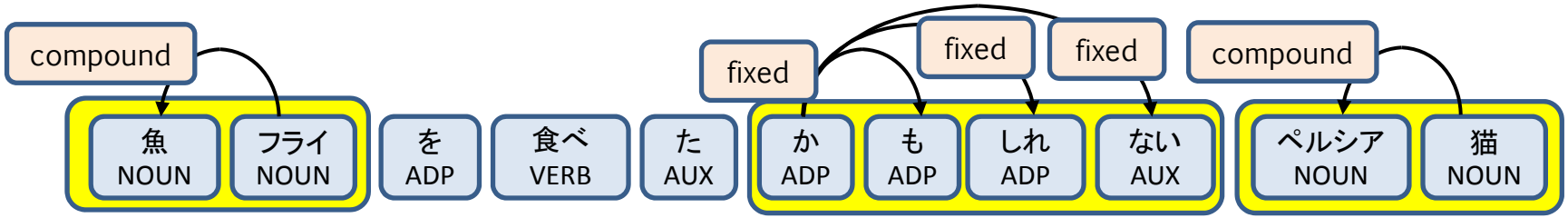


変換の概要

- (A) 単語の変換
- (B) 単語依存構造への変換
- (C) 依存関係ラベルの同定



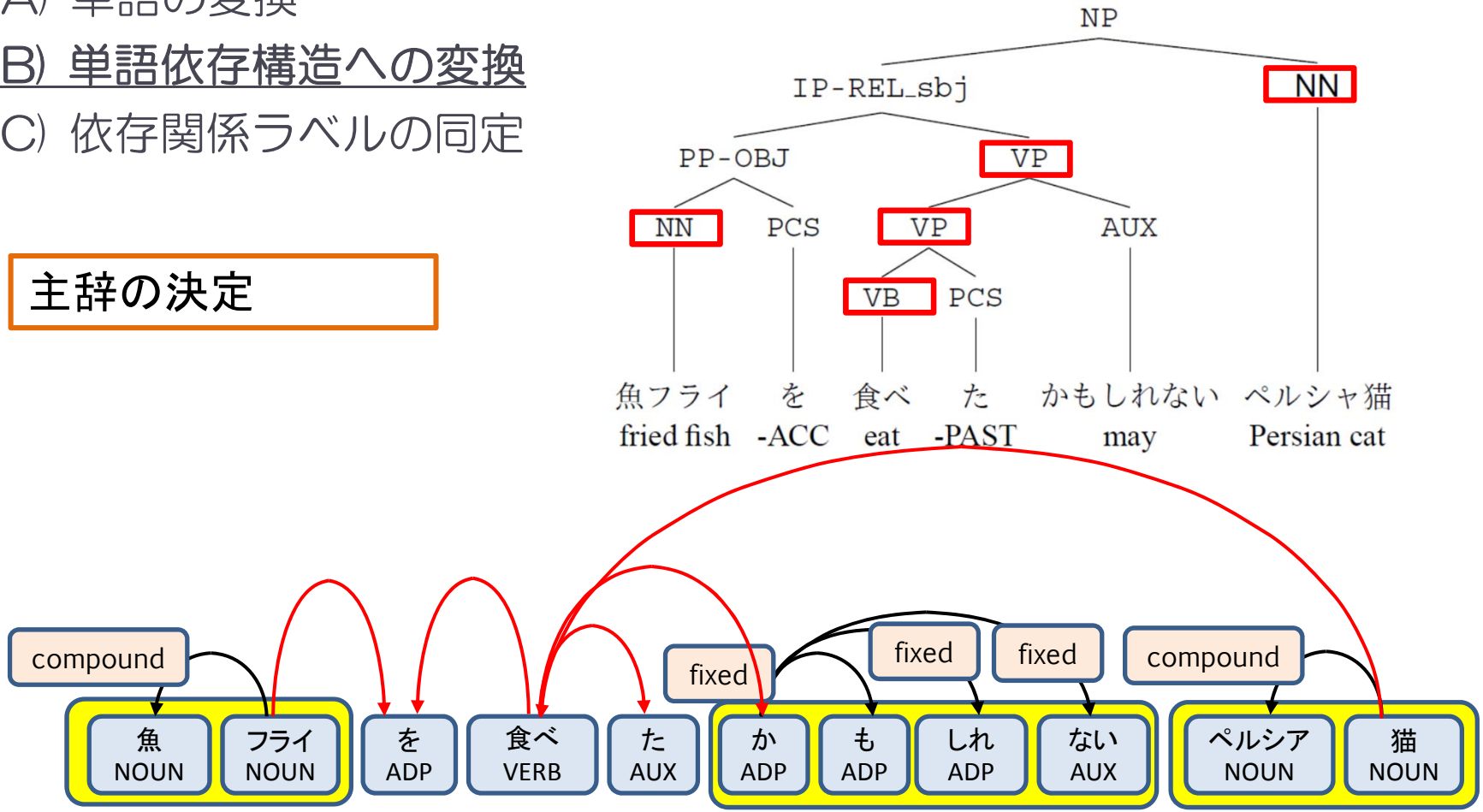
長単位-短単位分割
品詞の変換



変換の概要

- (A) 単語の変換
- (B) 単語依存構造への変換
- (C) 依存関係ラベルの同定

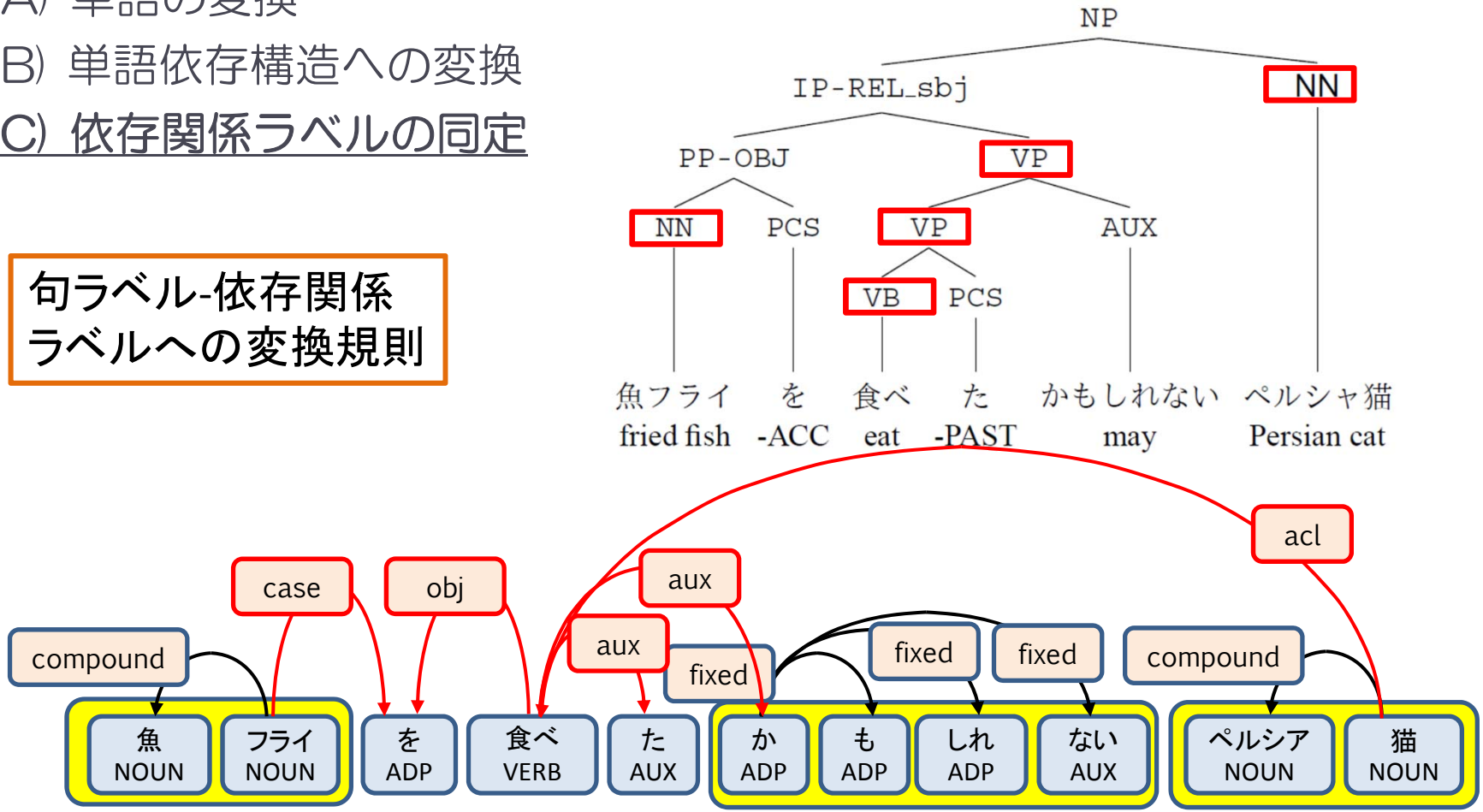
主辞の決定



変換の概要

- (A) 単語の変換
- (B) 単語依存構造への変換
- (C) 依存関係ラベルの同定

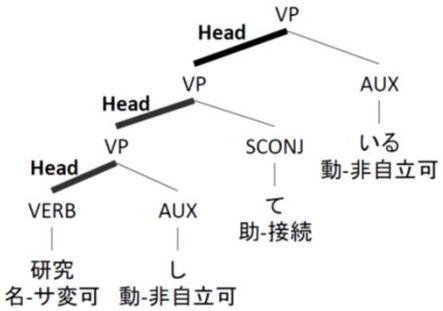
句ラベル-依存関係
ラベルへの変換規則

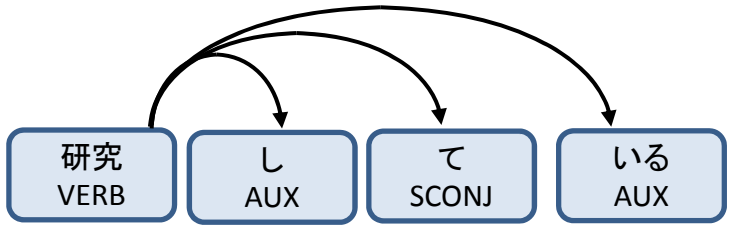


単語依存構造への変換

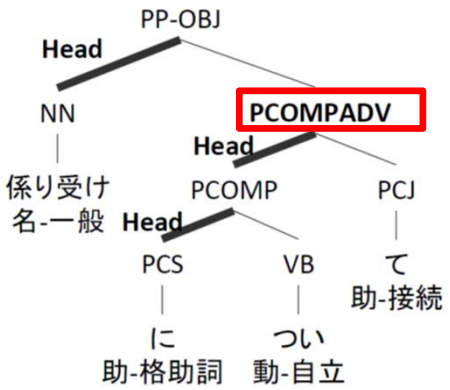
主辞の決定

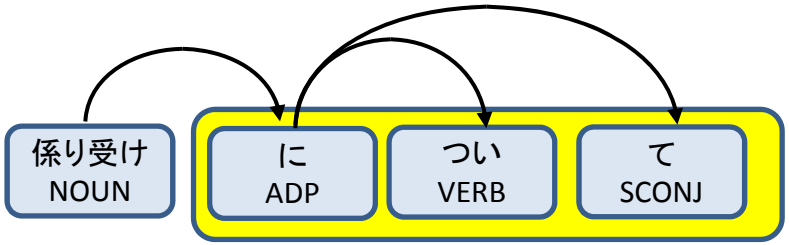
- 2分木の各分岐ごとに主辞決定規則を適用





- 機能語の複単語表現は，先頭主辞にする
  - Fixed タグ



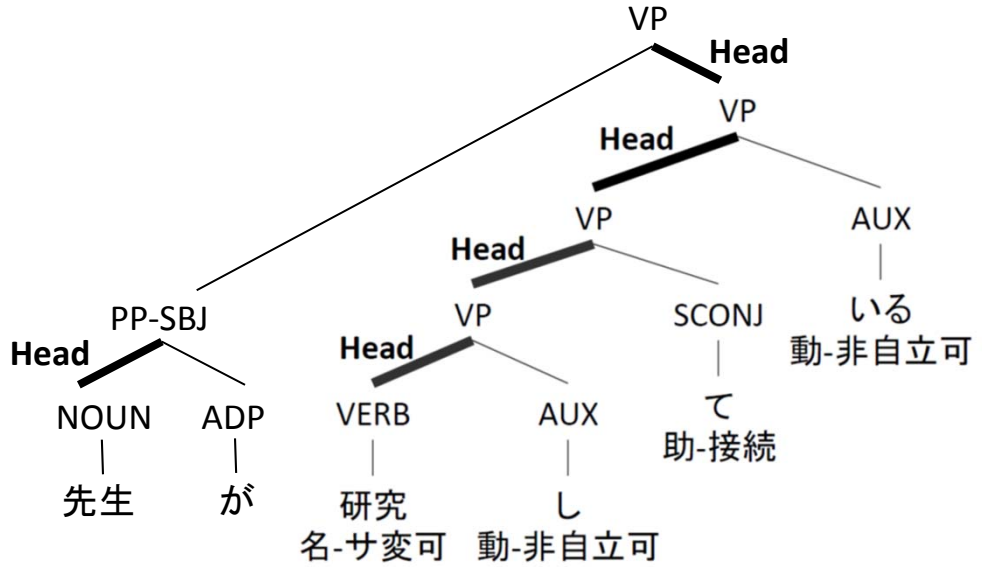


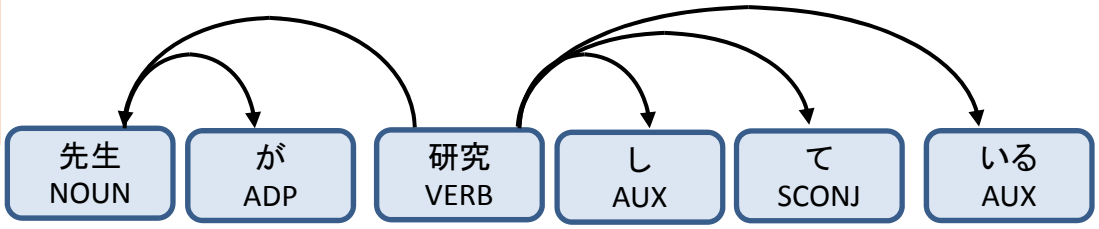
依存構造ラベルの同定

依存関係ラベル変換規則を部分木に順に適用する

- 各形態素から
  - 親をたどって自分が主辞でないノードCを探す
  - 規則に従ってラベルを出力
    - Cのラベル
    - Cの左の子Lのラベル
    - Cの姉妹ノードHのラベル
    - 形態素の基本形，品詞

(VP, VERB, *, AUX) => aux
(VP, VP, *, AUX) => aux
(VP, VP, *, SCONJ) => mark
(PP-SBJ, NOUN, *, ADP) => case
(VP, PP-SBJ, *, NOUN) = nsubj



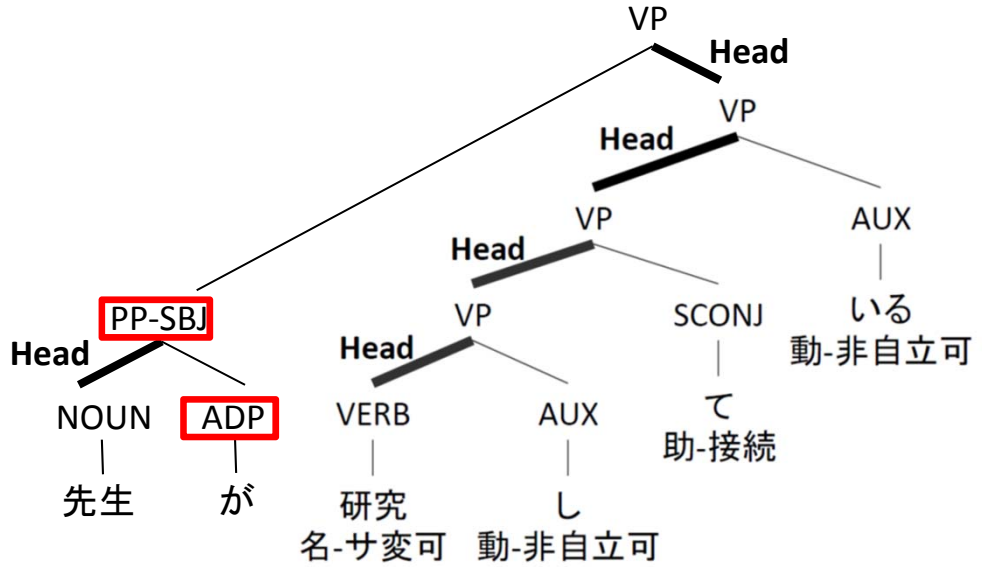


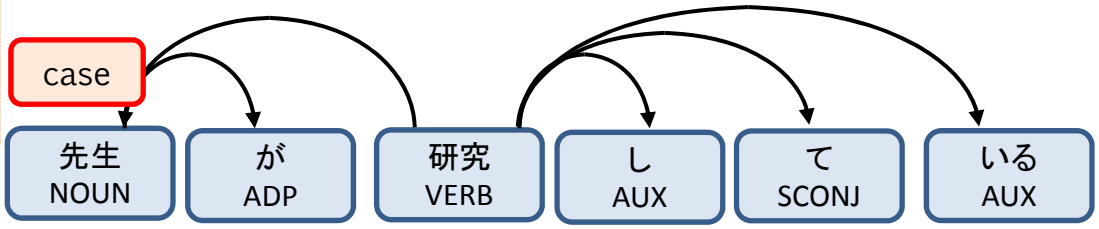
依存構造ラベルの同定

依存関係ラベル変換規則を部分木に順に適用する

- 各形態素から
  - 親をたどって自分が主辞でないノードCを探す
  - 規則に従ってラベルを出力
    - Cのラベル
    - Cの左の子Lのラベル
    - Cの姉妹ノードHのラベル
    - 形態素の基本形，品詞

(VP, VERB, *, AUX) => aux
(VP, VP, *, AUX) => aux
(VP, VP, *, SCONJ) => mark
(PP-SBJ, NOUN, *, ADP) => case
(VP, PP-SBJ, *, NOUN) = nsubj



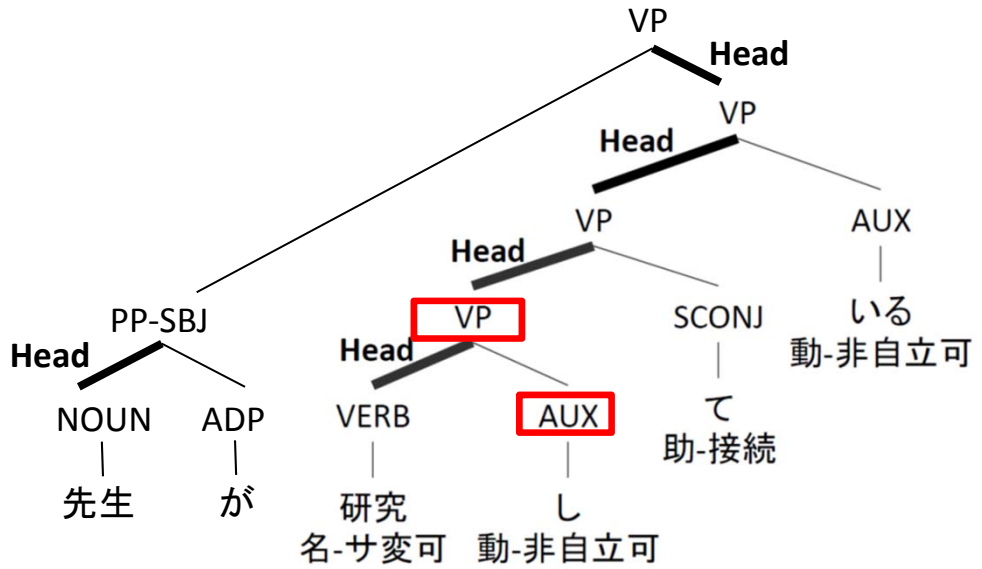


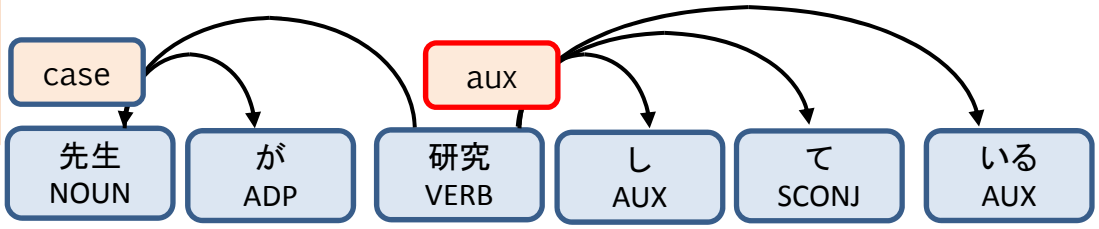
依存構造ラベルの同定

依存関係ラベル変換規則を部分木に順に適用する

- 各形態素から
  - 親をたどって自分が主辞でないノードCを探す
  - 規則に従ってラベルを出力
    - Cのラベル
    - Cの左の子Lのラベル
    - Cの姉妹ノードHのラベル
    - 形態素の基本形，品詞

(VP, VERB, *, AUX) => aux
(VP, VP, *, AUX) => aux
(VP, VP, *, SCONJ) => mark
(PP-SBJ, NOUN, *, ADP) => case
(VP, PP-SBJ, *, NOUN) = nsubj



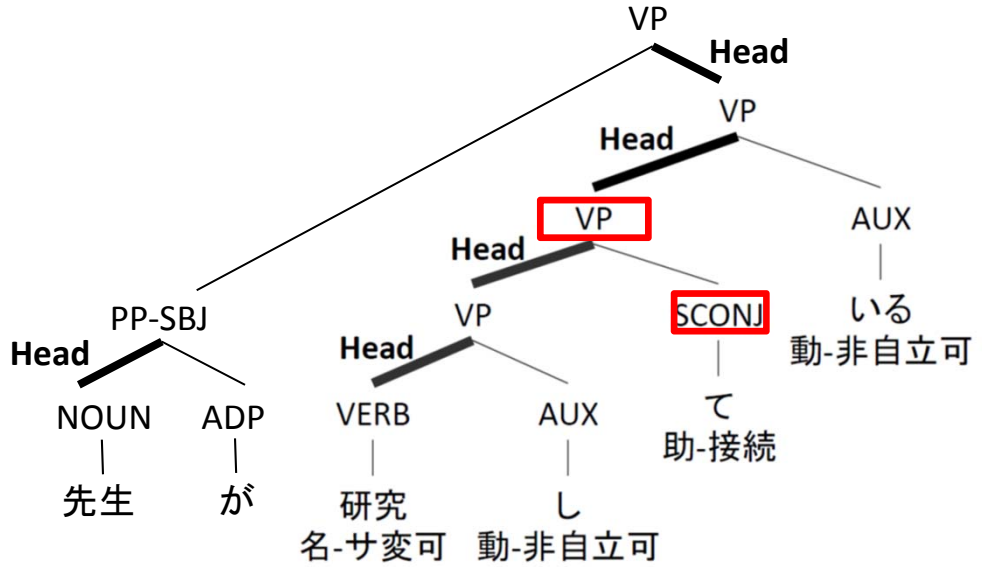


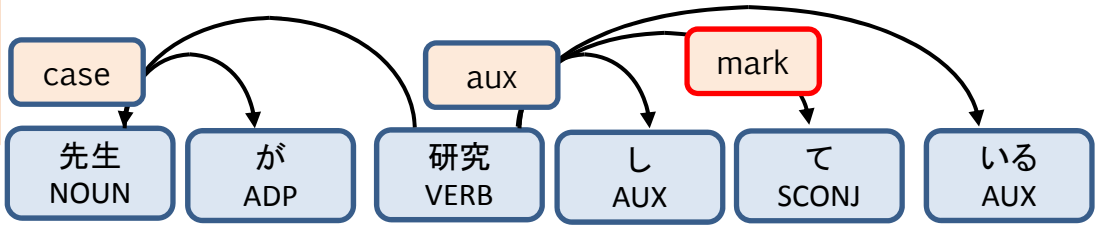
依存構造ラベルの同定

依存関係ラベル変換規則を部分木に順に適用する

- 各形態素から
  - 親をたどって自分が主辞でないノードCを探す
  - 規則に従ってラベルを出力
    - Cのラベル
    - Cの左の子Lのラベル
    - Cの姉妹ノードHのラベル
    - 形態素の基本形，品詞

(VP, VERB, *, AUX) => aux
(VP, VP, *, AUX) => aux
(VP, VP, *, SCONJ) => mark
(PP-SBJ, NOUN, *, ADP) => case
(VP, PP-SBJ, *, NOUN) = nsubj



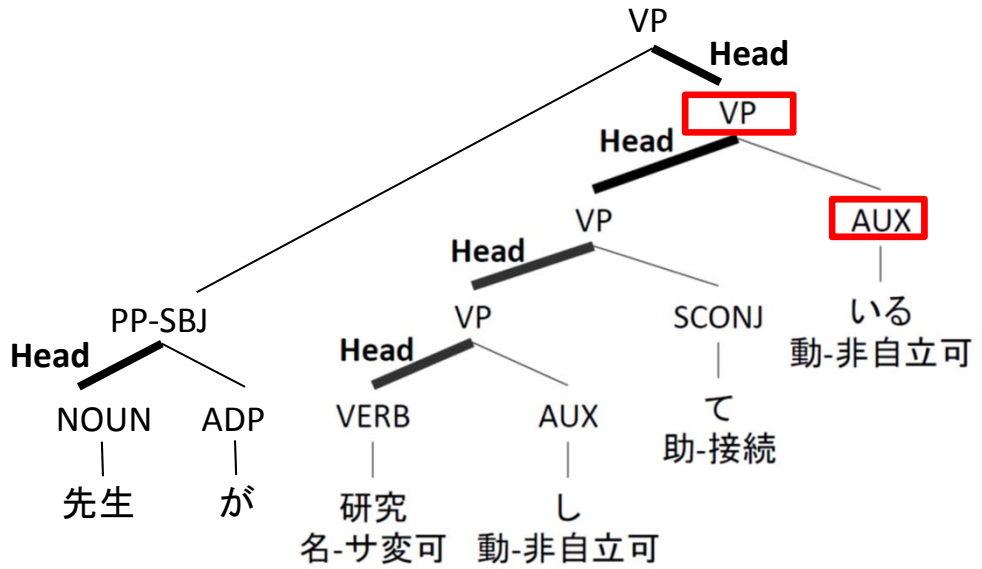


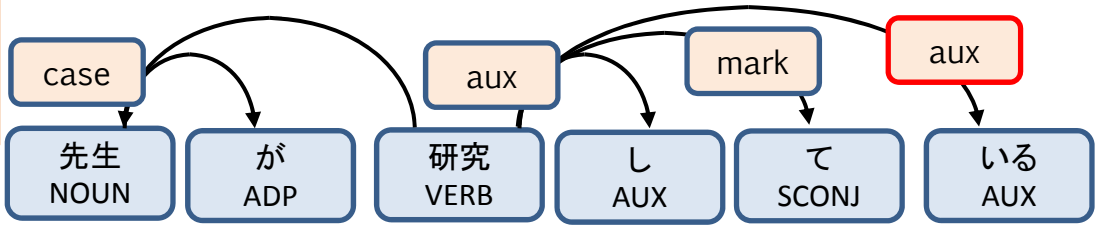
依存構造ラベルの同定

依存関係ラベル変換規則を部分木に順に適用する

- 各形態素から
  - 親をたどって自分が主辞でないノードCを探す
  - 規則に従ってラベルを出力
    - Cのラベル
    - Cの左の子Lのラベル
    - Cの姉妹ノードHのラベル
    - 形態素の基本形，品詞

(VP, VERB, *, AUX) => aux
(VP, VP, *, AUX) => aux
(VP, VP, *, SCONJ) => mark
(PP-SBJ, NOUN, *, ADP) => case
(VP, PP-SBJ, *, NOUN) = nsubj



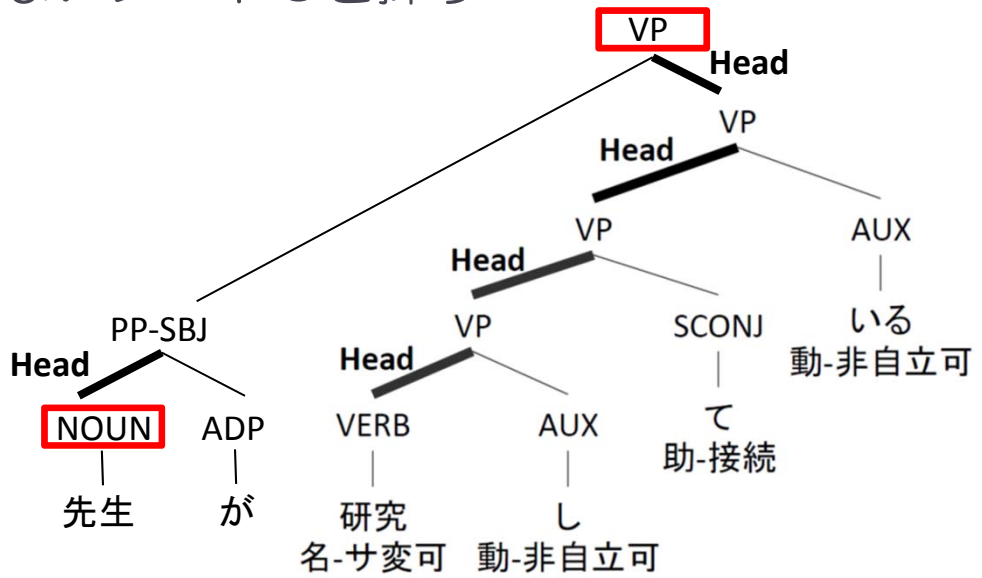


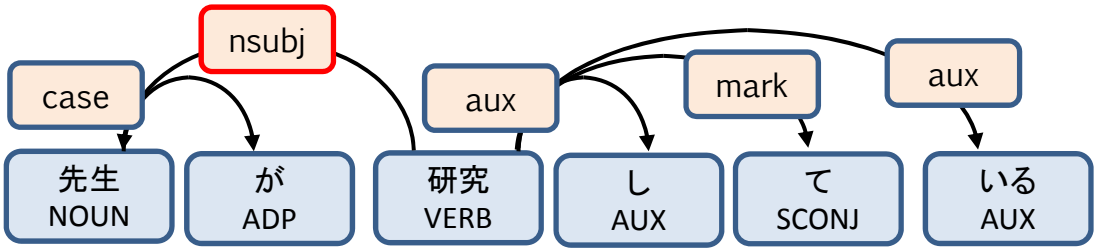
依存構造ラベルの同定

依存関係ラベル変換規則を部分木に順に適用する

- 各形態素から
  - 親をたどって自分が主辞でないノードCを探す
  - 規則に従ってラベルを出力
    - Cのラベル
    - Cの左の子Lのラベル
    - Cの姉妹ノードHのラベル
    - 形態素の基本形，品詞

(VP, VERB, *, AUX) => aux
(VP, VP, *, AUX) => aux
(VP, VP, *, SCONJ) => mark
(PP-SBJ, NOUN, *, ADP) => case
(VP, PP-SBJ, *, NOUN) = nsubj





変換のまとめ

既存コーパスから自動変換により構築

- (A) 単語の変換
- (B) 単語依存構造への変換
- (C) 依存関係ラベルの同定

句構造からの変換

- 品詞のマッピング
- 主辞決定規則
- 依存構造ラベル変換規則

おわりに

- 日本語の文法機能ラベル付き構文木
  - 日本語句構造ツリーバンク「楓」
    - 京都大学テキストコーパスの1万文
    - 2分木
    - 文法機能ラベル
  - 句構造から単語依存構造へ
    - UDとその他の単語依存構造
- Universal Dependencies への変換
  - 変換に必要な情報
    - 格関係，節の機能，並列構造，複単語表現
  - 変換方法の実際
    - 主辞規則（左右の子，親）
    - 句ラベル => 依存構造ラベル

今後の予定

- UDv1版 から UDv2版 への更新