日本語Universal Dependenciesへの複合辞情報付加の試み

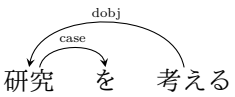
久保 大輝† 田中 貴秋‡ 進藤 裕之† 松本 裕治† 永田 昌明‡
†奈良先端科学技術大学院大学 情報科学研究科
‡ NTT コミュニケーション科学基礎研究所

†{kubo.daiki.kz7, shindo, matsu}@is.naist.jp
‡{tanaka.takaaki, nagata.masaaki}@lab.ntt.co.jp

1 はじめに

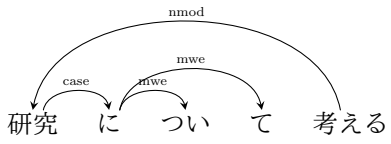
Universal Dependencies(UD)[1][1, 7] とは言語横断的な係り受け構造を設計する枠組みであり，2014 年には京都大学テキストコーパス [2] をもとにした日本語の UD コーパスが公開された．UD は，単語単位の依存構造を採用しており，原則として内容語を少なくとも一方に含む依存関係により統語構造を表す．そのため，複数の単語から構成される機能表現（複合辞）の扱いを適切に行う必要がある．例えば，文 1 では「研究」と「考える」という項と述語の関係が捉えられているが，文 2 では複合辞「について」の「つい」を内容語として扱ってしまい，重要な係り受け関係が直接捉えられない．適切に複合辞を考慮すると，文 3 のように「研究」と「考える」の関係を捉えることが出来る．

1. 研究を考える



2. 研究について考える

3. 研究について考える (複合辞を考慮)



UD には，依存構造のラベルとして複合辞に相当する mwe が定義されている [2] が，現状の日本語 UD コーパス (UD1.2) では，網羅的にアノテーションがされていないため，このコーパスを用いて複合辞を考慮した解析器を作るためには，より多くの複合辞に関する情報をコーパスに付与する必要がある．本研究は，複合辞を UD にアノテーションすることで，複合辞を考慮した構文解析のための言語資源の整備と構文解析の枠組みを整理することを目的とする．

複合辞のアノテーションを行うためには，同一の表記で複合辞として機能的意味で使われている場合と，内容的な意味で使われている場合を識別する必要がある．例えば，以下の文 1 と文 2 には「について」という同一表記の表現が現れているが，文 1 では「に/つい/て」の「つい」が，動詞「つく」という内容的な働きをしており，文 2 では「について」が 1 つの表現として機能的な働きをしている．

1. 親 について 歩く (内容的な働き)
2. 研究 について 話した (機能的な働き)

複合辞は，複数の形態素から構成され全体として機能的に働く表現 [5] であるが，定義や辞書の登録語が体系によって異なるため，使用する辞書によって解析結果が異なってしまう．例えば，IPADIC や JUMAN の辞書のように単語辞書に複合辞を登録すると，形態素解析が「どこまでを複合辞として認めるか」の問題に影響されてしまう．また，複合辞の解析は,「内容的用法と機能的用法の判別」の問題があるため，構文解析に関係づけて解析する必要がある．そこで我々は，最も基盤となる解析として，形態素として斉一な単語辞書である UniDic ベースの形態素解析を行い，その後段の階層で複合辞を含めた構文解析を行うこととする．

本稿では，UniDic[8] と機能表現辞書つつじ [5] をベースに複合辞辞書を構築し，辞書をもとに複合辞の

[1] http://universaldependencies.org/

[2] 本稿は version1 の UD に基づく

日本語 UD コーパスへの情報付加の試みについて報告する．

2 辞書の構築

本節では，日本語機能表現辞書つつじ [5] をもとに構築した新たな複合辞辞書の詳細について述べる．つつじは言語学的文献を参考にして得た見出し語 341 件について種々の異形を考慮した 16,801 種類の機能表現が収録されており，9 つの階層構造で見出し語，意味，文法的機能，機能語の交替，音韻的変化，とりたて詞の挿入，活用,「です/ます」の有無，表記の異なりを表現している辞書である．我々が構築する辞書は,「各々の見出し語は唯一の文法的機能を持つ」という方針をとり，つつじで見出し語に該当する第 1 階層ではなく，第 3 階層における 555 語の見出し語をもとに構築する．さらに見出し語およびその表記の異なりを，構成する短単位列に分解し，その標準形，品詞，読みの情報を追加する表 1 に，我々が構築した辞書の詳細を示す．また，本辞書の形式には JSON を採用する．mwe_id は複合辞を一意に識別するための ID を表している．末尾のアルファベットは複合辞の型を表しており，例えば，P は格助詞型，D は連体助詞型を意味する．headword は見出し語，mwe_pos は全体品詞，suw_lemma，suw_lemma_pos，suw_lemma_yomi はそれぞれ．headword を構成する短単位列の標準形，品詞，読みである．variation は headword の異表記，variation_lemma は variation の短単位標準形である．また，後節で詳細に述べるが，複合辞のコーパスへのマッピングは，suw_lemma と variation_lemma を用いて行う．meaning は複合辞の意味カテゴリ，left は複合辞の前方接続の制約である．接続制約は,「品詞, 活用型, 活用形, 語彙素」の 4 つからなり，前方の 2 形態素の制約を表す．また，表 2 につつじと本辞書の比較を示す．

表 1: 構築した辞書の詳細

属性	説明	属性値の例
mwe_id	MWE の ID	0011P
headword	見出し語	にとって
mwe_pos	全体品詞	格助詞型
suw_lemma	短単位標準形	[に, 取る, て]
suw_lemma_yomi	短単位標準形読み	[ニ, トル, テ]
suw_lemma_pos	短単位標準形の品詞列	[助詞-格助詞, 動詞-一般, 助詞-接続助詞]
variation	異表記	[にとりまして, にとり]
variation_lemma	異表記の短単位標準形	[[に，取る，ます, て],[に, 取る]]
meaning	意味カテゴリ	立場
left	左接続制約	[[名詞,*,*,*][*,*,*,*]]

2.1 機能表現の対象

つつじにおいて機能表現とは,「機能語」と「複合辞」からなる表現と定義されている．しかし，本研究は複合辞にのみ焦点を当てているため，辞書に登録する対象は「複合辞」のみとする．短単位一語からなる機能表現を機能語とし，我々が構築する辞書には登録しない [3]．この結果，エントリ数は，555 語から 427 語となった．また，機能表現の構文的な働きについての定義は，つつじにおける定義に則る．

表 2: つつじと本辞書の比較

	つつじ	本辞書
ファイル形式	XML	JSON
エントリ数 [4]	427	441
品詞体系	IPADic	UniDic
構成単語の形態素情報	なし	あり

2.2 品詞体系の変換

つつじには，機能表現の前方に接続される形態素の制約が,「品詞, 品詞細分類 1 , 品詞細分類 2, 品詞細分類 3, 活用型, 活用形, 原形」の形で IPADic の体系によって定義がされている．しかし，本研究でのアノテート対象である日本語 UD コーパスは，日本語の代表的な言語資源である「現代日本語書き言葉均衡コーパス」(BCCWJ) と同様に，UniDic の体系を採用している．我々はつつじで定義されている 96 個の制約に対して IPADic から UniDic 体系への対応付けを実施した．UniDic 体系における制約の定義は「品詞大分類-品詞中分類-品詞小分類-品詞細分類, 活用型, 活用形, 語彙素」とした．表 3 に対応付けの例を示す．

2.3 エントリの追加

つつじにおいて,「からといって」のような表現は接続助詞型であるが，これは「だからといって」のように変形すると接続詞型になることができる．一方で,「ところが」という表現は，つつじで接続詞型として登録されているが，これは接続助詞型にもなることができる．このように，つつじで接続詞型もしくは接続助詞型として登録されている表現の中で，その表現が接続助詞型もしくは接続詞型にもなりえる場合，新たにエントリとして本辞書に登録する．この結果，全 14 語のエントリを追加し，エントリ数は 427 語から 441 語に増加した．以下に具体的なエントリの追加の手順を示す，まず，接続助詞型から接続詞型へ変換する場合は，以下の手順で実施する．

1. 接続助詞型の内容を接続詞型にコピーする．

[3] 接続助詞型「ので」のように，つつじで一語であっても短単位列では 2 語以上になるものは複合辞として辞書に登録する

[4] 複合辞のみをカウント

表 3: 制約の対応付けと例

IPADic における制約		UniDic における制約
品詞, 品詞細分類 1, 細分類 2, 細分類 3, 活用型, 活用形, 原形		品詞大分類-中分類-小分類-細分類, 活用型, 活用形, 語彙素
例 1	動詞,*,*,*,*, 連用タ接続,*	動詞,*, 連用形-促音便,*
例 2	形容詞,*,*,*,*, 未然ウ接続,*	形容詞,*, 意志推量形,*

2. mweid の末尾を Q から C に, 品詞を接続助詞型から接続詞型に変更する.
3. 接続詞型の左接続制約を ”補助記号-読点,*,*,*; 補助記号-句点,*,*,*;None,None,None,None ”[5] とする.

次に，接続詞型から接続助詞型から接続詞型へ変換する場合は，以下の手順で実施する.

1. 接続詞型の内容を接続助詞型へコピーする.
2. mweid の末尾を C から Q に，品詞を接続詞型から接続助詞型に変更する.
3. 接続詞型の左接続制約に ”補助記号-読点,*,*,*; 補助記号-句点,*,*,*;None,None,None,None ” を加える.

3 アノテーションの事前検討

UD コーパスへのアノテーション実施の事前検討として，複合辞辞書のエントリをコーパスへ自動的にマッピングし，係り受け構造の自動付与を Stanford Parser[3] を用いて行った．それぞれの結果から，アノテーションの事前検討を行う.

3.1 コーパスへのマッピング

日本語 UD コーパスの全 9995 文に対して，構築した辞書のエントリ全 441 表現をマッピングの候補とし，コーパス中の各文へのマッピングを行う．マッピングの手順は，以下の通りである.

1. 各エントリの suw_lemma 及び variation_lemma の短単位列とコーパス中の各文内の短単位列が完全に一致する表現を探す.
2. マッチした表現の 2 つ前方の形態素及び直前の形態素の「品詞，活用型，活用形」と，辞書中の左接続制約を比較し，一致しないものを削除する.

コーパスへのマッピングを実施した結果，7314 短単位列がマッピングされた．また，マッチした複合辞のうち，コーパスにもともと付与されている複合辞を除くと，5833 短単位列であった．従って，我々が構築した辞書を用いることによって，UD コーパスへの複合辞のアノテーションが，より網羅的になることが示唆される.

[5]; は論理和の「OR」を意味し，None は,「いかなる形態素も接続してはならない」という意味を示す

3.2 係り受け構造の自動付与

第 1 節の文 3 のような複合辞を考慮した係り受け構造を自動で付与することを試みる．まず，複合辞がアノテーションされた日本語 UD コーパスを，複合辞が含まれている文と含まれていない文の 2 つの集合に分割し，複合辞が含まれていない文集合を用いて，Stanford Parser[3] によって係り受け解析のモデルを学習する．構築したモデルを用いて複合辞が含まれている文集合の係り受け解析を行った[6]．この際，複合辞は連結して 1 単語と扱って解析を行った．結果の中から，格助詞型・助動詞型・接続詞/接続助詞型をそれぞれ 10 個ずつランダムサンプリングを行い，(a) 複合辞とその主辞 (名詞や動詞) と，(b) 複合辞の主辞 (名詞や動詞) とさらにその主辞の 2 種類の係り受けの正解率 (UAS) を算出した．結果を表 6 に示す.

表 6: ランダムサンプリングの結果

	(a)	(b)	用例
格助詞型	8/10	6/10	「について」
助動詞型	9/10	9/10	「ことができる」
接続詞・接続助詞型	10/10	8/10	「ばかりか」

3.3 考察

短単位列と左接続制約のマッピングだけでは，一意に品詞が割り当てられないものが存在し (例えば，複合辞「という」に対して助動詞型と連体助詞型)，そのようなものについては人手で判定する必要がある．さらに，複合辞「を通して」や「ことは」のような，構成形態素の中に内容語を含む複合辞については，機能的な用法ではない箇所にマッピングされてしまう事例が多くみられ，人手による用法判定の必要性が示唆された．また，複合辞「ために」はつつじにおいて，接続助詞型としてのみ登録されているが，例えば,「不心得者のために失墜する」のような文において接続助詞型としてマッピングされており，これは誤りである．そこで，新たな見出し語として,「名詞+のために」の形で格助詞型の複合辞として登録する必要がある．このように，新たな見出し語の追加をする余地があることが明らかとなった．次に，係り受け構造の自動付与の結果においては，(a)(b) ともに，完全には自動で係り

[6]同一箇所に複数の品詞がマッピングされたものは除去

表 4: 品詞カテゴリとその用法

つつじにおける品詞	新たな品詞カテゴリ	用法
格助詞型, とりたて詞型 提題助詞型, 形式名詞型	格助詞型	名詞+○+動詞/形容詞
連体助詞型	連体助詞型	動詞/形容詞/名詞+○+名詞
接続助詞型, 接続詞型	接続詞型	文/節+○+文/節
助動詞型	助動詞型	動詞+○

表 5: アノテーション例

前文脈	注釈該当部	後文脈	選択肢	機能的用法か	品詞選択
中西忠夫専務	によると	、「地下でもポケベルは鳴ります」とのこと。	1 → P,0 →その他	○	1
しかし、三日までに提案	に応じた	のは一割強の五百五百四十四十人。	1 → D,0 →その他	×	×

受け構造を正確に付与できておらず，自動付与を行った後の人手による修正作業が必要であることが明らかとなった．(a) 複合辞とその主辞については，直前の名詞や動詞を主辞にすることが多いこともあり，高い精度で自動付与でき．(b) その主辞 (名詞や動詞) とさらにその主辞については，(a) より若干精度が低くなっており，人手による修正作業のコストがより多くかかることが示唆される．

4 アノテーションの方針

前節の考察を踏まえ，今後の複合辞のコーパスへのアノテーションの方針とその手順について述べる．まず,「には」のような構成形態素に内容語が含まれない複合辞は用法の曖昧性が無く,「に当たり」のような内容語が含まれている表現にのみ用法の曖昧性があると仮定し，後者のみをアノテーションの対象とする．そのため，コーパスでマッピングした 7314 短単位列の中から，内容語が含まれている複合辞を対象にアノテーションを行う．また，アノテーション作業における作業品質の一貫性とアノテーションの簡易化のため，つつじで定義されている 8 種類の品詞カテゴリを用法ごとに 4 種類に集約し，それぞれの用法を定義する (表 4)．アノテーションの手順は，まず「機能的用法であるかどうか」の判断を行い，機能的用法である場合は，該当する品詞を選択し，該当する品詞がない場合は、0(その他) を選択する．機能的用法でない場合は，品詞の選択は行わないこととする．表 5 に，アノテーションの例を示す．

5 関連研究

複合辞が付与されたコーパスは，BCCWJ においては，助詞相当 75 語，助動詞 55 語の複合辞が収録されている．また，現代語複合辞用例集 [4] の代表的複合辞一覧に基づいて，それらの派生形である 337 種類の機能表現を規定し，1995 年の毎日新聞の記事に対して内容的用法と機能的用法の区別をアノテートし，各複合辞ごとに最大 50 件の用例を収録した日本語複合辞用例データベース [6] や，BCCWJ2008 及び 2009 を用いて作成された BCCWJ 複合辞辞書 [9] がある．

6 おわりに

本稿は，複合辞の辞書構築と，日本語 Universal Dependencies へのアノテーションの検討を行った．構築した辞書エントリをコーパスへマッピングした結果から，日本語 UD コーパスへのより網羅的なアノテーションが可能であることが確かめられた．今後は，実際に UD コーパスへの人手によるアノテーション作業を行っていく予定である．

参考文献

[1] Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, et al. Universal dependency annotation for multilingual parsing. In ACL (2), pp. 9297, 2013.

[2] Daisuke Kawahara, Sadao Kurohashi, and Koiti Hasida. Construction of a Japanese relevance-tagged corpus. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), pp. 20082013, 2002.

[3] Danqi Chen and Christopher D Manning. A Fast and Accurate Dependency Parser using Neural Networks. In EMNLP, pp. 740-750, 2014.

[4] 国立国語研究所：現代語複合辞用例集 (2001)

[5] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理 14(5), pp. 123-146, 2007.

[6] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一. 日本語複合辞用例データベースの作成と分析. 情報処理学会論文誌 47(6), pp. 1728-1741, 2006.

[7] 金山博, 宮尾祐介, 田中貴秋, 森信介, 浅原正幸, 植松すみれ. 日本語 Universal Dependencies の試案. 言語処理学会第 21 回年次大会, pp. 505-508, 2015.

[8] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. 日本語科学 (22), pp.101-123, 2007.

[9] 近藤泰弘. BCCWJ 複合辞辞書について (＜小特集＞ 日本語学・日本語教育)(安田尚道教授退任記念号). 青山語文 42, pp. 10-15, 2012.