日本語 Universal Dependencies への複合辞のアノテーション

久保 大輝 †　田中 貴秋 ‡　進藤 裕之 †　松本 裕治 †
† 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@lab.ntt.co.jp

1　はじめに

機能表現とは，1 つの形態素からなる「機能語」と，複数の形態素から構成され，全体として機能的に働く「複合辞」からなり，機能表現の中でも「にあたって」，「について」のような表現に代表される「複合辞」においては，同一の表記で内容的な意味を持つ場合と機能的意味を持つ場合がある．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　(内容的用法)
2. 研究 について 話す　（機能的用法）

このような表現おいて，同一表記で内容的な用法の場合と機能的な用法の場合を識別する必要があり，そのためには，複合辞の辞書および，複合辞の用法を注釈付けした言語資源の整備は不可欠である．言語資源の 1 つに，Universal Dependencies(UD)[1, 2, 13] がある．UD は，言語横断的な係り受け構造を設計する試みであり，依存構造のラベルとして複合辞に相当する *mwe* が定義されているものの，日本語 UD においてはルールベースによって *mwe* のラベル付けがされており正確なものではない．

本稿では，日本語機能表現辞書「つつじ」をもとに，UD のための新たな機能表現辞書を構築し，その辞書を用いて人手による日本語 UD へのアノテーションを実施した結果について報告する．

2　日本語機能表現辞書「つつじ」

現在，無料で公開されている電子化された機能表現の辞書として，日本語機能表現辞書つつじ (以下，つつじ)[4] がある．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，全 16,801 種類の機能表現が収録され，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．

3　辞書の構築

本節では，機能表現辞書つつじをもとに構築した新たな機能表現辞書の詳細について述べる．本辞書は，各々の見出し語は唯一の機能を持つという方針をとるため，つつじの第 3 階層の語を見出し語とする．

3.1　機能表現の定義

つつじにおける機能表現は,「機能語」と「複合辞」からなる表現と定義されている．本研究は複合辞にのみ焦点を当てているため，機能表現を構成する表現は複合辞のみであると定義し，つつじにおける**数える必要あり**の機能語を全て除去する．しかし，機能表現の働きについての定義は，つつじの定義を則ることとする．

3.2　品詞体系の変換

機能表現の前後形態素の品詞・活用形等の制約が，つつじにおいて IPADic の体系で定義がされている．しかし，本研究でアノテート対象のコーパスである日本語 UD や，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) においては，UniDic[14] の体系を採用しており，今後の日本語における言語処理は UniDic が主流になると予想されることもあり，本辞書を広く使用してもらうためにも，制約を IPADic から UniDic に対応させる必要がある．

3.3 エントリの追加

4 日本語 UD へのアノテーション

4.1 機能表現のマッピング

4.2 アノテーションの仕様

5 コーパスの分析

6 関連研究

複合辞がアノテートされたコーパスは，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 [3] 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，つつじで定義されている意味体系を再構成した 116 種類の意味ラベルを定義し，BCCWJ の Yahoo!知恵袋ドメインの一部における各文の述部に付随する機能表現を対象にラベルを付与したコーパス [12] がある．

7 おわりに

本稿では，

参考文献

[1] Universal Dependencies contributors. Universal dependencies. https://universaldependencies.github.io/docs/, 2014.

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

[3] 国立国語研究所：現代語複合辞用例集 (2001)

[4] 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.

[5] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史. 日本語機能表現の自動検出と統計的係り受け解析への応用. 自然言語処理 14(5), pp. 167-197, 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] 土屋雅稔, 注連隆夫, 高木俊宏, 内元清貴, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一. 機械学習を用いた日本語機能表現のチャンキング. 自然言語処理 14(1), pp. 111138, 2007.

[8] 小早川健, 関場治朗, 木下明徳, 熊野正, 加藤直人, 田中英輝. 単語格子とマルコフモデルによる日本語機能表現の解析: 日本語機能表現辞書 「つつじ」 を用いて (解析). 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション 109(142), pp. 15-20, 2009.

[9] 鈴木敬文, 宇津呂武仁, 松吉俊, 土屋雅稔. 代表・派生関係を利用した日本語機能表現の解析. 情報処理学会研究報告 (2010-NL199-6), pp. 1-9, 2010.

[10] 長坂泰治, 宇津呂武仁, 土屋雅稔. 大規模日本語機能表現辞書の階層性を利用した機能表現検出. 言語処理学会第 14 回年次大会, pp. 837-840, 2008.

[11] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔. 『現代日本語書き言葉均衡コーパス』における複合辞の検出と評価, コーパス日本語学ワークショップ (2012), 2012.

[12] 上岡裕大, 成田和弥, 水野淳太, 乾健太郎. 述部機能表現に対する意味ラベル付与. 情報処理学会研究報告 第 216 回自然言語処理研究会, pp. 1-9, 2014.

[13] 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.

[14] 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.