UD Japanese BCCWJ: 現代日本語書き言葉均衡コーパスのUniversal Dependencies

大村 舞　　浅原 正幸
人間文化研究機構 国立国語研究所
{mai-om, masayu-a}@ninjal.ac.jp

1 はじめに

Universal Dependencies [6](以下 UD) は、多言語で一貫した構文構造とタグセットを定義し、言語間での共通した依存構造タグ付きコーパスを提供することを目的とした活動、あるいはそのコーパスのことを指す。共通した枠組みのコーパスを構築することで、多言語横断構文解析や、他の言語のコーパスを用いた言語横断的な学習、言語間の定量的な比較の実現を目指している。我々は UD の日本語版を設計する活動として、品詞体系、ラベル付き依存構造の定義の策定、その文書化と、参照用のコーパスの作成を進めている [1, 10]。

本稿ではこの UD 日本語版設計の活動の一環として、現代日本語書き言葉均衡コーパス [4]（以下 BCCWJ）に基いた日本語 UD コーパス **UD Japanese BCCWJ** について説明する。BCCWJ は日本語について入手可能な唯一の大規模均衡コーパスであり、BCCWJ から UD を構築することによって約 100 万単語規模のコーパスを提供することが可能となる。

尚本稿は文献 [14] からいくつか変更となった内容のみ示す。そのため、より詳細な変換規則などは文献 [14] を参照されたい。

2 日本語の依存構造コーパスとUniversal Dependencies

これまでにも日本語の依存構造解析（係り受け解析）のために京都大学テキストコーパス [3] 、日本語係り受けコーパス [7] などのコーパスが開発されている。日本語の依存構造コーパスは日本語の統語構造の基本的な単位である文節間の係り受け関係を表したデータが主となっている。

Universal Dependencies は、すべての構文構造を単語間の依存関係と関係のラベルで表現している。語順が自由な言語も含めて言語横断的に共通化した体系を確立するために、図 1 のように内容語間の依存構造を中心とした表現を用いる。異なる言語間で依存構造解析器の性能比較を行うだけでなく、言語学的に類型論的な分析が可能にすべく言語横断的な設計を目指している。現在の UD におけるアノテーション体系 (version 2.0) は、Google Universal Part-of-speech Tags [5] を基にして 17 種類の品詞ラベル Universal PoS tags が定義されている。さらに Universal Stanford Dependencies [9] を基にして 37 種類の係り受けのラベル Universal dependency relations が定義されている[1]。

現在 UD 基準の日本語依存構造タグ付きコーパスとしていくつかのコーパスが公開されている。**UD Japanese KTC** [10] は日本語句構造ツリーバンク [11] を変換した日本語版 UD コーパスである。このコーパスは日本語句構造ツリーバンクにある形態素や句構造などのアノテーションを用いて変換されたものである。**UD Japanese** は Wikipedia 由来の UD であり、**UD Japanese PUD** は 1000 文程度の多言語横断パラレルコーパスの UD である。また、日本語歴史コーパス (CHJ) [8] に基づいた **UD Japanese Modern** [14] も構築した。現在公開・開発中の日本語版 UD の一覧を表 1 に示す。このうち **UD Japanese, Japanese KTC, Japanese PUD** の 3 つのコーパスは 2018 年 1 月現在 `http://universaldependencies.org/` にて配布されており、**UD Japanese BCCWJ** や **UD Japanese Modern** も公開予定である。

3 現代日本語書き言葉均衡コーパスについて

現代日本語書き言葉均衡コーパス [4](BCCWJ) は、書籍全般、雑誌全般、新聞、白書、ブログ、 ネット掲示板、教科書、法律などのジャンルにまたがって 1 億

[1] いずれのラベルも http://universaldependencies.org/ にて一覧が公開されている

表 1: 日本語の Universal Dependencies 一覧

	単語数	基コーパス	備考
UD Japanese	189K	Wikipedia	
UD Japanese KTC	186K	毎日新聞 95 年度	日本語句構造ツリーバンクに基づいて構築
UD Japanese PUD	26K	-	パラレルコーパス
UD Japanese BCCWJ	1,098K	BCCWJ	大規模コーパス、依存構造 (BCCWJ-DepPara) に基づいて構築
UD Japanese Modern	14K	日本語歴史コーパス	古文コーパス

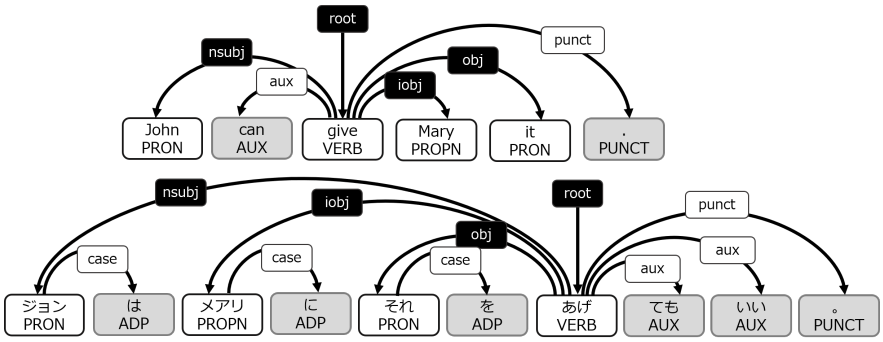

図 1: Universal Dependencies のイメージ、上が英文、下が日本語. 助動詞や格助詞など、英語と日本語の違いがあっても、内容語の関係は保たれている

430 万語のデータを格納したコーパスであり、現在、日本語について入手可能な唯一の均衡コーパスである。このうちコアデータである 1980 サンプル・57256 文には二種類の形態論情報（短単位・長単位）が付与されている。

BCCWJ は他の日本語統語コーパスと比較すると依存構造情報が付与されているデータのみでも約 100 万語収録されており、UD Japanese BCCWJ が構築されれば表 1 に示すとおり、大規模な日本語版 UD が公開されることになる。

BCCWJ には、短単位・長単位の形態論情報だけでなく、文節単位の依存構造・並列構造アノテーションである BCCWJ-DepPara [2] や述語項構造情報アノテーションである BCCWJ-PAS [13] が提供されている。UD Japanese BCCWJ はこれらの情報を利用することによって構築している。

4 UD Japanese-BCCWJ について

BCCWJ や BCCWJ-DepPara などに収録されている既存のアノテーションに基づき、変換プログラムを構築し変換することで、UD 本体の基準の変更や日本国内での議論に対応している。以下では UD Japanese-BCCWJ における単位認定・品詞割り当て・依存構造ラベル割り当てなどについて説明する。

4.1 単語認定

日本語は英語とは異なり、単語に分割されていない。そのためまず単語の認定について決める必要がある。従来の日本語依存構造解析で用いられている統語関係の単位としては**文節**が用いられている。BCCWJ のすべてのサンプルは**短単位・長単位**という言語単位に基づいて形態素解析されている。**短単位 (Short unit word, SUW)** は日本語の形態的側面に着目して規定した単位であり、語種ごとに規定した最小単位の線形結合に基づき定義されている。**長単位 (Lont unit word, LUW)** は日本語の構文的な機能に着目して規定した単位であり、文節の構成要素ともなっている。短単位・長単位・文節は図 2 のように短単位＜長単位＜文節という階層関係が成り立っている。

現在公開されている UD Japanese のリソースでは、UD の単語単位として BCCWJ の品詞体系である短単位を基本単位として採用している。ただし、UD ガイドラインで規定されている単語単位は syntactic words となるように単語単位を制定すると定義されており、

短単位	魚	フライ	を	食べ	た	か	も	しれ	ない	ペルシャ	猫
	NOUN	NOUN	ADP	VERB	AUX	PART	ADP	VERB	AUX	PROPN	NOUN
長単位	魚フライ		を	食べ	た	かもしれない				ペルシャ猫	
	NOUN		ADP	VERB	AUX	AUX				NOUN	
文節	魚フライを			食べたかもしれない						ペルシャ猫	

図 2: 短単位・長単位・文節の関係、短単位＜長単位＜文節という階層関係が成り立っている

UD や他の言語と基準を合わせることを踏まえると、短単位は統語関係を表現するのに尤もらしい単語単位と必ずしも言えない [2]。また過去の研究では短単位を基準として調査されているものは多い一方、長単位についてはあまり検討されていない。そのため、長単位ベースのリソースも公開することにより適した単語単位を検討する必要がある。

4.2 Universal PoS tags への変換

UD では全言語の品詞を集約するための体系として Universal PoS version 2.0 を採用している。Universal PoS version 2.0 では、17 種の品詞が定義されている。品詞の細分類や、性数・時制・格など文法的属性に関するものは、`FEATS` や `MISC` など列に言語ごとの個別に定義する属性値 (features) を持たせることで情報が失われないようにしている。

UD Japanese BCCWJ では UniDic [12] と Universal PoS tags との対応表を構築することで UD の品詞を定義する。BCCWJ の品詞体系は短単位と長単位で異なっている。短単位では**語彙主義 (lexicon-based)** 的な可能性に基づく品詞体系を採用している。例えば「名詞-普通名詞-副詞可能」は「名詞」用法も「副詞」用法もある語彙であることを意味する。長単位では文脈に基づいてこの用法の曖昧性を解消する**用法主義 (usage-based)** に基づく品詞を規定している。BCCWJ には長単位形態論情報としてこの用法主義に基づく品詞である「用法」の情報が付与されており、短単位からも参照できるようになっている。

UD のガイドラインには、語彙主義の品詞体系か用法主義の品詞体系に基づく品詞体系どちらを採用するべきかの制定は明示されていない。現在の方針では、長単位で用いられる用法主義に基づく品詞体系を採用することを検討している。そのため、長単位に基づくのコーパスでは用法主義に基づく品詞体系と Universal PoS tags への変換を行う。丁度長単位で採用されている用法主義に基づく品詞体系は Universal PoS tags とほぼ 1 対 1 で対応することが分かっている。

短単位に基づくコーパスの場合、基本的に前述した語彙主義に基づく品詞体系から Universal PoS tags への変換を行う。ただしいくつかの単語に関しては、長単位と同様に用法主義に基づく品詞体系を用いる。例えば、サ変名詞や形状詞の場合は語彙主義に基づく品詞体系ではなく、文脈に基づいて用法の曖昧性を解消する用法主義に基づく品詞を用いる。用法主義に基づく品詞のほうが、他の言語との対応がとりやすいという利点があるということと、語尾の有無などにより揺れが少なく `VERB`, `ADJ` とする条件を規定し易いからである。短単位に基づく変換規則に関しては [14] に掲載している。

4.3 文節から単語への依存構造の変換

BCCWJ-DepPara [2] のような文節依存構造の情報には、文節間の依存関係情報は含んでいるものの、Universal dependency relations に対応する単語間の依存構造に関する統語的な用法情報を含んでいない。そのため、Universal PoS tags に変換した後、文節依存構造の情報から UD における単語ベースの依存構造へ変換し、依存構造関係がある単語対の品詞情報に基づいて Universal dependency relations ラベル（以下 UD ラベル）を決定する。

文節依存構造を単語間の依存構造に変換する規則は以下のように規定している。

1. 文節内の主辞を決定し、文節内の他要素に関してはすべて文節内の主辞に係けるようにする。文節内の主辞として内容語と機能語が分かれる内容語の最右の語を採用する。
2. 文節間の依存関係については従来の依存関係を採用する。

単語間依存構造に変換した後、UD のガイドラインに従って、UD ラベルをルールベースで規定していく。

[2] 例えば `http://www.cjvlang.com/Spicks/udjapanese.html` では文節を用いたほうがよいと議論されている

[3] ただ後述のように BCCWJ-PAS に基づく深層格での付与は行わなくなったため、`nsubj, obj` などの割当規則に変更がある

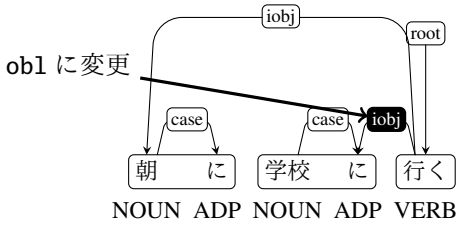

図 3: 表層格に基いてラベルを付与すると衝突する例、この場合 obl に変更する

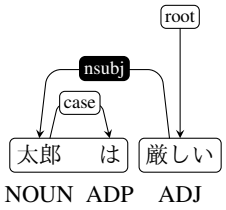

図 4: 主題を表す助詞「は」が現れた場合の例、この場合は nsubj を付与する

具体的な [14] に Universal dependency relations の割当の詳細を掲載している [3]。

UD の規定によると格関係を表すラベルは、深層格のような意味的な関係ではなく、統語関係に基いて用言情報を付与する必要がある。そのため、格関係を表現するラベル nsubj, obj, iobj などは格助詞「が」「を」「に」などの表層格情報に基いて付与する。例えば nsubj は格助詞「が」が先行している名詞句と述語との関係として割り当てる。同様に obj は格助詞「を」が先行している名詞句と述語との関係として割り当てる。このように、基本的には日本語の表層格と対応させながら依存関係として UD ラベルを割り当てていく予定である。

ただし図 3 のよう格助詞「に」の場合 iobj が 2 つ付与されてしまう可能性がある。この場合は BCCWJ-PAS の述語項構造関係による深層格情報を参照して、「学校」→「行く」の間の依存関係ラベルは iobj ではなく obl に変更する。このように一部の格情報については BCCWJ-PAS から述語項構造の情報を参照する。また、図 4 のように、主題を表す助詞「は」が現れた場合、nsubj としてラベルを付与する。

5 おわりに

本稿では日本語コーパスである現代日本語書き言葉均衡コーパス (BCCWJ) の文節依存構造を UD の体系へと変換したコーパスについて変換規則なども踏まえて紹介した。日本語と英語における統語関係の違いによりいくつか議論・検証するべき箇所はあるものの、現段階で決めた規則に基づき公開を予定している。

本稿執筆時点では、短単位に基づく UD の変換がほぼ完了し、長単位に基づく UD コーパスの実装へと移行している。また、長単位に基づく統語解析器も短単位に基づくデータ作業が終わり次第構築する予定である。構築後、短単位・長単位ベースの日本語 UD データを公開予定である。

謝辞

本研究の一部は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) および科研基盤 (A) によるものです。

参考文献

[1] Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. Universal dependencies version 2 for japanese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'2018)*, 2018. to appear.

[2] Masayuki Asahara and Yuji Matsumoto. BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58, 2016.

[3] Sadao Kurohashi and Makoto Nagao. *Building a Japanese Parsed Corpus – while Improving the Parsing System*, chapter 14, pp. 249–260. Kluwer Academic Publishers, 2003.

[4] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

[5] De Marneffe Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC'2014)*, pp. 4585–4592, 2014.

[6] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pp. 92–97, 2013.

[7] Shinsuke Mori, Hideki Ogura, and Tetsuro Sasada. A japanese word dependency corpus. In *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC'2014)*, pp. 753–758, 2014.

[8] Toshinobu Ogiso, Asuko Kondo, Yoko Mabuchi, and Noriko Hattori. Construction of the "Corpus of Historical Japanese: Meiji-Taisho Series I - Magazines". In *In Proceedings of Digital Humanities 2017 (DH2017)*, 2017.

[9] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pp. 2089–2096, 2012.

[10] Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. Universal dependencies for japanese. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'2016)*, pp. 1651–1658, 2016.

[11] Takaaki Tanaka and Masaaki Nagata. Constructing a practical constituent parser from a Japanese treebank with function labels. In *Proceedings of 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL'2013)*, pp. 108–118, 2013.

[12] 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, pp. 101–123, 2007.

[13] 植田禎子, 飯田龍, 浅原正幸, 松本裕治, 徳永健伸. 『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション. 第 8 回コーパス日本語学ワークショップ予稿集, pp. 205–214, 2015.

[14] 大村舞, 浅原正幸. 現代日本語書き言葉均衡コーパスの universal dependencies. 言語資源活用ワークショップ 2017, pp. 132–142, 2017.