Universal Dependencies における述語並列記述の展望

伊藤 薫
九州大学 言語文化研究院
ito@flc.kyushu-u.ac.jp

1 はじめに

1.1 Universal Dependencies とは

Universal Dependencies (UD) は通言語的に統一された方法で依存構造や品詞などについてアノテーションが付与されたコーパス群を開発する国際プロジェクトである [1]。UD は依存構造が内容語主辞で付与されていること、**Universal POS (UPOS)** と呼ばれる品詞体系や**依存関係タグ**が定義され、コーパスに付与されていることなどを特徴とする。また、人手アノテーションのしやすさやパージング精度の高さなど、言語学的妥当性以外も考慮して設計されている [2]。しかし、UD は進行中のプロジェクトであるため問題も多く抱えており、日本語 UD における並列表現もその 1 つである。本論では **Universal Dependencies (UD)** ver. 2.7 [3] における日本語の述語並列に関わる問題について、UD コーパスを用いた日英比較から得られたデータと、言語学で蓄積された並列に関する知見をもとに議論し、今後の展望について述べる。

1.2 UD における並列表現の定義

本節では、UD における並列表現について説明するため、公式サイト [3] に掲載されている依存関係タグに関する記述をもとに、解説を加えながら紹介する。UD において、並列表現は等位接続詞を表す品詞タグ CCONJ、等位項を表す関係タグ *conj*、及び、等位接続詞に掛かる関係タグ *cc* によって表現される。*conj* タグは等位接続詞で結合される等位項 (conjunct) であることを示すタグである。UD の方針として内容語を主辞としていることから、並列表現における最初の等位項が主辞とされ、それ以降の等位項は全て最初の等位項に依存すると定められている。最初の等位項とそれ以降の等位項は関係タグ *conj* で結ばれる。最後に、関係タグ *cc* は CCONJ とその主要部を結ぶ。UD では等位接続詞が省略される場合（e.g. 句読点類のみで並列が示される場合）にも並列構造とみなすことを認めており、この場合は CCONJ, *cc* がなく、*conj* のみで並列構造が表されることになる。

英語の並列表現を UD の体系で表す場合、CCONJ は ‘and’, ‘or’, ‘but’ など、語やそれより大きい構成要素を統語的に従属させることなく結び、両者の間の意味関係を表すものとされる。また、*conj* については UD 全体に向けた記述と大きな相違点は見受けられない。‘either’, ‘both’, ‘neither’ 等も CCONJ と見なされ *cc* の対象に含まれる点が英語に関して若干特徴的と言えるだろう [3]。

1.3 UD Japanese の並列アノテーション

前節で見たように、UD において関係タグ *conj* は最初の等位項を主辞として付与する、と定められているため、日本語のような右主辞の言語とは相性が悪い。日本語 UD では *conj* タグが強制的に左主辞となる仕様であることから生じる不自然さを回避するため、名詞句、動詞句の並列で *conj* タグを使用せずにコーパスが構築されている。例えば、「食べて走る」では「て」によって 2 つの動詞句が並列されているが、「て」を従属接続詞とみなし 2 つの動詞を副詞句による修飾 (*advcl*) としている [2]。

また、UD 2.7 における日本語コーパスで最大のものは UD Japanese-BCCWJ で、日本語 UD のデータのおよそ 4 分の 3 を占める。これは BCCWJ-DepPara という日本語書き言葉均衡コーパス (BCCWJ) のコアデータに対して追加アノテーションを施したものを変換することで作成されている。BCCWJ-DepPara では並列アノテーションの付与対象は名詞句の並列、部分並列とされており、日本語の節間の並列関係を従属関係と識別するのが困難なため述語並列は対象外とされている [4]。従って、今後上述の主辞に関わる問題が解決されたとしても、元コーパスに信頼のおけるアノテーションが施されている名詞句などの並列と比較して、他の日本語 UD における述

語並列のアノテーションはデータに反映されにくい状況が続くと考えられる。

2 述語並列の日英比較

前節では、UD における並列と日本語 UD における述語並列の記述に関する問題を概観した。多言語パラレルコーパス PUD treebank を用いて英語の並列表現が日本語 UD でどのように扱われるかを、翻訳と UD タグという観点から調査する。

2.1 PUD treebank

Parallel UD (PUD) treebank は 2017 年の記述によると 1,000 文、18 言語からなるパラレルコーパスである [1]。本論では英語と日本語の UD を対象としているため、PUD の英語版である **UD English-PUD** および日本語版である **UD Japanese-PUD** (以下、それぞれ **PUD-en**, **PUD-ja** と略記する) を使用した。1,000 文のうち 750 文は英語の文書から集められ、残りの 250 文は他の 4 言語からの翻訳である。なお、本論では多重に翻訳された文を除いた分、つまり、PUD-en において原文が英語である文 (文 ID が n01 または w01 で始まる文) を対象とした。収録されている文は Wikipedia とオンラインニュースから無作為に抽出されたものである。

2.2 手法

まず、PUD-en の英語を言文とする 750 文から動詞 (VERB) と動詞が関係タグ *conj* で結ばれている関係を全て抽出し、PUD-ja でどのような形式で翻訳されているかを調査した。[1)] PUD-en からデータを抽出する際には、オンラインで提供されているコーパスインターフェイス Grew[5] を使用した。抽出した文に関して収集した情報は文 ID、本文、係り元と係り先の表層形、そのうち等位接続詞を伴うものについては上記に加え CCONJ タグのついた語の表層形である。次に、抽出した英語文に対応する PUD-ja の文を文 ID によって検索し、目視で日本語訳を確認した。PUD-ja において注目したのは PUD-en における接続表現（主に CCONJ だが、句読点のみの場合なども含む）との対応であり、それらに対応する日本語の接続表現の表層形、及び、XPOS に記載されている国語研短単位品詞を集計した。また、活用のある品詞の場合は品詞に活用形を追記した。

並列される動詞句や節の構造が翻訳文でも保たれている場合は、(1) のように先に出現する動詞に直接依存する接続表現もしくは連用形[2)] を収集したが、原文とは大きく異なる構造になっている場合や、(2) のように 3 つ以上の等位項がある場合も見受けられた。(2) の場合は、*conj* 関係タグで結ばれる動詞は ‘works’ と ‘lives’、‘works’ と ‘breathes’ の 2 組だが、後者の組に対応する日本語文については、後方の等位項を含む節の直前に出現する接続表現を収集を試みた。つまり、‘works’ と ‘breathes’ に関わる接続表現を日本語文で探す際は、‘breathes’ に対応する「生きている」の直前にある接続表現「暮らし」（動詞-一般-連用形）を集計対象とした。[3)] また、適切な接続表現が見つからない場合は (3) のように「該当なし」[4)] とした。

(1) (a) A sacrococcygeal teratoma is a tumour that *develops* before birth **and** *grows* from a baby’s tailbone.
(b) 仙尾骨部のテラトーマは、赤ん坊の誕生前に発生**し**、尾てい骨から成長する腫瘍である。(ID: n01051006, 強調筆者)

(2) (a) It’s fair to say that Rocco Catalano *works*, lives **and** *breathes* retro.
(b) ロッコ・カタラーノ氏はレトロに働き、**暮らし**、生きていると言ってもよい。
(ID: n01087005, 強調筆者)

(3) (a) Isner, who produced some of his best tennis as he *leveled* at one set all **and** *forced* a decider, also paid his tribute.
(b) 最高のテニスでワンセットオールで対戦を余儀なくされた、イスナーもまた、賛辞を送っている。(ID: n01142007, 強調筆者)

1） 1.2 で述べたような理由により ver. 2.7 までの日本語 UD では *conj* タグが導入されていないため、逆方向の抽出、つまり、PUD-ja で *conj* とタグ付けされている動詞間の依存関係を抽出し、PUD-en でどのように翻訳されているかを調査することはできないため、本調査における翻訳の方向性には制約がかかっている。

2） 日本語 UD では、サ変動詞は語幹名詞+動詞「する」とされる。

3） この方針は後述する PUD-en における接続詞の集計方法とは異なるが、これは英語においては同じ語から *conj* タグで表示される 3 つ以上の並列としてコーパスで記述されているのに対し、PUD-ja では連体修飾関係を示す *advcl* タグで連鎖的にタグ付けされているという違いに基づく。

4） 「ワンセットオール」の後の「で」は格助詞とタグ付けされている。

表 1　翻訳文中で対応する接続表現の品詞別頻度

国語研短単位品詞	頻度 (%)
動詞-非自立可能-連用形	49 (34.5)
助詞-接続助詞	34 (23.9)
助動詞-連用形	16 (11.3)
動詞-一般-連用形	16 (11.3)
助詞-格助詞	8 (5.6)
接続詞	6 (4.2)
助詞-副助詞	3 (2.1)
(該当表現なし)	4 (2.8)
補助記号-読点	3 (2.1)
形容詞-非自立可能-連用形	2 (1.4)
助詞-終助詞	1 (0.70)
(連用形合計)	83 (58.5)

2.3　結果

PUD-en からは *conj* タグで結ばれる 175 組の動詞の組が抽出され、うち等位接続詞を伴う並列は 172 組であった。また、集計対象は *conj* タグの頻度であるため、3 つ以上の等位項が並列されている場合は「等位項 1 と等位項 2」、「等位項 1 と等位項 3」……のように主辞を基準として個別に計上した。収集された等位接続詞の集計結果を付録の表 2 に示す。なお、括弧で括られた等位接続詞は 3 つ以上の等位項を持ち、等位項と等位接続詞が隣接していないことを表す。また、表中に大文字で NONE と表記したデータでは等位接続詞が見られず、カンマのみで並列が表されていた。付録の表 2 に見られるように、英語の等位接続詞は 'and' が大半 (75%) を占め、次いで 'but', 'or'、更に少数の例としてカンマによる並列、'as' ('as well as' の一部) が確認された。

次に、PUD-ja において対応する接続表現の頻度を品詞ごとに表 1 に示す。対応する接続表現がない場合は表中に「(該当表現なし)」としてまとめた。表の末尾には品詞の区別なく活用形のみでまとめた場合の数値も付記した。表 1 に示した接続表現のうち、日本語 UD で CCONJ タグがついているものは「接続詞」のみであり、5 例[5] (3.5%) と稀である。また、品詞としてはサ変動詞の連用形が最も多くを占めており、品詞を問わず活用形のみで集計すると半数以上 (58.5%) を連用形による接続が占めている。連用形は一般的な日本語文法においては動詞の活用形の 1 つとされており並列表現の 1 形式でもある [6, 7] が、日本語 UD では *advcl* として主節に係る。

3　考察

3.1　日本語並列表現の多様性と UD

英語では表層形の種類が限られており、今回の調査では 'and', 'but', 'or', '(as well) as' の 4 種類のみであった。英語 UD の方針では 'both' や 'either' なども CCONJ に含まれるが、今回の結果には含まれなかった。一方、日本語では読点を除き 37 語が並列表現を含む英文の翻訳に使用されており、語幹ごとに別々に数えられる動詞の連用形を除いても 10 語[6] と、英語に比べ種類が多いのが特徴である。

また、現状の日本語 UD の仕様では並列表現が様々な UPoS に分散している。基本的に英語で CCONJ が出現する場合でも、日本語訳では CCONJ は出現せず、延べ語数で数えた場合に半数以上が連用形で表されることが明らかになった。また、連用形に次いで多い接続助詞は PUD-ja において SCONJ とされる。具体的には今回抽出された「し」「て」「が」は全て SCONJ だが、これらは理論によっては全て並列を担う表現とされうる。また、英語では CCONJ とされる 'or' に対応する表現は「か」だが、これらは PUD-ja において PART とされる。他にも、PART とされる「たり」の扱いにはそもそも日本語学の中でも様々な立場があり、少なくとも並列を表す接続助詞とする立場 [6]、タ系連用形とする立場 [7] があるが、PUD-ja のデータには国語研短単位品詞で助詞-副助詞とされているなど、基礎的な部分での困難もある。翻訳においても意味的な並列が保持されているのであれば、これらの扱いが日本語 UD における並列の記述方法に大きな影響を与えることになる。

3.2　日本語の並列表現と UD の体系

UD は世界中の個別言語を 1 つの体系で記述しようとするプロジェクトであるため、個々の言語への適用の際には問題が生じる場合もある。これらの問題は UD 以前に個別言語で整備されてきたデータの蓄積や UD への対応状況というローカルな事情によるものから、類型論的挑戦である UD の仕様そのものが抱える問題まで様々である。

今回観察したデータには接続助詞「か」が接続詞「あるいは」と同時に出現する場合もあり、この場合は英語のように「か」を CCONJ と見なすと UD 上では二重に CCONJ が出現することになる。これは英語の 'either ... or ...' に相当する。また、現在のとこ

5)　確認された表層形は「また」「しかし」「あるいは」「および」。

6)　「また」「しかし」「か」「が」「し」「たり」「て」「と」「(ため) に」「(もの) の」

ろ「て」「が」は日本語 UD において SCONJ としてタグ付けされ、SCONJ を伴う節と節を関係タグは多くの場合 *advcl* となり、少なくとも *conj* は不適である。加えて、上記 2 つよりも問題は少ないが連用形による接続の場合、品詞タグ VERB が付与される語の内部に並列を示す特徴が組み込まれていることになり、独立した CCONJ は出現しない。そのため、連用形による述語並列を日本語 UD において並列と認める場合、形式的には並列を表す表現があるものの、UPoS としては VERB であるため CCONJ を伴わず *conj* が出現する例が多くなる。

上述のように、日本語では連用形や助詞によって並列を表すことが多い。そのため、英語と親和性が高そうに思われる UD の基準からすると、日本語の並列表現は CCONJ を伴わない非典型的な形式として記述することになる場合が多くなると思われる。特に、連用形の場合は並列を表す形式が語に内在しているため、現在の UD の体系では表現が難しい。UD では語に内在する特徴 (Features) を記述することができ、動詞的な屈折特徴としては時制、法、相等が設定されている。Features には Verbform という動詞の形式を表すカテゴリがあり、日本語の連用形やテ形は Conv (converb, 副詞的な従属接続標示することを主な機能とする不定動詞 [8]) であり、日本語の連用形 [9] やテ形 [8][9] を converb としている文献も存在する。ただし、この方法の場合は統語的に従属となることを示すに留まり、意味的に並列であることを示すことはできない。

3.3 並列表現の認定基準

英語では *conj* タグが出現する例において等位接続詞が出現する場合が多く、カンマのみによる並列が少ないことを 2.3 で明らかにした。しかし、日本語の場合は述語並列の形式が英語に比べて多様であり、並列と従属の線引きも難しく、何を以て並列とするかも理論によって異なる。例えば、節の並列で言えば「主節に対して対等に並ぶ関係で結びつく節」を「並列節」とする統語面を重視する定義 [7] や、「2 つ以上の異なる自体の時間的前後関係が解釈の際に問題にならない時、その事態は並列関係にある」という意味的側面を重視する定義 [6] がある。また、BCCWJ-DepPara が節の分類の際に依拠している文法書 [10] では接続表現の形式ごとに従属節内部に出現可能な要素を検討して 3 つに分類しており、並列は従属節が表す意味の一部に組み込まれている。

また、形式のみで等位接続と従属接続が明確に分かれると思われる英語においても、同様の問題は存在する。例えば、‘You drink another can of beer and I’m leaving.” という例文では、前半の節が後半の節の条件として解釈でき、統語的には並列だが意味的には従属と見なせるとされる [11]。また、この議論を発展させて並列と従属は統語と意味の両面から見るべきだという提案もあり、上記 ‘and’ の例を pseud-coordination、日本語の「て」による並列を pseudo-subordination（統語的には従属、意味的には並列）とする研究もある [12]。並列は多面的な現象であり、UD の体系の中で日本語の並列を記述するためには並列のどのような側面を重視して認定するかを考慮する必要がある。

4 終わりに

本論では日本語 UD における述語並列の問題について議論し、現在の仕様では英語の並列表現が日本語訳の大半において従属節となっており、連用形や一部の接続助詞の扱いが重要であることを示した。この状態は pseudo-subordination という観点から、現在の日本語 UD の仕様は統語面を重視した仕様になっていると評価できる。Croft[13] は information packaging を依存関係、意味内容を語彙のタグセットと組み合わせるべきであると主張しており、依存関係を重視すれば日本語では *conj* を使う機会は減ると予測されるが、3.2 で示した通り現在の仕様では直接的に語彙のタグセットで並列であることを示す手段はない。逆に機能的側面を重視するならば、英語の ‘and’ が pseudo-coordination となる例もあるため、典型的に意味的な並列を担う語彙項目に関して、現在 *advcl*, SCONJ となっている一部の並列表現形式を *conj*, CCONJ へ変更するのも一案である。

いずれにしても、UD においては日本語の述語並列について並列と見なすのであれ従属と見なすのであれ、何らかの方針に基づきタグ付けをすることになる。認定基準は典型的な用法、統語テスト、機能面で見た並列用法と従属用法の比率など様々に考えられるが、UD 全体との整合性やコンバージョンの容易さなどを考慮して決定する必要がある。

謝辞 本研究の一部は JSPS 科研費 19K13180 の助成および国立国語研究所コーパス開発センター共同研究プロジェクトの支援を受けたものです。

参考文献

[1]Anders Björkelund, Agnieszka Falenska, Xiang Yu, and Jonas Kuhn. Ims at the conll 2017 ud shared task: Crfs and perceptrons meet neural networks. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 40–51, 2017.

[2]浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇有吾, 松本裕治. Universal dependencies 日本語コーパス. 自然言語処理, Vol. 26, No. 1, pp. 3–36, 2019.

[3]Universal Dependencies contributors. Universal dependencies, 2020-10 閲覧. https://universaldependencies.org/.

[4]浅原正幸, 松本裕治. 『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション. 自然言語処理, Vol. 25, No. 4, pp. 331–356, 2018.

[5]Bruno Guillaume. Grew— graph rewriting for nlp, 2020-12 閲覧. https://grew.fr/.

[6]中俣尚己. 日本語並列表現の体系. ひつじ書房, 東京, 2015.

[7]益岡隆志, 田窪行則. 基礎日本語文法—改訂版— . くろしお出版, 東京, 1992.

[8]Martin Haspelmath. The converb as a cross-linguistically valid category. In Martin Haspelmath and Ekkehard König, editors, *Converbs in Cross-linguistic Perspective*, pp. 1–55. Mouton de Gruyter, Berlin, 01 1995.

[9]清瀬義三郎則府. 連結子音と連結母音と–日本語動詞無活用論. 国語学, Vol. 86, pp. 42–56, 1971.

[10]南不二男. 現代日本語の構造. 大修館書店, 東京, 1974.

[11]Peter W. Culicover and Ray Jackendoff. Semantic subordination despite syntactic coordination. *Linguistic Inquiry*, Vol. 28, No. 2, pp. 195–217, 1997.

[12]Etsuyo Yuasa and Jerry M. Sadock. Pseudo-subordination: A mismatch between syntax and semantics. *Journal of Linguistics*, Vol. 38, No. 1, pp. 87–111, 2002.

[13]W. Croft, D. Nordquist, Katherine Looney, and Michael Regan. Linguistic typology meets universal dependencies. In *TLT 2017*, 2017.

A 付録

表 2 PUD-en で収集された CCONJ の形式と頻度

等位接続詞	頻度 (%)
‘and’	132 (75.43)
‘but’	28 (16.00)
‘or’	7 (4.00)
(‘and’)	3 (1.71)
NONE	3 (1.71)
‘as’	1 (0.57)
(‘as’)	1 (0.57)