一般論文

Peer-Reviewed

国語研長単位に基づく日本語 Universal Dependencies

大村　舞[†]・若狭　絢[†]・浅原　正幸[†,††]

Universal Dependencies (UD) は言語横断的に単語の依存構造に基づくツリーバンクを構築するプロジェクトである．全言語で統一した基準により，品詞・依存構造アノテーションデータの構築が 100 言語以上の言語について進められている．分かち書きをしない言語においては，基本単位となる**構文的な語 (syntactic word)** を規定する必要がある．従前の日本語の UD データは，形態論に基づく単位である国語研短単位を採用していた．今回，我々は新たに構文的な語に近い単語単位である**国語研長単位**に基づく日本語 UD である **UD_Japanese-GSDLUW**，**UD_Japanese-PUDLUW**，**UD_Japanese-BCCWJLUW** を構築したので報告する．

キーワード：係り受け構造，分かち書き，Universal Dependencies, 日本語，長単位

Universal Dependencies for Japanese Based on Long-Unit Words by NINJAL

Mai Omura[†], Aya Wakasa[†] and Masayuki Asahara[†,††]

Universal dependencies (UD) are part of an international project that aims to construct cross-lingual dependency treebanks. The consistent annotation standards of grammar (parts of speech, morphological features, and syntactic dependencies) are defined across different languages and compiled as treebanks of more than 100 languages. The languages written without word delimitation must define the word units of their **syntactic words** on the UD guideline. The preceding UD Japanese resources are based on the short-unit words by NINJAL, which is defined by their lexicon-based morphology. This study introduces UD Japanese resources **UD_Japanese_BCCWJ-GSDLUW**, **UD_Japanese_PUDLUW**, and **UD_Japanese_BCCWJLUW** based on the long-unit words by NINJAL, which are more suitable as syntactic words than NINJAL's short-unit words in Japanese.

Key Words: *Dependency Structure, Word Delimitation, Universal Dependencies, Japanese, Long Unit Word*

† 人間文化研究機構国立国語研究所, National Institute for Japanese Language and Linguistics, Japan
†† 東京外国語大学, Tokyo University of Foreign Studies

本論文は，Universal Dependencies Workshop 2021 "Word Delimitation Issues in UD Japanese" Omura, Wakasa and Asahara (c) Association for Computational Linguistics (CC BY 4.0) 2021 年 12 月 および言語処理学会第 28 回年次大会『国語研長単位に基づく UD Japanese』 (c) 大村・若狭・浅原 (CC BY 4.0) 2022 年 3 月 の発表内容に基づき言語資源の継続的な修正および比較実験の再試行を行ったものである．

(C) The Association for Natural Language Processing. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1　はじめに

Universal Dependencies (UD) (Nivre et al. 2016) は，言語横断的に品詞・形態論情報・依存構造をアノテーションする枠組およびコーパスである．UD プロジェクトの研究目標として，多言語の統語解析器開発，言語横断的な言語処理技術の開発，さらには類型論的な言語分析 (de Marneffe et al. 2021) などがあげられている．UD では，データ構造やアノテーション作業を単純化するため，またくだけた文や特殊な構造に対して頑健な表現を実現するために，句構造 (phrase structure) ではなく，図 1 のような語の間の依存関係と依存関係ラベルで表現する依存構造を採用している．UD のガイドラインを基に，現代語のみならず，古語・消滅危機言語・クレオール・手話などを含めた 100 言語以上の依存構造アノテーションデータが構築され，公開されている[1]．2022 年 8 月現在でも，言語横断性を高めるために UD の基準について活発に GitHub[2]上やワークショップで議論され，ラベルの統廃合が行われながらもアノテーションやガイドラインが更新し続けられている．

この UD の枠組では，依存構造関係を付与する基本単位として，音韻的な単位や文字・形態素ではない**構文的な語 (syntactic word)** を語として用いることを規定している．英語やフランス語といった空白を用いて分かち書きをする言語においては（縮約形態などを除いて）空白を語の単位認定として用いることが多い．一方，語の境界を空白などで明示しない東アジアの言語においては，どのような単位を構文的な語 に規定すべきかという問題があり，これらの言語では，一度語の基本単位を定義してから，UD を構築している．現代中国語 (Xia 2000; Leung et al. 2016) や韓国語の UD (Chun et al. 2018)，トルコ語・古チュルク語 (Kayadelen et al. 2020; Derin and Harada 2021) などでも，言語ごとにコーパスや形態素解析などによって語の単位認

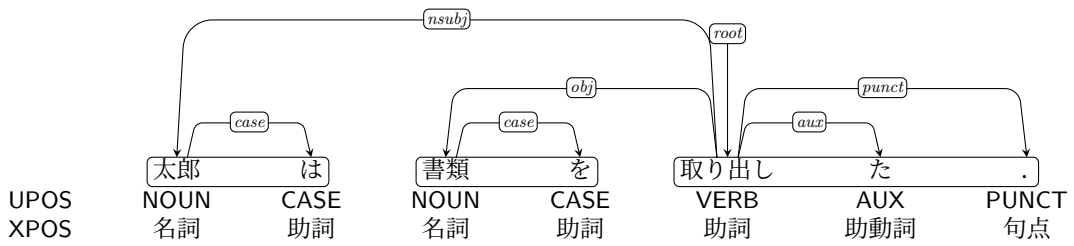


図 1　日本語 UD の例.「文節係り受け構造」で採用されている単位「文節」(枠で囲んである単位）とは異なり「自立語（内容語)」と「付属語（機能語)」を分解した単語単位を UD では想定する． UPOS が UD の定義する品詞, XPOS は言語依存の品詞（日本語 UD では Unidic 品詞. 図の例では品詞の詳細を略するときがある．）

[1] https://universaldependencies.org/
[2] https://github.com/UniversalDependencies/docs/issues

定を行い，UD の言語資源が構築されている．

UD Japanese（日本語 UD）Version 2.6 以降では，その基本単位として**国語研短単位**（Short Unit Word, SUW: 以下 **短単位**）を採用している (浅原 他 2019)．短単位 は『現代日本語書き言葉均衡コーパス』(BCCWJ) (Maekawa et al. 2014)・『日本語日常会話コーパス』(CEJC) (Koiso et al. 2022) をはじめとした形態論情報つきコーパスでも単位として採用されている．短単位に基づく形態素解析用辞書として，約 97 万語からなる UniDic (伝 他 2007) も公開されている．また，170 万語規模の単語埋め込み NWJC2vec (Asahara 2018) でも短単位が使われており，短単位を基準として言語処理に必要な基本的な言語資源が多く整備されている． この短単位に基づく言語資源の豊富さから， 実用上は短単位に基づく処理が好まれる傾向にあった．しかし，グレゴリー・プリングルによるブログ記事[3]や Murawaki (2019) では，単位として短単位を採用している既存の UD Japanese コーパスは「形態素」単位であり，UD の原則にあげられる「基本単位を構文的な語とする」という点において不適切であることを指摘している．

国語研においては，形態論情報に基づいて単位認定し，「可能性に基づく品詞体系」が付与されている短単位とは別に，文節に基づいて単位認定し，「用法に基づく品詞体系」が付与されている**国語研長単位**（Long Unit Word, LUW: 以下 **長単位**）を規定している．しかし，長単位に基づくコーパスの構築は， 短単位に基づくコーパスの構築より長時間の作業を要する[4]という問題がある．言語資源としては， BCCWJ や CEJC には長単位に基づいた形態論情報が付与されているとはいえ，短単位と比べると利用可能な言語資源やツールが少ないため，長単位に基づく依存構造が解析器によって生成できるのかという問題もある．

日本語における語の単位認定の検証のためには，実際に短単位のみではなく，長単位に基づく 日本語 UD 言語資源を整備することが必要である．本研究では，長単位に基づく日本語 UD の言語資源を整備したので報告する．UD 全体と日本語における単位認定について説明しながら，既存の言語資源・解析器によって長単位に基づく日本語 UD の構造が生成しやすいかを短単位 UD と比較して検討する．

2　Universal Dependencies と日本語の単位認定

2.1　Universal Dependencies による単位認定

Universal Dependencies (UD) (Nivre et al. 2016) は，言語横断的に品詞，形態論情報，依存構造を付与するための枠組およびコーパスである．UD の枠組として，品詞体系 (UPOS) は Google Universal Part-of-speech Tags (Petrov et al. 2012) を， 形態論情報 (FEATS) は Interset

[3] http://www.cjvlang.com/Spicks/udjapanese.html
[4] これは, 短単位と比較すると, 自動解析の精度が担保されておらず, 長単位のアノテーション修正の作業ができる人材も少ないなどといった理由が挙げられる．

interlingua for morphosyntactic tagsets (Zeman 2008) を，依存関係ラベル (DEPREL) は Universal Stanford Dependencies (de Marneffe et al. 2014) を基に構成されている．さらに UD では，多言語横断の枠組の実現のために，すべての構文構造を図 1 のように語の間の依存関係と依存関係ラベルで表現する．そのため，依存関係を表すための**「語 (words)」の単位認定**が重要になる．

UD のガイドラインでは，この「語」の単位について，語同士は依存関係が成立しているという前提でアノーテションすることを求めている．さらに，「語は**構文的な語 (syntactic word)** である」と定義しており，構文的な語は形態素ではないと説明している．形態論的な特徴は語の属性としてアノテーションし，語を形態素に分割しない．これは，音韻論上他の語に依存する拘束形態素であっても統語的に独立した語（スペイン語の “dámelo” → “da me lo”）や縮約形態 (“au” → “à le”) を元の個別の単位に分解したいという理由に基づく[5]．このような UD の規定する語を本稿では「構文的な語」として説明する．語の分割が空白により規定できる言語については，より細かく分割すべき接語 (clitics) やまとめあげるべき数的表現・略語 (“20 000” や “e. g.”) に対して適切なドキュメンテーションを行うことが求められている．

日本語や一部の東アジア諸言語では，空白を用いずに記述するため，明示的な空白を用いた単位認定が困難である．UD のアノテーションを作成するうえで単位認定は非常に重要であるため，言語またはツリーバンクごとにアノテーションする単位を規定している．たとえば，日本語と同様に空白による境界のない現代中国語 (Xia 2000; Leung et al. 2016) や韓国語の UD (Chun et al. 2018) では，既存の単語分割コーパスに基づいて単位を規定している．古チュルク語を含むトルコ語 (Kayadelen et al. 2020; Derin and Harada 2021) の UD は，語の境界を表す空白は存在するものの，日本語と同じく膠着語であり，動詞などは接尾辞を付着させて文法関係を示している．そのため空白だけではなく，さらに UD の示す構文的な語に合うように接尾辞などを区切り，単位認定を行っている．

2.2　日本語の単位認定

日本語は，形態論的変化が豊富でありながら，語境界に空白を使わない言語である．日本語における単位は分析手法に応じて定義される傾向にあり，自明な語境界があるわけではない．また，日本語の単位認定は工学的にも重要と考えられており，形態素解析の研究が自然言語処理の分野で盛んに行われていた．日本語の形態素解析では，内部的に辞書を用いて解析が行われることがほとんどであり，さまざまな種類の辞書が構築されて公開されている．日本語の形態素解析用辞書としては，IPADIC 辞書[6], JUMAN 辞書[7]，Sudachi 辞書 (Takaoka et al. 2018; 坂本 他 2018)，IPA 辞書から新語を拡張した辞書 mecab-ipadic-NEologd (佐藤 他 2017) なども公

[5] https://universaldependencies.org/docs/u/overview/tokenization.html
[6] https://ja.osdn.net/projects/ipadic/
[7] https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN

開されている．

日本語 UD は，他言語 UD と同様に，既存の辞書やコーパスを利用して UD の形式に自動変換することを目指している．語境界を新たに認定するよりも，既存の単位に基づいて品詞などの対応を取るほうが UD の基準の変更に追随するためにも望ましい．日本語 UD では，Version 2.6 より，UniDic (伝 他 2007) の語彙項目の単位，すなわち**国語研短単位（短単位）**を単位とする方針とした (浅原 他 2019)．本稿で報告する長単位 UD の**国語研長単位（長単位）**も，短単位と同様に設計されたものである．次節では，短単位と長単位といった国語研による語の単位認定について説明をする．

2.3　国語研による語の単位認定

国語研短単位や国語研長単位は，語彙調査のために制定された単位である (国立国語研究所 2006)．図 2 で示す通り，短単位は最小単位から，長単位は文節から規定されている．この節では最小単位，短単位，文節，長単位の説明をする．

国立国語研究所では，現代語において意味を持つ最小の単位のことを**最小単位**と定義する．最小単位はその語種に基づき定義される．日本語は，漢語・和語・外来語・固有名詞・数値表現・記号などの語種からなる．漢語は 1 文字 1 語とし，和語・外来語・固有名詞はそのもっとも短い単位を 1 語とする．数値表現は十進の位取りで発音できる単位に分割する．記号は 1 文字 1 語とする．このような操作的な規則に則り，最小単位を制定している．

短単位は，この最小単位に基づき，同じ語種同士の 1 回結合までを 1 単位として定義する．図 2 の例では，「学」と「年」はそれぞれ漢語であり，結合した「学年」を 1 短単位とする．和語である「用い」と付属語である「られ」を結合しないで，それぞれを 1 短単位とする．短単位は，基準が分かりやすく，ゆれが少ないという特徴があり，頻度の計数にふさわしい単位とされている．この短単位に対して UniDic 体系に基づく形態論情報（品詞・活用型・活用形・語

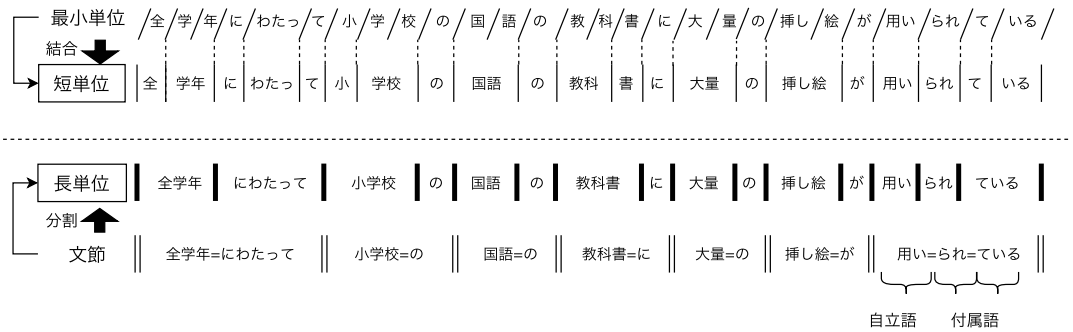


図 2　最小単位，短単位，長単位，文節の例（BCCWJ の PB33_00032 より）．2.3 節でも説明する通り，最小単位から短単位，文節から長単位が規定されている．

形・語彙素）が定義される．短単位に付与される UniDic 品詞は「可能性に基づく品詞」であり，その語彙素（表層形）がなりうるすべての品詞用法を考慮したものである．たとえば，「日曜」などの単位は文脈によって「名詞-一般」にも「副詞-一般」にもなりえるため，「名詞-副詞可能」といった品詞を定義する．短単位は語彙分析を目的としており，文脈に依存した用法を考慮した品詞を定義していない．

一方**長単位**は，まず**文節**境界を認定したのちに，文節内の短単位要素の結合により認定する単語単位である．文節は日本語における係り受け構造（依存構造）情報における基本単位に適した境界として採用されており，この文節を基に係り受け構造の整備が行われている (浅原，松本 2018)．日本語の文節単位の係り受け構造は，係り受けが交差せず，倒置などをのぞいて右主辞であり，簡単なアルゴリズムで組み上げられるという性質を持つ．しかしながら，文節自体には品詞が設定されておらず，文節間同士の依存関係も基本的に定義されていない．そのため，文節のみでは，直接 UD に変換するための情報が本質的に欠けている．文節は，１つの自立語と接頭辞・接尾辞・助詞・助動詞などの複数の付属語から構成されるとし，その各要素を長単位として認定する．図 2 の文節「用いられている」は，長単位自立語と「用い」と長単位付属語「られ」「ている」により構成される．長単位の規程集 (小椋 他 2008) では，長単位としてみなす複合辞 (Multi-word Expressions) が認定されており，複合辞をなす機能表現は 1 単位とする．また長単位には，文脈によりどのような統語的なふるまいかについて曖昧性解消を行った「用法に基づく品詞」が付与される．長単位は，文節を基に「用法に基づく品詞」が付与可能な自立語構成素と付属語への分割を行った単位とも言える．

短単位と比較すると，長単位は依存構造の基本単位である文節を基準とした単位のため，構文的な語に近いと考えられる．また，たとえば，短単位における品詞「名詞-副詞可能」は，長単位においては文脈に基づき「名詞-一般」もしくは「副詞-一般」のいずれの用法なのかを判別して付与するため，統語上における品詞の曖昧性も解消できている可能性が高い．さらに，短単位は前述のとおり，形態素に基いているのみで，膠着語としての性質を考慮していない．そのため，接尾の語により品詞が変化する場合に対応できないものがある．これは浅原 他 (2019) でも問題としてあげられている点である．これは長単位の区切り方で解消することができると考えられる．

図 3 に短単位と長単位の違いの 1 例をあげる．上が短単位，下が長単位の例である．図 3 から分かるように短単位は「力強さ」を「力強」「さ」，「両立する」が「両立」「する」の 2 つに認定されており，長単位ではそれぞれ「力強さ」「両立する」という 1 つの長単位として認定されている．接尾辞の「さ」は前にくる形容詞を名詞化するため，統語的には「力強さ」で名詞と品詞認定されるべきだが，短単位では実現できていない．長単位では「力強さ」を 1 長単位とし，「名詞」と認定されている．また，短単位では「両立」と「する」で分かれてしまっているため，「する」によって動詞化されていることを確認する必要がでてくる．実際 **UD_Japanese-BCCWJ**

では「する」が名詞に接続されているかをプログラムで抽出して判定した後，UPOS の VERB を付与するような変換規則を設けている (Omura and Asahara 2018)．一方で，「両立する」として 1 つの長単位として認定すれば，長単位の品詞体系から, 自然と動詞として品詞認定できる．このように長単位は文脈上の品詞の曖昧性を解消しているため，短単位と比べて, UD の示す構文的な語に近いと考えられる．

2.4　日本語の単位認定について利用できる言語資源・解析器

統語解析では，統計的機械学習や深層学習により自動解析器を構築するものが多く，そのため，学習時に利用する言語資源の豊かさは重要な点である．前述した最小単位・短単位・長単位・文節について，表 1 に各単位認定で利用できる言語資源・解析器について示す．表中 TTR は BCCWJ 中の総語数に対する異なり語数の比率 (Type Token Ratio) であり，この値が高いほ

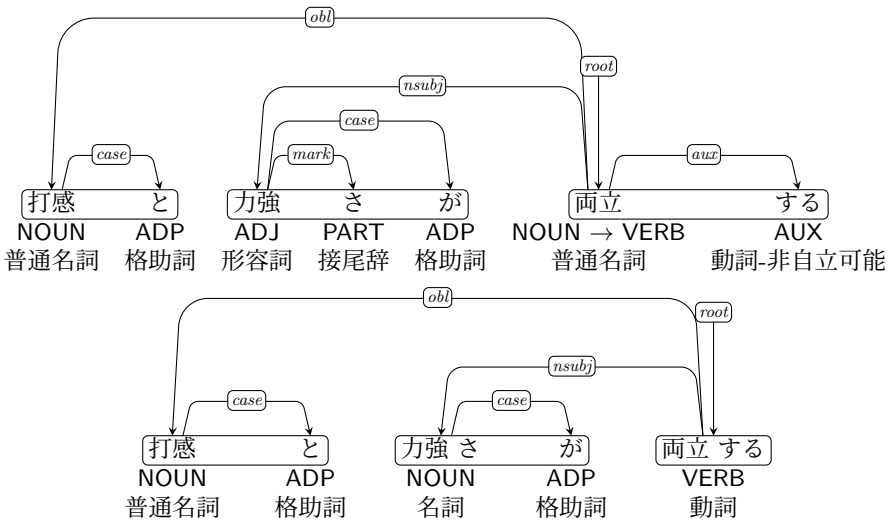


図 3　例文「打感と力強さが両立する」. 上が短単位の例, 下が長単位の例である．（BCCWJ の PM41_00172 より一部抜粋） 短単位の場合,「両立」だけではなく,「する」まで参照してから「VERB」を付与する規則を設けている．

	辞書	単語分割器	単語埋め込み	TTR
最小単位	N/A	N/A	N/A	N/A
短単位	UniDic	MeCab	NWJC2vec	0.00176
長単位	N/A	Comainu/Vaporetto	N/A	0.02922
文節	N/A	CaboCha/Comainu	N/A	0.22221

表 1　各単位認定で利用できる言語資源・解析器．N/A は存在しない，あるいは計算できないことを示している．

ど，その単位の出現確率が小さくなることを意味する．

最小単位は，現在のところ公開されている言語資源・解析器は存在しない．短単位は，形態素解析器 MeCab (Kudo et al. 2004) と形態素解析用辞書 UniDic を用いることにより生成される．ほかにも Asahara (2018) によって公開された NWJC2vec などの単語埋め込みが利用できる．

長単位・文節については，TTR の大きさからも分かる通り，語彙が膨大であるため辞書は存在しない．しかし, 長単位は, 短単位から構成規則により生成が可能なため，中・長単位解析器 Comainu (小澤 他 2014) や Vaporetto (赤部 他 2022) といった解析器などにより生成することができる． 文節は, CaboCha (工藤，松本 2002) などのツールでも解析ができる．TTR が大きいほど単語埋め込みの構築の難易度も高くなることが Asahara (2018) によって指摘されており，長単位や文節について単語埋め込みは構築されていない．

2.5　これまでの日本語 Universal Dependencies の歴史

2023 年 1 月までに，現代日本語 UD リソースとして 7 種類のコーパスが公開されている．これまで公開された日本語 UD を表 2 に示す．

UD_Japanese-KTC (Tanaka et al. 2016) は京都大学テキストコーパスを元に作られた UD コーパスである．**UD_Japanese-KTC** は長単位に近い単語境界で単語を構成しており，人手によって句構造木が付与され，その句構造木に基づいて UD の依存構造木を構築している．現在 UD プロジェクトでは Version 2 が公開されているが，**UD_Japanese-KTC** については Version 1 でメンテナンスが止まっている．

UD_Japanese-BCCWJ (Omura and Asahara 2018) は BCCWJ に基づいた UD コーパスである．Version 2.2 より公開された．BCCWJ では短単位・長単位・文節および文節単位の係り受け構造の情報が提供されている．**UD_Japanese-BCCWJ** はこの BCCWJ からの形態素情報および係り受け構造を元に Omura and Asahara (2018) らが提案している規則によって，短単位に基づく依存構造木へと自動変換されたものである．

UD_Japanese-GSD と **UD_Japanese-PUD** は Version 1.4 まで Google (McDonald et al.

名称	語数	単語単位	Version	Copyright	文の種類
UD Japanese-KTC	189k	LUW ライク	–v1	内容分離	新聞
UD Japanese-GSD	186k	SUW	v2.0–	CC-BY-SA	Wikipedia
UD Japanese-PUD	26k	SUW	v2.0–	CC-BY-SA	Wikipedia 対訳
UD Japanese-BCCWJ	1,098k	SUW	v2.2–	内容分離	新聞・雑誌・白書・ブログ等
UD Japanese-GSDLUW	186k	LUW	v2.9–	CC-BY-SA	Wikipedia
UD Japanese-PUDLUW	26k	LUW	v2.9–	CC-BY-SA	Wikipedia 対訳
UD Japanese-BCCWJLUW	1,098k	LUW	v2.9–	内容分離	新聞・雑誌・白書・ブログ等

表 2　公開現代日本語 UD のコーパス. SUW は短単位, LUW は長単位の略称である．

2013) が管理し，Version 2.0 から 2.5 までは IBM の単語分割器 (Kanayama et al. 2000) により単語分割し，修正したものであった．Version 2.6 より国立国語研究所がメンテナンスおよび公開しているコーパスである．あらかじめ **UD_Japanese-GSD** の Version 1.4 の権利者の許諾を得て，さらに交渉によりライセンスを CC BY-NC-SA から CC BY-SA に変更したうえで，Version 1.4 データの例文を元に，国立国語研究所にて国語研短単位形態論情報と文節係り受け情報を人手により付与している．UD の依存構造木については，文節係り受け木から Omura and Asahara (2018) らの規則によって **UD_Japanese-BCCWJ** と同様に変換したものである．

今回，さらに **UD_Japanese-GSD**, **UD_Japanese-PUD**, **UD_Japanese-BCCWJ** の例文に基づき，長単位に基づく日本語 UD コーパスを構築した．BCCWJ には長単位形態論情報が付与されていたが，新たに **UD_Japanese-GSD** と **UD_Japanese-PUD** にも長単位形態論情報を付与した．次節では，長単位に基づく日本語 UD コーパスの構築について紹介する．

3 長単位に基づく日本語 Universal Dependencies

UD の理念に即した単位に基づく日本語の言語資源を構築すべく，我々は長単位を単位認定とした日本語 UD コーパスを開発した．**UD_Japanese-GSD**, **UD_Japanese-PUD** に対して，新たに国語研長単位形態論情報を付与し，UD の Version 2.9（2021 年 11 月）から，BCCWJ, GSD, PUD の長単位に基づく 日本語 UD を，**UD_Japanese-BCCWJLUW**[8], **UD_Japanese-GSDLUW**[9], **UD_Japanese-PUDLUW**[10] として公開した．2023 年 1 月現在 Version 2.11 が最新版である．

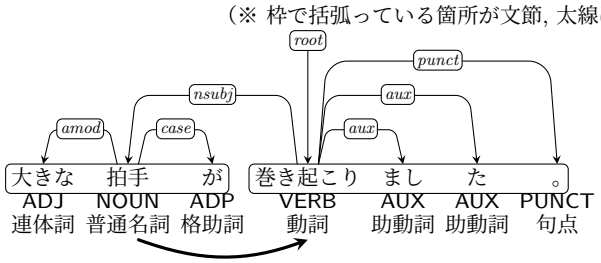
前節で説明したとおり，**UD_Japanese-GSD** と **UD_Japanese-PUD** は，Version 2.6 公開時に国語研長単位に基づく形態論情報（単語境界・品詞・語彙素）と文節に基づく係り受け構造を人手により整備した． さらに，国語研長単位に基づく形態論情報（単語境界・品詞・語彙素）の整備を継続して進めてきた． 長単位整備時に，長単位と短単位でデータに一貫性の不備があった場合は，短単位形態論情報もさらに修正した．そしてこの国語研短単位・長単位形態論情報が付与された文節係り受けデータ GSD, PUD を Omura and Asahara (2018) の提案した規則によって，文節係り受け構造から，長単位に基づく UD 基準の依存構造コーパスに変換し，**UD_Japanese-GSDLUW** と **UD_Japanese-PUDLUW** を構築した．**UD_Japanese-GSD** と **UD_Japanese-PUD** の元となっているこの文節係り受けコーパスは拡張 CaboCha 形式 (松吉 他 2014) ファイルのオープンデータとして公開[11] されている．

[8] https://github.com/UniversalDependencies/UD_Japanese-BCCWJLUW/
[9] https://github.com/UniversalDependencies/UD_Japanese-GSDLUW/
[10] https://github.com/UniversalDependencies/UD_Japanese-PUDLUW/
[11] https://github.com/udjapanese/UD-Japanese-GSDPUD-Cabocha

BCCWJ は短単位の境界と品詞，長単位の境界と品詞が整備されており，浅原，松本 (2018) のデータを組み合わせることで文節係り受けデータも獲得できる． そこから GSD と PUD と同様に Omura and Asahara (2018) の規則によって変換を行い **UD_Japanese-BCCWJLUW** を構築した．

この Omura and Asahara (2018) らの変換規則は短単位と長単位の語彙素や品詞，形態的な特徴に基づいたものである． 図 4 に Omura and Asahara (2018) らの変換規則の一部[12]と例を掲載している．例文にアノーテションされた短単位と長単位の語彙素や品詞・係り受け構造から，単語間依存構造と情報を抽出し，図 4 の下部の表で示しているような規則を単語ごとに適応している．適応結果から得られた UPOS および DEPREL のラベル付与することで UD の依存構造を生成している．Omura and Asahara (2018) らの変換規則は，Unidic 品詞だけではなく，文節や係り受け構造も参照しつつ変換されている．日本語における UPOS と DEPREL についての説明は文献 (浅原 他 2019) を参照すること．

Omura and Asahara (2018) の時点では， 変換規則は短単位のみの変換を想定していた．長単位 UD コーパス構築に際して，長単位のための追加の規則が必要か検討したものの，既存の変換規則が長単位の規則を内包していたことが分かった[13]．そのため，長単位 UD でも，Version



変換規則	UPOS
依存構造の親から順に適応している	
...	...
品詞が「句読点」である	PUNCT
...	...
品詞が「連体詞」である	ADJ
品詞が「名詞」である	NOUN
...	...
品詞が「助詞-格助詞」である	ADP
品詞が「^動詞-.*」｜文節の末尾	VERB
...	...

変換規則	DEPREL
「係り先の子」を対象の語として規則を適応している	
文節の末尾である｜文節の主辞である	root
UPOS が「PUNCT」	punct
...	...
文節の主辞｜体言である｜助詞「が」が付属	nsubj
...	...
文節が機能語ではない｜UPOS が「ADJ」	amod
UPOS が「ADP」	case
UPOS が「AUX」	aux
...	...

図 4　係り受け構造から UD 構造への変換規則の例．上記は規則の一部でかつ説明を簡略化したものである．（上記の例は UD Japanese-GSDLUW 中の dev-s63)

[12] 詳細な規則の一覧は `https://udjapanese.github.io/UD_converion_table/index.html` に記載している．
[13] これは短単位に基づく日本語 UD であっても， 短単位品詞が「可能性に基づく品詞」の場合， 実際は長単位品詞を用いて用法の曖昧性解消を行っていたためである．

2.10 時点で，そのまま数点の規則の変更のみ[14]で， 長単位でも同じように適応している．

表 3 に構築した長単位 UD の文数・文節・単語の統計情報を示している． 比較のために短単位 UD の統計情報も掲載している．長単位は 1 語以上の短単位により構成されているため，表 3 から分かるように，長単位の語数は短単位の語数より少ない．しかしながら， ほとんどの語において短単位と長単位は 1 対 1 対応している．BCCWJ においては 81.3%が， GSD においては 79.0% が，短単位と長単位が同一単位となっている．

表 4 と表 5 には日本語 UD の UPOS と DEPREL のコーパスごとの頻度割合を示している． 短単位と長単位を比較すると，表 4 から，長単位の UPOS は PROPN, SCONJ, SYM や NOUN の割合が減少していることが分かる．表 5 から，長単位の DEPREL は *compound, fixed, nummod,mark* などの割合が減少しているのが分かる． いずれも複合名詞や複合動詞でよく用いられていたアノテーションであり， 長単位で 1 単位となった際に消える関係のため減少する．一方で長単位の AUX や ADP，*aux*，*case* などの割合が増えているのは，短単位が結合して 1 長単位が格助詞や助動詞といった付属語に変化したためである．たとえば, 図 5 の事例で説明すると，「吉田」「あや子」は長単位の場合, 複合名詞「吉田あや子」と変化するため，「吉田」「あや子」を結んでいた *compound* が消えている．また長単位「ている」は短単位の場合「て」「いる」の 2 単語で *mark* と *fixed* を用いているが, 長単位「ている」と結合されたことで「助動詞」となるため, UPOS が AUX に変化し, *aux* が付与される．

データ		単語単位	文	文節	語
BCCWJ	train	SUW	40,801	308,679	908,738
		LUW	—	308,648	715,759
	dev	SUW	8,427	60,722	178,306
		LUW	—	60,697	145,398
	test	SUW	7,881	56,350	166,859
		LUW	—	56,332	134,475
GSD	train	SUW	7,050	57,357	168,333
		LUW	—	57,174	130,298
	dev	SUW	507	4,203	12,287
		LUW	—	4,186	9,531
	test	SUW	543	4,588	10,429
		LUW	—	4,568	13,034
PUD	test	SUW	1,000	10,008	28,788
		LUW	—	9,971	22,910

表 3　日本語長単位ベース UD の統計情報．（文数・文節数・単語数，比較のために短単位ベース UD の数値も掲載）

[14] UD の言語資源を公開するにあたって validator (`https://universaldependencies.org/validation-rules.html`) を通す必要がある．UD の validator 側の更新に合わせて，規則の精緻化や validator に必要な例外規定の整備をリリース毎に行うため都度変更は発生することが多い．

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM
GSD	1.98%	21.62%	1.22%	10.93%	0.42%	0.51%	0.01%	30.05%	2.67%
GSDLUW	2.60%	27.37%	1.61%	12.24%	0.55%	0.66%	0.01%	23.33%	2.08%
PUD	2.13%	23.07%	1.12%	11.53%	0.50%	0.73%	0.00%	25.79%	2.28%
PUDLUW	2.78%	28.38%	1.49%	12.29%	0.63%	0.91%	0.00%	20.50%	1.96%
BCCWJ	2.14%	20.03%	1.51%	9.74%	0.41%	0.48%	0.07%	29.24%	3.11%
BCCWJLUW	2.70%	24.61%	1.95%	11.17%	0.57%	0.61%	0.10%	22.66%	2.24%

	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
GSD	0.65%	0.57%	3.69%	9.93%	4.13%	0.67%	10.96%	0.00%
GSDLUW	0.45%	0.70%	3.03%	12.61%	2.15%	0.07%	10.57%	0.00%
PUD	0.73%	1.54%	4.73%	10.61%	3.44%	0.77%	11.02%	0.00%
PUDLUW	0.31%	1.88%	4.36%	13.22%	1.37%	0.02%	9.90%	0.00%
BCCWJ	1.18%	0.90%	2.87%	11.69%	4.49%	1.53%	10.57%	0.03%
BCCWJLUW	1.06%	1.08%	2.55%	14.52%	2.68%	1.27%	10.20%	0.03%

表 4　日本語 UD Version 2.10 の統計情報.（UPOS の割合）

	acl	advcl	advmod	amod	appos	aux	case	cc	ccomp
GSD	3.62%	3.74%	1.18%	0.23%	0.00%	8.90%	21.33%	0.42%	0.20%
GSDLUW	4.65%	4.77%	1.58%	0.12%	0.00%	11.42%	27.37%	0.55%	0.24%
PUD	3.82%	3.21%	1.09%	0.29%	0.17%	9.60%	22.56%	0.50%	0.26%
PUDLUW	4.78%	3.98%	1.46%	0.16%	0.21%	11.79%	28.31%	0.63%	0.30%
BCCWJ	3.62%	3.85%	1.43%	0.25%	0.00%	7.56%	19.65%	0.41%	0.22%
BCCWJLUW	4.58%	4.80%	1.87%	0.12%	0.00%	10.21%	24.59%	0.56%	0.27%

	compound	cop	csubj	dep	det	discourse	dislocated	fixed	mark
GSD	14.14%	1.26%	0.08%	0.04%	0.51%	0.01%	0.14%	4.41%	4.06%
GSDLUW	1.42%	0.81%	0.11%	0.03%	0.66%	0.01%	0.20%	0.01%	2.59%
PUD	10.55%	1.21%	0.03%	0.04%	0.73%	0.00%	0.11%	5.02%	3.37%
PUDLUW	0.91%	0.48%	0.03%	0.06%	0.91%	0.00%	0.14%	0.03%	1.67%
BCCWJ	14.67%	1.20%	0.11%	0.99%	0.48%	0.03%	0.10%	4.26%	5.04%
BCCWJLUW	3.00%	0.94%	0.14%	1.14%	0.61%	0.04%	0.15%	0.03%	3.73%

	nmod	nsubj	nummod	obj	obl	parataxis	punct	reparandum	root
GSD	6.75%	4.12%	1.45%	2.74%	6.55%	0.00%	9.93%	0.00%	4.18%
GSDLUW	8.09%	5.47%	0.39%	3.53%	7.99%	0.00%	12.61%	0.00%	5.39%
PUD	8.01%	5.16%	1.50%	2.93%	5.74%	0.00%	10.61%	0.00%	3.47%
PUDLUW	9.13%	6.52%	0.26%	3.68%	6.98%	0.00%	13.22%	0.00%	4.36%
BCCWJ	6.92%	3.78%	1.16%	2.62%	5.41%	0.00%	11.69%	0.00%	4.55%
BCCWJLUW	8.06%	4.88%	0.22%	3.30%	6.51%	0.00%	14.52%	0.00%	5.74%

表 5　日本語 UD Version 2.10 の統計情報.（DEPREL の割合）

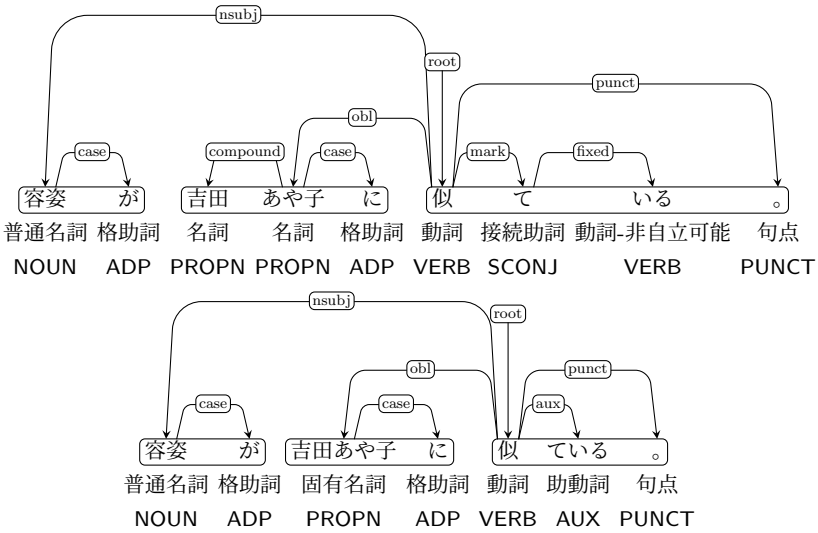


図 5　UD Japanese-GSD/GSDLUW 中の train-s1430 の事例．上が短単位, 下が長単位である．長単位になると *compound* と *fixed* の関係がなくなる．また,「ている」のように結合し助動詞 (AUX) に変化している．

4　オープンソース解析器による比較実験

この節では短単位および長単位という単位の異なる日本語 UD コーパスについて，公開されている解析器により統語解析を行う．正解率に基づき，どの段階の構造が，どの程度再現可能かを調査する． 統語解析の解析結果を比較することにより，短単位と長単位による解析精度の違い，長単位の再現度の困難さなどを確認することが目的である．

4.1　実験設定

この実験では短単位 UD と長単位 UD の比較のため, 短単位の UD である **UD_Japanese-GSD** と長単位の UD である **UD_Japanese-GSDLUW**（いずれも Version 2.10）を用いた． GSD データは CoNLL-2017 の Shared Task にも用いられていた経緯から，訓練・開発・評価データに分割されており，実験データの設定もこの分割に基づいて実施した．

図 6 に統語解析の流れを示す． 生文を入力してから単語分割, 品詞付与とレンマ推定をし, 依存構造解析を行う形[15]で統語解析を行い, UD を生成する．解析器としては UDPipe, MeCab, Comainu を用いた． UDPipe では単語分割, 品詞推定+レンマ推定, 依存構造解析を行い, MeCab と Comainu は単語分割および品詞推定 (XPOS まで) を行う．UD では言語依存の品詞を XPOS

[15] 日本語では単語分割と品詞付与およびレンマ推定, すなわち形態素解析として同時に行われるものが多いが, 本稿では UDPipe のフェイズに合わせた形で説明する．

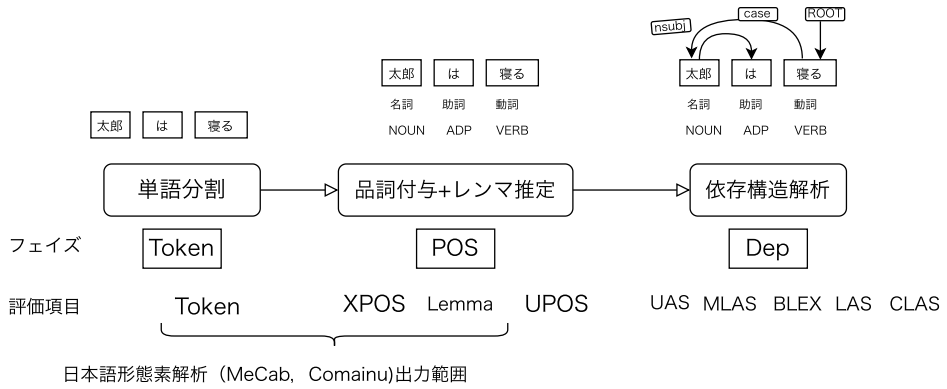


図 6　本実験における統語解析の流れ．この 3 段階の流れは UDPipe のモデルに合わせている．

として指定できるため, 日本語 UD では Unidic 品詞（国語研短単位形態論情報, 国語研長単位形態論情報）を提供している．

UDPipe (Straka and Straková 2017) は，生文あるいは UD コーパスを入力として単語分割，品詞付与+レンマ推定，依存構造解析ができる解析器である．UDPipe ではそれぞれの段階ごとにモデルを学習し構築している．UDPipe はニューラルネットモデルをフェイズごとに組み合わせて実装されており，単語分割には Bidirectional LSTM artificial neural network (Graves and Schmidhuber 2005) を，品詞付与+レンマ推定には MorphoDiTa (Straková et al. 2014) を，依存構造解析には Parsito (Straka et al. 2015) が使われている．**UD_Japanese-GSD**, **UD_Japanese-GSDLUW** の訓練データ (Version 2.10) を用いて UDPipe Version 1.2.0 により訓練したモデルを用いる[16]．UDPipe では，単語埋め込みを依存構造解析段階でのみ使用することができるため，単語埋め込みとして，短単位に基づく単語埋め込み NWJC2vec を用いた．なお NWJC2vec は短単位で構築されているが, 短単位と長単位では共通する形態素も存在するため, 影響をみるために, 長単位のモデルでも NWJC2vec を用いている．本実験での NWJC2vec は 300 次元の Skip-gram のモデルを使用した．

MeCab (Kudo et al. 2004) は国語研短単位形態論情報（単語分割，品詞付与+レンマ推定）を付与するために用いる．利用したバージョンは MeCab-0.996 と UniDic-2.1.0 であった．Comainu (小澤 他 2014) は国語研長単位形態論情報（単語分割，品詞付与+レンマ推定）を付与するために用いる．利用したバージョンは 0.72 であった．なお，Comainu は内部で MeCab を呼び出して利用しており，MeCab の解析結果も Comainu の結果から引用している．MeCab も Comainu

[16] UDPipe の最新版は version 2 系列であるが，2022 年 6 月時点でモデルを学習するインターフェースなどが見当たらず，また単語分割の機能は実装されていないため，利便性から UDPipe version 1.2.0 を用いている．

もすでに訓練されたモデルをそのまま用いた[17].

評価も図 6 に示す 3 つの段階「単語分割」「品詞付与+レンマ推定」「依存構造解析」の段階での比較で行った．1 つ目は未解析文を入力にしたすべての解析（単語分割，品詞付与+レンマ推定，依存構造解析）を行うものである．2 つ目は正解の単語分割 (Gold) を入力にして品詞付与+レンマ推定・依存構造解析を行うものである．3 つ目は正解の単語分割 (Gold) と正解の品詞タグ (Gold) を入力に依存構造解析のみを行うものである．

評価プログラムには CoNLL 2018 Shared Task にて使用された評価スクリプト[18]を用いた．このスクリプトは出力結果と正解同士の結果を内部でアライメントし，その一致率を求めた上で，単語分割，品詞付与，レンマ推定，依存構造解析結果それぞれについての F 値を出力する．本稿の実験結果もこの F 値を正解率として示している． 以降は各段階に焦点を当てて結果と考察を示すが, 全体の結果は付録 A の表 10 を参考にされたい.

4.2 単語分割および品詞付与+レンマ推定

表 6 に単語分割および品詞タグ付け + レンマ推定の結果を示す．Words は単語分割の F 値，UPOS は UD 品詞付与の F 値， XPOS は UniDic 品詞付与の F 値，Lemmas はレンマ（UniDic 語彙素）推定の F 値である．なお，品詞付与+レンマ推定については XPOS と UPOS のそれぞれの正解率を求めているが，MeCab と Comainu では XPOS のみ生成し，UPOS のみ UDPipe で出力する形になっている． また示している Comainu の結果は短単位の解析誤りが含まれている．

まず，単語分割においては MeCab と Comainu を比較すると，長単位 (LUW) のほうが正解

Token	POS			Words	UPOS	XPOS	Lemmas
	XPOS	UPOS					
MeCab	MeCab	UDPipe	SUW	96.84%	94.30%	95.81%	94.21%
Comainu	Comainu	UDPipe	LUW	97.26%	94.49%	96.14%	93.58%
UDPipe	UDPipe	UDPipe	SUW	96.18%	93.96%	93.19%	94.56%
			LUW	95.34%	92.87%	92.39%	92.97%
Gold	UDPipe	UDPipe	SUW	—	97.34%	96.37%	97.81%
			LUW	—	97.09%	96.50%	96.90%

表 6 実験結果：単語分割および品詞タグ付け + レンマ推定. 各項目は F 値を表している.

[17] これは Comainu-0.72 により GSD を用いてモデルを訓練し用いた結果と付属モデルでの解析結果を比較した際，主に品詞推定において再訓練したものが著しく低かったためである．これは Comainu の付属モデルに用いた訓練テキスト量が GSD のテキスト量よりも多いためと考えられ，また品詞体系も差異があるため完全な比較が難しい．我々の目的とする再現性の比較には不要と判断した.

[18] https://universaldependencies.org/conll18/evaluation.html

率が高く[19]，品詞付与の正解率も UPOS, XPOS ともに長単位のほうが高い．また, UDPipe と MeCab と Comainu を比較すると， MeCab と Comainu のほうが精度が高く, XPOS が特に高い．日本語形態素解析では, 辞書などを用いながら UniDic 品詞 (XPOS) を推定することに特化しているため，XPOS のほうが 汎化の高い UPOS よりも性能が高くなる傾向にある．

次に， UDPipe を用いた場合の短単位 (SUW) と長単位 (LUW) を比較すると，単語分割において短単位のほうが正解率が高いことが分かった．品詞タグ付けやレンマ推定においても，短単位のほうが高いことが分かる．前述の通り短単位は短く揺れが少ない傾向にあるため，辞書知識を用いない UDPipe において，短単位のほうが正解率が高いのは直感的である．全体として分割精度が低いと後続の結果にも影響がでるため，短単位のほうが結果として正解率が高くなっていると考えられる．

単語分割を正解とした入力を与えた場合には， XPOS においては長単位のほうが短単位よりも正解率が高いことが分かった．これは辞書を用いずに文脈のみで機械学習により推定する場合には，「可能性に基づく品詞」の推定が困難で，「用法に基づく品詞」のほうがより推定しやすいことによる．

レンマ推定においては，一貫して数字表現や複合辞の語彙素の復元が難しく，長単位のほうが正解率が低い．レンマ推定の際，Gold データの場合と解析器によるものを比較すれば 3–4% も落ちていることが分かり，単語分割の正確性は重要であることが伺える．

4.3　依存構造解析

表 7 に依存構造解析結果を示す．評価指標として UAS，LAS，CLAS，MLAS，BLEX を示している．NWJC2vec を利用していないものと利用したものの結果について **(UDPipe) w/o vec** と **(UDPipe) w/ vec** にて示している．

UAS (Unlabeled Attachment Score) は，親への依存関係 (dependency attachment) が正しい場合に正解として，正解率を評価する指標である．LAS (Labeled Attachment Score) は，単語の係り先と親への依存関係ラベル (universal dependency label) が正しい場合に正解として，正解率を評価する指標である．CLAS (Content-Word Labeled Attachment Score) は，LAS の評価において，機能語 (functional words) から自立語 (content words) への依存関係の重みを 0 にしたものである．つまり，自立語の評価を主とした指標である．自立語と認定される依存関係は 29 種として定義されている[20]のみとする．MLAS (Morphology-Aware Labeled Attachment

[19] うまく分割できてないものの大半が「数字表現」である．日本語 UD コーパスでは数字表現を前処理で統合しひとつの数字表現としており， UniDic では短単位・長単位関係なく数字表現は 1 桁ずつ分割するという違いがあり，その結果，単語数が多くなる短単位が多少評価が不利になっている．そのためこの正解率は参考程度となる．

[20] この 29 種類は nsubj obj iobj csubj ccomp xcomp obl vocative expl dislocated advcl advmod discourse nmod appos nummod acl amod conj fixed flat compound list parataxis orphan goeswith reparandum root dep となっている．

Token	POS	Dep		UAS	LAS	CLAS	MLAS	BLEX
MeCab	MeCab+UDPipe	UDPipe	SUW	86.95%	84.90%	77.39%	74.81%	73.87%
Comainu	Comainu+UDPipe	w/o vec	LUW	88.42%	86.69%	79.14%	75.63%	73.74%
MeCab	MeCab+UDPipe	UDPipe	SUW	87.77%	86.05%	79.30%	76.73%	75.78%
Comainu	Comainu+UDPipe	w/ vec	LUW	88.43%	86.74%	79.09%	75.80%	73.67%
UDPipe	UDPipe	UDPipe	SUW	84.37%	82.56%	75.92%	73.33%	74.15%
		w/o vec	LUW	84.31%	82.96%	74.10%	70.65%	70.65%
UDPipe	UDPipe	UDPipe	SUW	85.10%	83.57%	77.58%	74.98%	75.85%
		w/ vec	LUW	84.38%	83.06%	74.19%	70.84%	70.65%
Gold	UDPipe	UDPipe	SUW	91.03%	88.88%	83.39%	80.60%	80.88%
		w/o vec	LUW	93.22%	91.35%	83.50%	79.86%	78.87%
Gold	UDPipe	UDPipe	SUW	91.71%	89.92%	85.09%	82.33%	82.65%
		w/ vec	LUW	93.27%	91.38%	83.45%	80.04%	78.74%
Gold	Gold	UDPipe	SUW	92.04%	90.92%	84.98%	84.46%	84.98%
		w/o vec	LUW	93.97%	93.34%	86.05%	85.65%	86.05%
Gold	Gold	UDPipe	SUW	93.34%	92.57%	87.53%	87.23%	87.53%
		w/ vec	LUW	94.07%	93.57%	86.44%	86.07%	86.44%

表 7 実験結果：依存構造解析の結果. 各項目は F 値を表している.

Score) は，CLAS の拡張であり, 親への依存関係のみならず， 子の機能語の依存関係がすべて対応しており， いくつかの形態論的属性 (FEATS) も正しい場合に正解として，正解率を評価する指標である． この指標での依存関係は係り先と関係ラベルが一致しているものだけを評価する．ただし内容語でも機能語でもない関係ラベルのものは除かれて評価されている．BLEX (Bilexical Dependency Score) は，MLAS と似た指標だが，形態論的属性 (FEATS) の代わりにレンマ (LEMMA) が一致している場合に正解とする．その際，子の機能語の依存関係は評価しない．なお， 出力と正解データで単語分割の区切りが一対一となるとは限らないため, 一致率ではなく， F 値として評価する.

まず， UAS, LAS を確認する．単語分割に UDPipe を用いたものを除いて，長単位のほうが短単位よりも UAS, LAS ともによいことが分かる． これは， 品詞付与のときも同様であるが, UDPipe の単語分割の性能の差が依存構造解析に残っていることに基づく．単語分割に正解を与えた場合 (Token: Gold)， いずれの場合も依存構造解析の性能において 0.73–2.47%程度， 短単位よりも長単位のほうが性能がよかった．また品詞に正解を与えた場合 (Token: Gold, POS: Gold) も, 短単位よりも長単位のほうが性能がよかった．これは，依存構造解析の単位として，長単位の分割および長単位に付与されている UPOS のほうが， 構文解析しやすいことが示唆される．定性的にも，短単位に付与されている「可能性に基づく品詞体系」は依存構造を推定するための情報として曖昧な品詞体系であり，長単位に付与されている「用法に基づく品詞体系」が有効と考えられる.

CLAS, MLAS, BLEX を確認すると，短単位のほうが長単位よりも正解率がよい．とくに BLEX は，レンマ解析結果の正解も必要なため，より長単位の評価に厳しい結果となっている．これは CLAS, MLAS, BLEX における自立語認定される語が，短単位のほうが長単位よりも比率が多いことに起因している．とくに短単位では長単位のときに 1 語と認定されているものが複数の語となり，かつその依存関係（*compound* や *fixed*）も自立語として正解率に集計されている．実際，**UD_Japanese-GSD** で自立語として認定される語の割合は 103769/217954=0.4761，**UD_Japanese-GSDLUW** で自立語として認定される語の割合は 66116/174543=0.3788 となっており，短単位のほうが自立語認定された語の割合が 1 割程度多い．これが，正解率に影響している可能性が高い．

自立語の依存構造が正しく認定されているかを確認するために，文節中の自立語同士の依存関係のみを抽出し，結果を出した[21]．短単位 UD と長単位 UD の文節の数はほとんど一致しており，文節中には自立語はひとつと定義されているので，この結果は日本語の係り受け構造解析とほぼ同様の結果となる．それぞれの評価指標を計算しなおすと表 8 のようになった．解析器がその単語が主辞かどうかを判定しないため，語の認定がずれている結果は正確に評価できない．そのため，語が一致しているもののみに限定している．短単位と比べて長単位のほうが自立語の依存構造においては解析精度がよいことが分かる．

表 9 に，単語埋め込み NWJC2vec を用いた際に，どの程度依存構造解析の正解率が向上するかについて示す．表 7 の w/ vec と w/o vec の差分を示したものである．短単位に基づく NWJC2vec は，短単位の依存構造解析正解率を 0.73-1.65% 程度向上させる．一方，長単位の依存構造解析正解率は 0.01–0.23% 程度しか向上させることができない．NWJC2vec は短単位に

Token	POS	Dep		CLAS-C	MLAS-C	BLEX-C
Gold	UDPipe	UDPipe	SUW	76.24%	73.97%	73.33%
		w/o vec	LUW	82.02%	78.66%	77.81%
Gold	UDPipe	UDPipe	SUW	78.40%	76.08%	75.62%
		w/ vec	LUW	81.90%	78.64%	77.59%
Gold	Gold	UDPipe	SUW	78.08%	77.50%	78.08%
		w/o vec	LUW	84.70%	84.40%	84.70%
Gold	Gold	UDPipe	SUW	81.54%	81.18%	81.54%
		w/ vec	LUW	84.96%	84.70%	84.96%

表 8　実験結果：文節内自立語主辞の依存構造解析結果 (F 値)．表 7 の一部に対して文節内自立語主辞の依存関係のみで再評価した結果．CoNLL 2018 Shared Task の CLAS, MLAS, BLEX の定義と異なるため ‘-C‘を付与している．

[21] 具体的には文節内自立語主辞のみに限定した．これは日本語 UD の MISC 列にある `BunsetuPositionType` の `SEM_HEAD` を抽出することと同値である．

単語分割	UAS				LAS			
(Token →)	MeCab/	UDPipe	Gold	Gold	MeCab/	UDPipe	Gold	Gold
(POS →)	Comainu	UDPipe	UDPipe	Gold	Comainu	UDPipe	UDPipe	Gold
SUW	+0.82%	+0.73%	+0.68%	+1.30%	+1.15%	+1.01%	+1.04%	+1.65%
LUW	+0.01%	+0.07%	+0.05%	+0.10%	+0.05%	+0.10%	+0.03%	+0.23%

表 9　実験結果：単語埋め込みの利用による依存構造解析正解率の向上．（表 7 に基づく）

基づき構築された単語埋め込みである．これは依存構造解析において，単語単位が揃っている単語埋め込みは正解率の向上に貢献することを示唆している．

5　おわりに

本稿では日本語 UD における単語分かち書きの問題について，国語研短単位と国語研長単位という単語単位を比較することで検討を行った．日本語分かち書き基準として以前より採用されていた国語研短単位について紹介するとともに，UD が掲げる理念に即した単位認定「構文的な語」にふさわしい単位として国語研長単位があることを紹介した．実際に，長単位に基づく現代日本語の UD リソース **UD_Japanese-BCCWJLUW**, **UD_Japanese-GSDLUW**, **UD_Japanese-PUDLUW** を構築し，共有・公開を行った．さらに長単位に基づく日本語 UD について，公開されているツール・言語資源および **UD_Japanese-GSDLUW** を用いて，その再構成可能性について検討を行った．結果，既存の形態素解析器 MeCab・Comainu とともに短単位に基づく単語埋め込み NWJC2vec を用いた設定において，短単位と長単位とで最終的な係り受けの性能において差がないことを確認した．

構文的な語としての長単位の出力が実応用的に有用であり， UD の枠組にふさわしいものであるかについてはまだ議論の余地が残っている．さらに，類型論の研究を考えた場合に，他言語においても同様の「構文的な語」を規定できるかという問題がある．今後他言語を含めた有用性を検証する必要があると考えられる．

検討事項があるとはいえ，今回の取組みにより，既存のアノテーション情報を元に短単位の日本語 UD コーパスと長単位の日本語 UD コーパスを構築することが可能となった．本基準を用いると，たとえば，短単位・長単位形態論情報・文節係り受けなどが付与されていれば，短単位・長単位に基づく『日本語日常会話コーパス』(CEJC) なども UD 対応することが可能であろう． 今後拡張 CaboCha 形式の日本語コーパスを UD 基準に変換するプログラムもオープンソースソフトウェアとして公開予定である．また公開された長単位の日本語 UD コーパスを元に 松田 他 (2022) や安岡 (2022) の研究のような長単位解析系のツールの開発なども取り組ま

れている．長単位の日本語 UD コーパスによって，日本語の語の単位認定について，より進んだ研究が取り組まれることを期待する．

謝　辞

本研究の実施にあたって，松田寛氏・金山博氏・宮尾祐介氏・田中貴秋氏・松本裕治氏・伊藤薫氏の助言を受けました．ここに記して謝意を表します．本研究は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（2016–2021）・基幹型共同研究プロジェクト（2022–2027）「アノテーションデータを用いた実証的計算心理言語学」の成果です．また，科研費 17H00917, 18H05521 の支援を受けました．

参考文献

赤部晃一，神田峻介，小田悠介，森信介 (2022)．Vaporetto: 点予測法に基づく高速な日本語トークナイザ．言語処理学会第 28 回年次大会発表論文集, pp. 256–261, オンライン開催．言語処理学会．[K. Akabe et al. (2022). Vaporetto: Ten Yosokuho ni motozuku Kosoku na Nihongo Tokunaiza, Proceedings of the 28th Annual Meeting of the Association for Natural Language Processing, pp. 256–261. Online. Association for Natural Language Processing].

Asahara, M. (2018). “NWJC2Vec: Word Embedding Dataset from ‘NINJAL Web Japanese Corpus’.” *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, **24**, pp. 7–22.

浅原正幸, 松本裕治 (2018).『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, **25** (4), pp. 331–356. [M. Asahara and Y. Matsumoto (2018). Bunsetsu-based Dependency Relation and Coordinate Structure Annotation on ‘Balanced Corpus of Contemporary Written Japanese’. Journal of Natural Language Processing, 25 (4), pp. 331–356.].

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 (2019). Universal Dependencies 日本語コーパス．自然言語処理, **26** (1), pp. 3–36. [M. Asahara et al. (2019). Japanese Universal Dependencies Corpora. Journal of Natural Language Processing, 26 (1), pp. 3–36.].

Chun, J., Han, N.-R., Hwang, J. D., and Choi, J. D. (2018). “Building Universal Dependency Treebanks in Korean.” In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2194–2202, Miyazaki, Japan. European Language Resources Association.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). "Universal Stanford Dependencies: A cross-linguistic typology." In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 4585–4592, Reykjavik, Iceland. European Language Resources Association.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). "Universal Dependencies." *Computational Linguistics*, **47** (2), pp. 255–308.

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 (2007). 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 (特集コーパス日本語学の射程). 日本語科学, **22**, pp. 101–123. [Y. Den et al. (2007). The Development of an Electronic Dictionary for Morphological Analysis and Its Application to Japanese Corpus Linguistics. Japanese Linguistics, 22, pp. 101–123.].

Derin, M. O. and Harada, T. (2021). "Universal Dependencies for Old Turkish." In *Proceedings of the 5th Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pp. 129–141, Sofia, Bulgaria. Association for Computational Linguistics.

Graves, A. and Schmidhuber, J. (2005). "Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures." *Neural Networks: The Official Journal of the International Neural Network Society*, **18** (5), pp. 602–610.

Kanayama, H., Torisawa, K., Mitsuishi, Y., and Tsujii, J. (2000). "A Hybrid Japanese Parser with Hand-crafted Grammar and Statistics." In *Proceedings of COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 411–417, Saarbrücken, Germany. European Language Resources Association.

Kayadelen, T., Ozturel, A., and Bohnet, B. (2020). "A Gold Standard Dependency Treebank for Turkish." In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 5156–5163, Marseille, France. European Language Resources Association.

Koiso, H., Amatani, H., Den, Y., Iseki, Y., Ishimoto, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., Usuda, Y., and Watanabe, Y. (2022). "Design and Evaluation of the Corpus of Everyday Japanese Conversation." In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*, pp. 5587–5594, Marseille, France. European Language Resources Association.

小澤俊介, 内元清貴, 伝康晴 (2014). 長単位解析器の異なる品詞体系への適用. 自然言語処理, **21** (2), pp. 379–401. [S. Kozawa et al. (2014). Adaptation of Long-Unit-Word Analysis System to Different Part-Of-Speech Tagset. Journal of Natural Language Processing, 21 (2), pp. 379–401.].

工藤拓, 松本裕治 (2002). チャンキングの段階適用による日本語係り受け解析. 情報処理学会

論文誌, **43** (6), pp. 1834–1842. [T. Kudo and Y. Matsumoto (2002). Japanese Dependency Analysis Using Cascaded Chunking. IPSJ Journal, 43 (6), pp. 1834–1842.].

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). "Applying Conditional Random Fields to Japanese Morphological Analysis." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 230–237, Barcelona, Spain. Association for Computational Linguistics.

Leung, H., Poiret, R., Wong, T.-s., Chen, X., Gerdes, K., and Lee, J. (2016). "Developing Universal Dependencies for Mandarin Chinese." In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 20–29. The COLING 2016 Organizing Committee.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguti, M., Tanaka, M., and Den, Y. (2014). "Balanced Corpus of Contemporary Written Japanese." *Language Resources and Evaluation*, **48** (2), pp. 345–371.

松田寛，大村舞，浅原正幸 (2022). UD Japanese に基づく国語研長単位解析系の構築. 言語処理学会第 28 回年次大会発表論文集, pp. 1618–1622, オンライン開催. 言語処理学会. [H. Matsuda et al. (2022). UD Japanese ni motozuku Kokugoken Chotani Kaisekikei no Kochiku. Proceedings of the 28th Annual Meeting of the Association for Natural Language Processing, pp. 1618–1622. Online. Association for Natural Language Processing].

松吉俊，浅原正幸，飯田龍，森田敏生 (2014). 拡張 CaboCha フォーマットの仕様拡張. 第 5 回コーパス日本語学ワークショップ予稿集, pp. 223–232, 東京. 国立国語研究所. [S. Matsuyoshi et al. (2014). RFC: Requirements in Extended CaboCha Formats. Proceedings of the 5th Japanese Corpus Llinguistic Workshop, pp. 223–232, Tokyo. National Institute for Japanese Language and Linguistics.].

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). "Universal Dependency Annotation for Multilingual Parsing." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Murawaki, Y. (2019). "On the Definition of Japanese Word." *arXiv preprint arXiv:1906.09719*.

国立国語研究所 (2006). 日本語話し言葉コーパスの構築法. 国立国語研究所報告. 124 号. 国立国語研究所. [National Institute for Japanese Language and Linguistics (2006). Construction of The Corpus of Spontaneous Japanese. The National Language Research Institute Occasional Papers, No. 124. National Institute for Japanese Language and Linguistics.].

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). "Universal Depen-

dencies v1: A Multilingual Treebank Collection." In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666, Portorož, Slovenia. European Language Resources Association.

小椋秀樹, 小磯花絵, 冨士池優美, 原裕 (2008). 『現代日本語書き言葉均衡コーパス』形態論情報規程集 改定版. 国立国語研究所. 国立国語研究所内部報告書；LR-CCG-07-04, [H. Ogura et al. (2008). Gendai Nihongo Kakikotoba Kinko Kopasu Keitairon Joho Kiteishu Kaiteiban. National Institute for Japanese Language and Linguistics; LR-CCG-07-04.].

Omura, M. and Asahara, M. (2018). "UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese." In *Proceedings of the 2nd Workshop on Universal Dependencies (UDW 2018)*, pp. 117–125, Brussels, Belgium. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). "A Universal Part-of-Speech Tagset." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2089–2096, Istanbul, Turkey. European Language Resources Association.

坂本美保, 川原典子, 久本空海, 高岡一馬, 内田佳孝 (2018). 形態素解析器『Sudachi』のための大規模辞書開発. 言語資源活用ワークショップ 2018 発表論文集, 3 巻, pp. 118–129, 東京. 国立国語研究所. [M. Sakamoto et al. (2018). Large Scale Dictionary Development for Sudachi. Proceedings of Language Resources Workshop 2018, Vol. 3, pp. 118–129, Tokyo. National Institute for Japanese Language and Linguistics].

佐藤敏紀, 橋本泰一, 奥村学 (2017). 単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会発表論文集, pp. 875–878, 茨城. 言語処理学会. [T. Sato et al. (2017). Tango Wakachigaki Jisho mecab-ipadic-NEologd no Jisso to Joho Kensaku ni okeru Kokateki na Shiyohoho no Kento. Proceedings of the 23rd Annual Meeting of the Association for Natural Language Processing, pp. 875–878, Ibaraki.].

Straka, M., Hajič, J., Straková, J., and Jr, J. H. (2015). "Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle." In *Proceedings of 14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pp. 208–220, Warsaw, Poland. Instytut Podstaw Informatyki PAN.

Straka, M. and Straková, J. (2017). "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe." In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99, Vancouver, Canada. Association for Computational Linguistics.

Straková, J., Straka, M., and Hajič, J. (2014). "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition." In *Proceedings of 52nd Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2014), pp. 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Takaoka, K., Hisamoto, S., Kawahara, N., Sakamoto, M., Uchida, Y., and Matsumoto, Y. (2018). "Sudachi: a Japanese Tokenizer for Business." In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Tanaka, T., Miyao, Y., Asahara, M., Uematsu, S., Kanayama, H., Mori, S., and Matsumoto, Y. (2016). "Universal Dependencies for Japanese." In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1651–1658, Portorož, Slovenia. European Language Resources Association.

Xia, F. (2000). "The Segmentation Guidelines for the Penn Chinese Treebank (3.0)." Tech. rep., Institute for Research in Cognitive Science. IRCS Technical Reports Series.

安岡孝一 (2022). Transformers と国語研長単位による日本語係り受け解析モデルの製作. 情報処理学会研究報告 : 第 128 回人文科学とコンピュータ研究発表会研究会報告, 2022-CH-128 巻, pp. 1–8, オンライン開催. 情報処理学会. [K. Yasuoka (2022). Transformers to Kokugoken Chotani niyoru Nihongo Kakariuke Kaiseki Moderu no Seisaku. Proceedings of the 128th IPSJ SIG Computers and the Humanities Symposium, 2022-CH-128, pp. 1–8, Online. Information Processing Society of Japan.].

Zeman, D. (2008). "Reusable Tagset Conversion Using Tagset Drivers." In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 213–218, Marrakech, Morocco. European Language Resources Association.

付録

A　依存構造解析の全体表

本稿で行った実験結果の全体表表 10 を付録として掲載する. この表は本稿の表 6 と表 7 を統合したものである.

Token	POS	Dep		Words	UPOS	XPOS	Lemmas	UAS	LAS	CLAS	MLAS	BLEX
MeCab/Comainu	MeCab/Comainu	UDPipe w/o vec	SUW	96.84%	94.30%	95.81%	94.21%	86.95%	84.90%	77.39%	74.81%	73.87%
	+UDPipe(UDPOS)		LUW	97.26%	94.49%	96.14%	93.58%	88.42%	86.69%	79.14%	75.63%	73.74%
MeCab/Comainu	MeCab/Comainu	UDPipe w/ vec	SUW	96.84%	94.30%	95.81%	94.21%	87.77%	86.05%	79.30%	76.73%	75.78%
	+UDPipe(UDPOS)		LUW	97.26%	94.49%	96.14%	93.58%	88.43%	86.74%	79.09%	75.80%	73.67%
UDPipe	UDPipe	Train w/o vec	SUW	96.18%	93.96%	93.19%	94.56%	84.37%	82.56%	75.92%	73.33%	74.15%
			LUW	95.34%	92.87%	92.39%	92.97%	84.31%	82.96%	74.10%	70.65%	70.65%
UDPipe	UDPipe	UDPipe w/ vec	SUW	96.18%	93.96%	93.19%	94.56%	85.10%	83.57%	77.58%	74.98%	75.85%
			LUW	95.34%	92.87%	92.39%	92.97%	84.38%	83.06%	74.19%	70.84%	70.65%
Gold	UDPipe	UDPipe w/o vec	SUW	—	97.34%	96.37%	97.81%	91.03%	88.88%	83.39%	80.60%	80.88%
			LUW	—	97.09%	96.50%	96.90%	93.22%	91.35%	83.50%	79.86%	78.87%
Gold	UDPipe	Train w/ vec	SUW	—	97.34%	96.37%	97.81%	91.71%	89.92%	85.09%	82.33%	82.65%
			LUW	—	97.09%	96.50%	96.90%	93.27%	91.38%	83.45%	80.04%	78.74%
Gold	Gold	Train w/o vec	SUW	—	—	—	—	92.04%	90.92%	84.98%	84.46%	84.98%
			LUW	—	—	—	—	93.97%	93.34%	86.05%	85.65%	86.05%
Gold	Gold	UDpipe w/ vec	SUW	—	—	—	—	93.34%	92.57%	87.53%	87.23%	87.53%
			LUW	—	—	—	—	94.07%	93.57%	86.44%	86.07%	86.44%

表 10 実験結果：統語解析全体の結果. すべて F 値で示している. Gold が正解データを与えた場合である. 単語分割の結果は後続の結果（品詞付与+レンマ推定, 依存構造解析）にも影響を与え, 正解データを与えた場合のほうが結果がよくなることが分かる.

略歴

大村　舞：国立国語研究所プロジェクト非常勤研究員.

若狭　絢：国立国語研究所プロジェクト非常勤研究員.

浅原　正幸：国立国語研究所・東京外国語大学教授.

（2022 年 8 月 1 日　受付）
（2022 年 10 月 3 日　採録）