UD Japanese-BCCWJの構築と分析

UD Japanese-BCCWJ の構築と分析

大村 舞 (国立国語研究所コーパス開発センター) *
浅原 正幸 (国立国語研究所コーパス開発センター)

Construction and Analysis of UD Japanese-BCCWJ

Mai Omura (National Institute for Japanese Language and Linguistics)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

自然言語処理の分野では多言語かつ言語横断的な言語研究が盛んに取り組まれている。その言語横断的な言語研究の取り組みとして Universal Dependencies (UD) がある。本論文では、日本語のコーパスである UD Japanese-BCCWJ について紹介をする。UD Japanese-BCCWJ は現代日本語書き言葉均衡コーパス (BCCWJ) に付随する係り受け情報などを組み合わせて、UD へと変換、構築した BCCWJ の Universal Dependencie である。これは日本語の UD の中でも 1980 文章、57,256 文、約 126 万単語を含む最大規模また複数のレジスターを内包したデータセットである。UD Japanese-BCCWJ の特徴について説明する。また UD Japanese-BCCWJ の構築手順について説明し、現状における問題点について議論する。

1. はじめに

Universal Dependencies (以下 UD) (Zeman et al. 2017) とは、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プロジェクト及びそのコーパス、枠組みのことを指す。我々は日本語版 UD を設計する活動として、日本語コーパスに対する品詞体系、ラベル付き依存構造の定義の策定、その Github 上での文書化と、参照用のコーパスの作成に着手している。

2018 年 7 月現在日本語版 UD では表 1 のように 5 種類の UD が公開されている (この表は文献 (Asahara et al. 2018) を参照して作成した)。日本語ウィキペディアから構築した **UD Japanese-GSD**、他言語間パラレルコーパスから構築された **UD Japanese-PUD** (Zeman et al. 2017)、Kaede treebank (Tanaka and Nagata 2013) から変換して構築した **UD Japanese-KTC** (Tanaka et al. 2016)、さらに「日本語歴史コーパス明治・大正編 I 雑誌 (CHJ) (Ogiso et al. 2017)」から構築した **UD Japanese-Modern** (Omura et al. 2017)、そして本稿で説明する **UD Japanese-BCCWJ** が公開済みである。

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 (BCCWJ) (Maekawa et al. 2014) に基いて構築された日本語 UD コーパス **UD Japanese-BCCWJ** について紹介する。UD Japanese-BCCWJ は他の日本語版 UD コーパスよりも大規模

* mai-om@ninjal.ac.jp

表 1　公開されている UD Japanese の一覧 (2018 年 7 月執筆時点)。

ツリーバンク	単語数	バージョン	Copyright	媒体
UD Japanese-BCCWJ	**1273k**	v2.2	内容分離	新聞、書籍、雑誌、ブログ etc.
UD Japanese-KTC	189k	v1.2	内容分離	新聞
UD Japanese-GSD	186k	v2.1	CC-BY-NC-SA	ウィキペディア
UD Japanese-PUD	26k	v2.1	CC-BY-SA	ウィキペディアのパラレルコーパス
UD Japanese-Modern	14k	v2.2	CC-BY-NC-SA	19 世紀の雑誌 (Ogiso et al. 2017)

で、また UD 上で公開されているコーパスの中でも、2 番目に大規模でかつ [(1)]、表 2 で示すような 6 種類のドメインのテキストで構成されたコーパスである。

本稿では UD Japanese-BCCWJ の構築、つまり、BCCWJ から UD の統語構造に変換する手順について説明していく。図 1 に BCCWJ の係り受け構造から UD の単語間係り受け構造に変換する手順の概略を示す。BCCWJ と UD には、品詞体系の違い、係り受け構造と単語間係り受け構造といった違いがある。そのため、これらの違いを考慮して変換する必要がある。そのためには BCCWJ に収録されている形態論情報のみではなく、係り受け構造や、並列構造の情報 (Asahara and Matsumoto 2016)、述語項構造情報 (植田ほか 2015) などを用いる必要がある。

日本語版 UD のプロジェクトでは BCCWJ から UD への変換を行ったことで、UD Japanese-BCCWJ を構築した。そして UD Japanese-BCCWJ や他の日本語版 UD を比較することで、日本語における統語構造と UD における統語表現の違いを比較、評価し、それらの結果についてプロジェクト内で議論を行っている。その結果を対外報告することで、UD プロジェクトに UD のフレームワークについて提言し、日本語版 UD のフレームワークの検討・改善に取り組んでいる。そこで本稿では UD Japanese-BCCWJ において問題となった点も取り上げていく。

2.　日本語における統語構造データと Universal Dependencies

表 2 に日本語版 UD の一覧を示している。現在、UD Japanese-BCCWJ を加えたことで、日本語版 UD は全 UD 内でも 2 番目に大規模な UD コーパスとなっている。公開されているコーパスとして **UD Japanese-KTC** (Tanaka et al. 2016)、**UD Japanese-GSD**、**UD Japanese-PUD** (Zeman et al. 2017)、**UD Japanese-Modern** (Omura et al. 2017) が存在する。これらの方針としては、既存の日本語統語データを用い、UD のフォーマットに自動変換することで低コストで日本語版 UD の構築を実現している。

UD 以外の、存在している日本語の統語構造コーパスには、京都大学テキストコーパス (Kurohashi and Nagao 2003) 、日本語係り受けコーパス (Mori et al. 2014)、Kaede treebank (Tanaka and Nagata 2013) などが存在する。これらのコーパスに共通していることとして、日本語の文節係り受け構造を元にして構築されていることが挙げられる。文節係り受け構造では、文節と

(1) 2018 年 7 月現在 http://universaldependencies.org/ 調べ。最大規模のコーパスはチェコ語の UD Czech-PDT である。

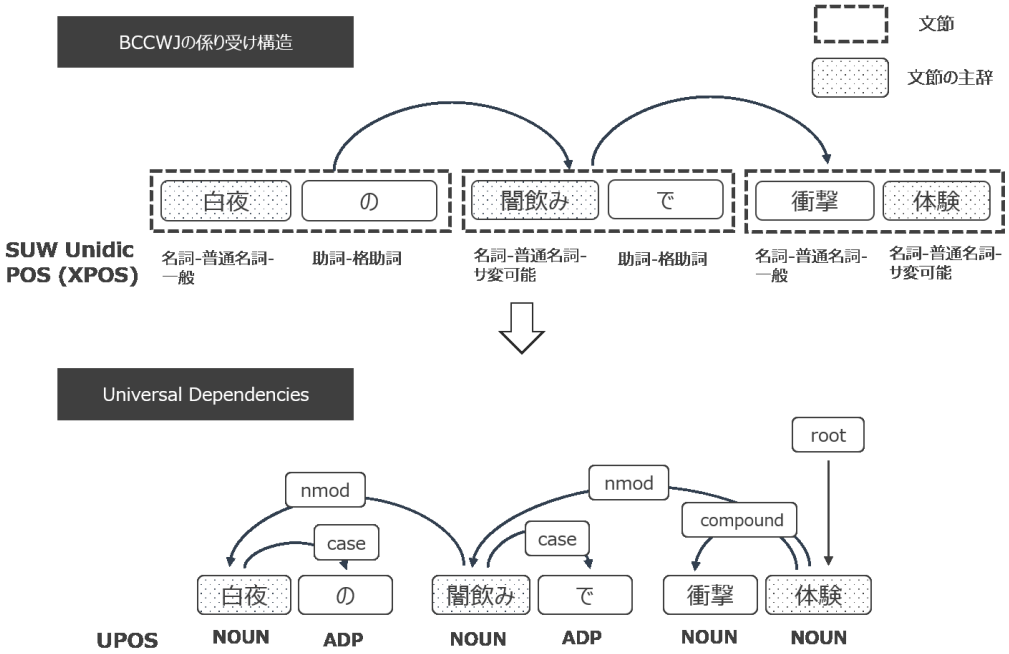

図 1　BCCWJ から UD Japanese-BCCWJ への変換の概要 (サンプルは PB_00001 から)。上の例が BCCWJ、下の例が UD Japanese-BCCWJ を表現している。

いう単語のグループ[(2)]を構成し、文節間の係り関係を記述する形で表現された統語構造であり、図 1 の上部図のような統語構造を持っている。UD Japanese-BCCWJ の基となる現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa et al. 2014) においてもこのような係り受け構造で統語構造を表現している。

一方 Universal Dependencies (UD) では、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、句構造を考慮せず、すべての構文構造を単語間の係り関係とその係り関係のラベルで表現する。異なる言語間で係り受け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。そのため図 1 の下部図のような、内容語間の係り受け構造を中心とした表現を採用している。

3.　現代日本語書き言葉均衡コーパス (BCCWJ)

現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa et al. 2014) は、1 億 430 万語のデータを格納した、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。サンプルの幅についても、書籍全般、雑誌全般、新聞、白書、ブログ、 ネット掲示板、教科書、法律などいった多領域のジャンル [(3)]が収録されている。

すべての収録サンプルは自動形態素解析によって言語単位、品詞付与が施されている。それぞれのサンプルは 2 種類の形態素、短単位 (Short Unit Word, SUW) と長単位 (Long Unit Word,

(2)　例えば図 1 の場合「白夜/の」「闇のみ/で」「衝撃/体験」が文節である。

(3) BCCWJ においてはこれをレジスターと呼んでいるがここでは他言語の UD とも比較するためジャンルという語で統一する。

表 2　BCCWJ のコアデータに収録されているレジスターの略称一覧

略称	説明
OC	Yahoo!知恵袋
OW	白書
OY	Yahoo!ブログ
PB	書籍
PM	雑誌
PN	新聞

表 3　BCCWJ コアデータのジャンルの分布。略称は表 2 を参照のこと。

ジャンル		OC	OW	OY	PB	PM	PN	合計
文章数	train	421	45	214	58	63	286	1,087
	dev	259	9	129	13	12	27	449
	test	258	8	128	12	11	27	444
	total	938	62	471	83	86	340	1,980
文数	train	2,838	4,456	3,278	7,196	9,546	13,487	40,801
	dev	1,650	780	1,920	1,131	1,510	1,436	8,427
	test	1,619	589	1,722	1,351	1,486	1,114	7,881
	total	6,107	5,825	6,920	9,678	12,542	16,037	57,109
単語数	train	50,415	168,909	51,310	174,394	177,947	300,786	923,761
	dev	29,961	31,471	32,164	27,315	30,328	29,528	180,767
	test	29,624	26,421	28,485	29,612	28,183	26,434	168,759
	total	110,000	226,801	111,959	231,321	236,458	356,748	1,273,287

LUW) という言語単位で解析されてそれぞれ公開されている。短単位は日本語の形態的側面に着目した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。

さらにこれらのデータに対して、BCCWJ の中 1% のサンプルは人手によって解析の誤りを修正されている。この修正されたデータを「コアデータ」と呼ぶ。BCCWJ のコアデータは 1980 文書、57,256 文が収録されており、UD Japanese-BCCWJ はこのコアデータを元に変換している。表 2 に BCCWJ のコアデータに収録されているジャンルの略称の一覧を示し、表 3 に BCCWJ のコアデータの統計を示す。

BCCWJ ではさらに、文節レベルの係り受け構造の情報を BCCWJ-DepPara (Asahara and Matsumoto 2016) で提供している。BCCWJ-DepPara には文節という単語単位のレイヤー情報、文節同士の係り関係の情報、単語間の並列関係の情報などが収録されている。また、BCCWJ-PAS (植田ほか 2015) によって、述語に対する格関係情報を記述した述語項構造という情報も提供されている。述語項構造は UD 関係ラベルを付与する際に参照している。UD Japanese-BCCWJ では形態素の情報、係り受け構造、述語構造などの情報を用いて UD への変換を試みている。

魚フライを食べたかもしれないペルシャ猫
"the Persian cat that may have eaten fried fish"

SUW	魚 NOUN *fish*	フライ NOUN *fry*	を ADP *−ACC*	食べ VERB *eat*	た AUX *−PAST*	か PART	も ADP	しれ VERB *know*	ない AUX *−NEG*	ペルシャ PROPN *Persia*	猫 NOUN *cat*
LUW	魚フライ NOUN *fried fish*		を ADP *−ACC*	食べ VERB *eat*	た AUX *−PAST*	かもしれない AUX *may*				ペルシャ猫 NOUN *Persia cat*	
bunsetsu	魚フライを			食べたかもしれない						ペルシャ猫	

図 2　短単位 (SUW)、長単位 (LUW)、文節の違いを表した例。

4.　BCCWJ から UD への変換手順

図 1 で分かるとおり、BCCWJ と UD の統語構造には違いがある。ひとつは、BCCWJ で使われている品詞体系 UniDic (伝ほか 2007) と UD で採用されている品詞体系 Universal POS(UPOS) (Petrov et al. 2012) とで異なるという点である。そして、BCCWJ は文節係り受けという文節単位の係り受け構造を採用しているのに対し、UD の統語構造は単語間の係り受け構造が要求されている。そして、UD では単語間に 37 種類ものある Universal Dependency Relations (ここでは依存関係ラベルと呼ぶ) という係り関係のラベルを付与する必要があるが、BCCWJ で用いる係り受け構造の情報にはここまで厳密に設定されていない [(4)]。そのため、これらの違いを考慮して変換する必要がある。本稿では、以下の手順で自動的に変換を試みた。

1. 単語単位を認定する。
2. UniDic の品詞体系 UPOS に変換する。
3. 文節係り受け構造を単語間依存構造に変換する。
4. 依存関係ラベルを付与する

それぞれの手順について、以降の節で説明する。

4.1　単語単位の認定

日本語は英語と異なり、単語の区切りが明示的に示されているわけではない。そのため、日本語版 UD における単語を決める必要がある。UD のガイドラインによると「統語的な単語 (syntactic words)」を単語として認定することが求められている。

前述のとおり、BCCWJ には短単位と長単位という言語単位が制定されている。また長単位を組み合わせた文節という単位も制定されている。文節は係り受け構造の単語単位にもなっている。そこで短単位、長単位、文節いずれかあるいはそれらを組み合わせた言語単位を UD で求められている単語とすることにした。図 2 に短単位、長単位、文節の例をあげている。単語認定について考えると、図 2 を例にした場合、例えば「魚フライを」という句は、短単位は「魚/フライ/を」の 3 つの単語に長単位は「魚フライ/を」という 2 つの単語に、そして文節は「魚フライを」という 1 つの単語となる。例から分かるように、短単位、長単位、文節には「短

(4) UD Japanese-BCCWJ で用いる文節係り受け構造の情報 BCCWJ-DepPara には、単語同士係り関係にあるか、並列構造にあるかなどの情報が付与されている。

表 4　Universal PoS version 2.0 (UPOS) の変換規則の一部。さらに具体的なものは (大村・浅原 2017) にも掲載している。

短単位の品詞	短単位基本形	長単位の用例	UPOS
^形容詞-非自立可能		形容詞-一般	AUX
^形容詞-非自立可能		助動詞	ADJ
^名詞-普通名詞-サ変可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-サ変可能		動詞-一般	VERB
^連体詞	^[こそあど此其彼] の		DET
^連体詞	^[こそあど此其彼]		PRON
^動詞-非自立可能	為る		AUX
^動詞			VERB
^名詞-固有名詞			PROPN
^名詞-普通名詞-副詞可能		副詞	ADV
^名詞-普通名詞-副詞可能			NOUN
接頭辞			NOUN
接尾辞			NOUN[(5)]

単位 <= 長単位 <= 文節」という階層関係があることがわかる。また後の 4.2 節でも述べるとおり、短単位と長単位ではそれぞれ異なった品詞体系を持っている。

UD Japanese-BCCWJ では短単位を統語的な単語として認定することにした。これは BCCWJ においては最小で基本的な言語単位、品詞体系を有している。ただし、後の節で説明するとおり、長単位のほうが求められている統語的な単語として、あるいは他言語比較の観点からして合っている可能性が高い。詳しくは 6.1 節にて説明する。

4.2　品詞の変換

UD では品詞体系として Universal PoS version 2.0 (UPOS) (Petrov et al. 2012) が採用されている。これらは多くの言語を定義するための 17 種類の品詞が制定されている。日本語版 UD でもこの UPOS を付与するために、BCCWJ で採用されている UniDic (伝ほか 2007) 品詞体系という品詞から UPOS に変換することで品詞の変換を実現する。

前述したとおり、この UniDic の品詞体系は短単位、長単位で異なっている。BCCWJ における短単位では**語彙主義**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**に基づく品詞を規定している。さらに短単位に対して、長単位を参照して長単位形態論情報として「用法」の情報が付与されている。短単位を単語として採用したため、品詞体系も短単位の語彙主義的な可能性に基づく品詞体系を採用する。

しかし、UD の品詞体系の標準にあわせる、あるいは他言語同士の比較をするという観点からすると長単位の用法主義に基づく品詞が求められる。例えば「する」を付与することで動詞化する「名詞-普通名詞-サ変可能」という品詞、「な」を付与することで形容詞化する「名詞-普通名詞-形状詞可能」という品詞が短単位の品詞体系には存在する。しかし、長単位の品詞体系であった場合、長単位は動詞であれば「XX する」のような言語単位が 1 つで構成され、これは確実に「動詞」であることが確定する。

(5) 日本語における接尾辞の品詞体系には「接尾辞」と書かれていても機能的なものから名詞的なものと幅があるため一概に NOUN を付与するのには議論の余地がある。現状 NOUN を付与することとする。

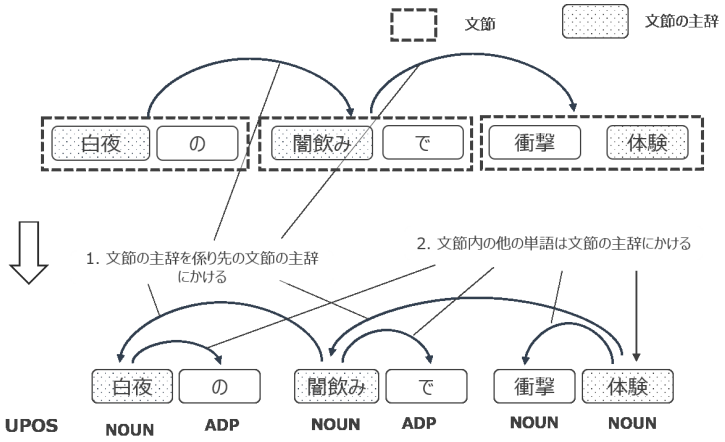

図 3　文節係り受けから UD の単語間係り受けへの変換の概略図

表 4 に UniDic 短単位の品詞体系から UPOS へ変換する規則を示す。表 4 で示している変換規則は短単位の品詞体系に基づいて変換しており、6.1 節で議論する通り、長単位で採用されている用法主義に基づく品詞体系を採用した場合さらに規則は単純になる。しかし、いくつかの理由により現状では用法主義に基づく品詞体系は採用していないものの、いずれ公開予定である。理由についても 6.1 節で説明する。

4.3　統語構造の変換

UD における単語間依存構造を得るために、日本語の統語構造である文節間係り受け構造を用いて変換する。BCCWJ には BCCWJ-DepPara (Asahara and Matsumoto 2016) という文節間係り受け構造・並列構造の情報が提供されている。BCCWJ-DepPara には文節の情報、係り受け関係の情報が収録されている。

BCCWJ の文節係り受け構造から UD の単語間係り受け構造に変換するために、文節間の係り関係のみではなく、それ以外の単語間でも係り情報を加える必要がある。BCCWJ-DepPara には文節の他にも「文節の主辞」(図 3 の網掛け部分) が設定されている。そこで図 3 のように、1. 文節の主辞同士でまず係り関係を結び、そして、2. それ以外の文節内単語に関しては文節の主辞にかける、という手順で文節係り構造から単語間係り構造に変換する。このとき、日本語の係り受け構造の場合、矢印は「係り元」から「係り先」にかかるような向きで表現するが、UD の場合矢印の向きが逆、つまり「係り先」から「係り元」に矢印が向く図になることに注意すること(6)。

日本語において文節の主辞は、図 3 の「衝撃/体験」の「体験」のように、文節の主辞は右側に置かれやすい傾向にある。これは日本語においては、主体となる名詞句は右側におき、補助的な要素は左側に置かれやすいからである。同様に日本語における文節間の係り関係は「左から右に」にかかりやすい。一方で、英語などの言語の場合「右から左」に向かう係り関係が存

(6) UD の単語間係り受け構造の図表現が「係り先」から「係り元」の方向になるだけで、後述のフォーマットのとおり、係り元の単語について、係り先を記述する形 (列 HEAD 参照) になっている。

表 5　依存関係ラベルの付与規則の一部。簡略的に書かれており実際の実装ではより詳細に設定されている。さらに具体的なものは (大村・浅原 2017) にも掲載している。ただし全ては掲載されていない。

ラベル付与ルール	ラベル
その係り元単語は係り先がなく (文末の文節である) でさらに文節の主辞である	root
その係り元単語は UPOSNUMMOD を持っている。	nummod
その係り元単語は UPOSADV を持っている	advmod
係り先単語は VERB を持っており、格助詞「が」が文節内にある	nsubj
係り先単語は VERB を持っており、格助詞「を」が文節内にある	obj
その係り元単語は UPOSVERB 持っており、その係り先単語は UPOSVERB を持っており、文節をまたがっている	aux
その係り元単語は UPOSVERB 持っており、その係り先単語は UPOSVERB を持っており、文節内の関係である	compound

在する場合がある。例えば並列表現の場合は、左に係り先をおいた表現を採用している。この違いが日本語版 UD における並列構造に影響を与えていることを 6.2 節にて議論する。

BCCWJ-DepPara には係り受け構造の情報や並列構造の情報は含まれているものの、UD で定義するように指定されている依存関係ラベル (Marneffe et al. 2014) のような詳細な係り関係の情報は含まれていない。依存関係ラベルには、例えば nsubj、obj、iobj、amod のような係り関係を定義するラベルが存在している(7)。そのため BCCWJ から用いることのできる情報などを利用して、単語間の係り関係に依存関係ラベルを付与する必要がある。表 5 に依存関係ラベルの付与規則の例をあげる。係り先単語について、文節の情報、格情報あるいは並列関係の情報などのを組み合わせることで依存関係ラベルを付与している。

nsubj、obj などのような統語構造の項は、格助詞などが (いわゆる助詞「が」「を」「に」など) 付与されているか否かで依存関係ラベルを付与する。UD の方針としては、あくまで統語構造を表現するものであるため、助詞の標識がある場合は、格標識に基づいて依存関係ラベルを付与する。しかし、日本語は英語とは異なり、必ずしも格標識「が」や「は」「を」などが文上の主体を表しているとは限らない。例えば「は」は通常であれば「私は学校に行く」と言ったとおり「私」が nsubj であるようにラベルを付与することができる。しかし、「象は鼻が長い」といった文の場合、「象」は Topic marker であるため、nsubj を付与すべきかどうかは不明瞭である。また「3 時に公園に行く」といったような文章だった場合、「に」という格助詞が衝突してしまう。この場合、BCCWJ-PAS (植田ほか 2015) の述語構造情報を参照する必要がある (8)。

なお現在のルールでは、 csubj、advcl、acl といった節に関するラベルを付与することができない。なぜならば、英語と比較して日本語は節かどうかの境界が曖昧だからである。節にかんしては 6.3 節にて議論をする。将来、この節の同定に関しても検討する必要がある。

BCCWJ-DepPara にはさらに、並列構造の情報が含まれており、並列の情報を用いて並列の情報 cc や conj を付与することになる。しかし、この並列構造情報を用いても、UD において

(7) 具体的な依存関係ラベルは http://universaldependencies.org/u/dep/index.html 参照。
(8) 日本語版 UD における格標識に関しては Asahara et al. (2018) の 3.4 節にて問題点を議論している。

```
# sent_id = OC01_00001-1
# text = 詰め将棋の本を買ってきました。
1 詰め 詰める VERB 動詞-一般 _ 2 aux _ BunsetuPosition=B|JPYomi=ツメル|BunsetuPositionType=CONT|SpaceAfter=No
2 将棋 将棋 NOUN 名詞-普通名詞-一般 _ 4 nmod _ BunsetuPosition=I|JPYomi=ショウギ|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
3 の の ADP 助詞-格助詞 _ 2 case _ BunsetuPosition=I|JPYomi=ノ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
4 本 本 NOUN 名詞-普通名詞-一般 _ 6 obj _ BunsetuPosition=B|JPYomi=ホン|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
5 を を ADP 助詞-格助詞 _ 4 case _ BunsetuPosition=I|JPYomi=ヲ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
6 買っ 買う VERB 動詞-一般 _ 8 advcl _ BunsetuPosition=B|JPYomi=カウ|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
7 て て SCONJ 助詞-接続助詞 _ 6 mark _ BunsetuPosition=I|JPYomi=テ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
8 き 来る VERB 動詞-非自立可能 _ 0 root _ BunsetuPosition=B|JPYomi=クル|BunsetuPositionType=ROOT|SpaceAfter=No
9 まし ます AUX 助動詞 _ 8 aux _ BunsetuPosition=I|JPYomi=マス|BunsetuPositionType=FUNC|SpaceAfter=No
10 た た AUX 助動詞 _ 8 aux _ BunsetuPosition=I|JPYomi=タ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
11 。 。 PUNCT 補助記号-句点 _ 8 punct _ BunsetuPosition=I|JPYomi=_|BunsetuPositionType=CONT|SpaceAfter=No

# sent_id = OC01_00001-2
# text = 駒と盤は持っていません。
1 駒 駒 NOUN 名詞-普通名詞-一般 _ 3 nmod _ BunsetuPosition=B|JPYomi=コマ|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
2 と と ADP 助詞-格助詞 _ 1 case _ BunsetuPosition=I|JPYomi=ト|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
3 盤 盤 NOUN 名詞-普通名詞-一般 _ 5 iobj _ BunsetuPosition=B|JPYomi=バン|BunsetuPositionType=SEM_HEAD|SpaceAfter=No
4 は は ADP 助詞-係助詞 _ 3 case _ BunsetuPosition=I|JPYomi=ハ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
5 持っ 持つ VERB 動詞-一般 _ 0 root _ BunsetuPosition=B|JPYomi=モツ|BunsetuPositionType=ROOT|SpaceAfter=No
6 て て SCONJ 助詞-接続助詞 _ 5 mark _ BunsetuPosition=I|JPYomi=テ|BunsetuPositionType=FUNC|SpaceAfter=No
7 い 居る AUX 動詞-非自立可能 _ 5 aux _ BunsetuPosition=I|JPYomi=イル|BunsetuPositionType=FUNC|SpaceAfter=No
8 ませ ます AUX 助動詞 _ 5 aux _ BunsetuPosition=I|JPYomi=マス|BunsetuPositionType=FUNC|SpaceAfter=No
9 ん ず AUX 助動詞 Polarity=Neg 5 aux _ BunsetuPosition=I|JPYomi=ズ|BunsetuPositionType=SYN_HEAD|SpaceAfter=No
10 。 。 PUNCT 補助記号-句点 _ 5 punct _ BunsetuPosition=I|JPYomi=_|BunsetuPositionType=CONT|SpaceAfter=No
    ....
```

図 4　BCCWJ の UD サンプル (OC01_00001)。上記のようにタブ区切りのテキストファイルになる。

表 6　CoNLL-U 形式の各列の説明

列	フィールド名	説明
1	ID	1-origin の ID (ROOT が 0)
2	FORM	書字形出現形
3	LEMMA	語彙素読みをローマ字にしたもの
4	UPOSTAG	品詞 Universal POS
5	XPOSTAG	品詞 BCCWJ の短単位品詞
6	FEATS	その他品詞情報 (“|” で OR を表現、順不同)
7	HEAD	係り先 ID
8	DEPREL	依存関係ラベル
9	DEPS	Secondary Dependency (List, Head-deprel pairs)
10	MISC	その他 (表 7 参照)

表 7　UD Japanese-BCCWJ における MISC フィールドの項目の一覧

ラベル	説明
BunsetuBILabel	文節の開始か中間かを表現 (B=開始、I=中間)。
BunsetuPositionType	文節の種類
LUWBILabel	長単位の開始か中間かを表現 (B=開始、I=中間)。
LUWPOS	UniDic 長単位品詞体系

解決できない点が存在する。この問題は 6.2 節で議論する。

4.4　フォーマット

以上の節で説明した通りの手順を経て、UD Japanese-BCCWJ は図 4 のようなフォーマットに変換される。このフォーマットはタブ区切りの UTF-8 の文字コードでエンコードされた CoNLL-X フォーマットに基づいている。それぞれの項目については表 6 に説明している。

UD では MISC フィールドを用いることで、さまざまな情報を付与させることができる。そのため、統語構造の情報として重要と思われる情報、長単位の情報、文節の情報を付与させる予定である [9]。表 7 に UD Japanese-BCCWJ の MISC フィールドで付与される情報の項目について説明している。

[9] 現行で公開されているバージョンでは付与されていないが、開発版には付与する予定である。

表 8　単語間係り受け解析の結果 (評価指標 UAS)。

train \ test	OC	OW	OY	PB	PM	PN	all.
OC	89.70	81.99	88.46	87.93	88.45	87.21	**90.49**
OW	80.21	**88.62**	78.08	83.66	84.74	84.95	88.55
OY	86.35	79.54	86.15	84.62	85.67	84.66	**88.21**
PB	89.23	86.23	88.34	**91.56**	90.91	90.63	91.48
PM	87.28	85.57	86.64	89.65	**89.74**	89.32	89.67
PN	86.40	87.66	85.88	88.65	89.31	**91.20**	90.83
all.	86.64	84.84	85.71	87.74	88.18	88.00	**89.89**

5. ジャンルごとの係り受け構造解析

UD Japanese-BCCWJ では 6 種類ものジャンルについて比較的大規模な量の UD が提供される。他の UD でも複数のジャンル収録されて UD も公開されているが、ある程度の量、数千文単位で収録されているものは少ない。UD Japanese-BCCWJ のデータの規模について検討するために、実験として単語間係り受けの解析結果を示すことにする。本稿では形態素解析の結果は示さない。理由としては、既存の形態素解析 (例えば MeCab(Kudo et al. 2004)) を用いて UniDic 品詞体系に品詞を付与することが可能であり、さらに前述のとおり、Unidc 品詞体系から UPOS に変換するのは規則ベースで簡単に変換することができるからである。

単語間係り受け解析を行うツールとして UDPipe (Straka and Straková 2017) を用いた。UDPipe では UD コーパスを元にモデルを構築、解析結果を出力できるツールである。さらに構築したモデルを用いて、単語分割、タグ付け、見出語認定、そして係り受け解析を行うことができる。係り受け解析には Parsito (Straka et al. 2015) という手法が採用されており、これはニューラルネットワークを用いた手法である。使用した UDPipe のバージョンは 1.2.1-devel を使い、オプションはつけずにトレーニング、評価を行った。実際に用いた訓練、テストデータの量は表 3 に示した通りである。評価指標としては Unlabeled attachment score (UAS) を用いた。UAS は係り元単語の係り先が合っているかを計算し、その正解割合を出したものである。

表 8 に結果を示す。表の列はそのジャンルのみで構築したモデルを表しており、行がテストに用いたジャンルのデータを表現しており、'all' はすべてのデータを使った場合を表している。つまり表示されている値は、列のジャンルで訓練したモデルに対して行のテストデータで評価した結果を表現している。

表 8 をみてわかるとおり、OW、PB、PM、PN といった 200,000 単語以上収録されているジャンルにおいては、同一のジャンルのモデルで評価した結果が評価が最も高い。一方、量が比較的少ない OC、OY(100,000 単語程度のもの) はすべてのデータで学習したものの精度が高くなっていることが分かる。そのため、必ずしも大規模な文章量があれば精度が良くなるというわけではなく、ある程度規模があれば、同一のジャンルでトレーニングしたモデルの方が精度がよくなる、といった結果を確認することができた。UD Japanese-BCCWJ を用いることでこのように、量による違い、ジャンルによる違いでの比較を行うことができることが分かる。

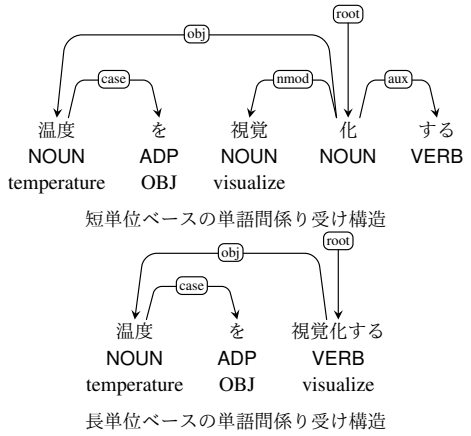

図 5　短単位と長単位の品詞体系による違いの例

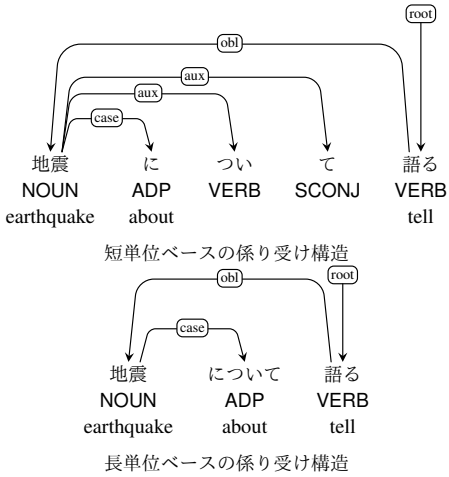

図 6　短単位と長単位における複単語表現の違い

6. 議論

この節では UD Japanese-BCCWJ において構築した際に検討する必要ができた内容などについて議論する。本稿では UD Japanese-BCCWJ を中心に説明してあり、日本語版 UD について全体的な議論については文献 (Asahara et al. 2018) で議論している。

6.1 単語認定単位について

UD における単語単位の認定は日本語版 UD において議論すべき問題のひとつである。前述の通り、BCCWJ で用いることができる単語単位には短単位、長単位、文節が存在する。現行の UD Japanese-BCCWJ では短単位を採用している。UD プロジェクトにおける単語とは、「統語的な単語 (syntactic word)」であると規定されている。UD Japanese-BCCWJ では短単位を採用しているものの、この統語的単語としては短単位よりも長単位の方が近いと考えられる。

例えば、短単位と長単位では品詞体系が異なり、これは長単位の方が syntactic word に合っている可能性がある。図 5 は短単位の場合と長単位の場合で UD にしたときの例である。短単位の場合、「可視/化/する」という語が 3 単語に分かれてしまい、それぞれ、NOUN、NOUN、VERB と UPOS をバラバラに与えられる。そのため、「可視化する」というフレーズが動詞であるかどうかを表現するのに係り関係を細かく設定する必要がでてくる。一方で長単位の場合、これは「可視化する」というひとつの単語になり、長単位は用法主義に基づく品詞であるため、「動詞」であると品詞体系からも確定する。

さらに図 6 のように、複単語表現「について」という表現も、短単位の場合は 3 つの単語で構成される一方で、長単位であればひとつにまとまってくれるため、機能語と名詞句との関係も簡素に表現できる。このように、元々長単位の品詞は構文に基づいて構成されているのもあり、UD の「統語的な単語」にあっていると考えられる。

しかし、現状は短単位を UD Japanese-BCCWJ では採用している。ひとつは長単位を厳密に解析できるツールがないこと、もうひとつの理由としては、複合表現の中でも、必ずし

も UD に合うような「統語的な単語」でない可能性があるためである。今後長単位で UD Japanese-BCCWJ を構築することで、これらの問題について検討する必要があるだろう。

6.2 並列構造

並列構造も UD、特に日本語や韓国語などで問題になっている。理由は 2 つあり、1 つ目の理由としては、日本語は主辞を右側に置く言語であるのに対して、英語は主辞となる句を左に置く言語であるため、並列構造のルールに反してしまう、という点である。2 つ目の理由として、例えば conj は名詞句の並列の並列を表現しており、名詞並列句であるか否かを考えなくてはいけないものの、UD Japanese-BCCWJ の場合、名詞並列句であるか、動詞並列句であるかの情報がない、という点である。

例えば、「と」という接続表現がある。 基本的には、英語でいう “with” の意味合いだと考えられるだろう。この with の意味合いの場合、UD では図 7 の上記の例のように nmod を付与する。しかし、必ずしもこの「と」が “with” の意味合いであるとは限らない。例えば、図 7 の中間の例のような「パンとジャム」の場合、「パンに（つける）ジャム」という意味合いが考えられるため、この「と」という接続表現は「with」の意味合いと考え nmod になると考えられる。一方で「パンとごはん」の場合、「ごはん」と「パン」とを並列に並べているだけである、と考えられるためこれは並列表現であるとみなし conj でつなげるべきである。しかしこの区別をするための情報は BCCWJ において付与されていないため、nmod でつなぐ表現であるのか、conj でつなぐ表現であるのかの区別が難しい。

また、前述のとおり日本語は「左から右にかかる」右主辞傾向の言語である。一方で英語、UD における基準では「右から左にかかる」左主辞傾向の言語である。そのため、UD の規定に従うならば図 7 の中間の例のような表現にする必要がある。しかし、現状の手順では図 7 の下部の図のような表現になってしまい、UD の規定に反してしまう。そのため左主辞の構造への変換という手順が必要となり、実直に実装することが難しいと言えるだろう。

6.3 節 (Clause)

UD の依存関係ラベルでは単語と句、節を分けるようにデザインされている。しかし、日本語では、単語、句、節との境界が曖昧である。なぜならば、日本語の文には主語も含めて、必ずしも明示的な格要素を書く必要がないためである。

図 8 に日本語における節と形容詞節の例をあげる。図 8 の上の例は名詞主題がついた形容詞節である。しかし、下の例は形容詞は修飾しているのか、叙述的であるのかが断定できない。なぜならば、日本語では、名詞叙述形容詞の名詞主題は省略できるからである。図 8 の一番下の例の場合、おそらく「しっぽ」などが補われると考えられるが、全体的に赤い猫である可能性もあるだろう。いずれであるかは、文脈から判断するしかない。このように、単純な修飾か、形容詞節であるかの区別は現状難しいため、すべての名詞句につく形容詞には acl を付与している。

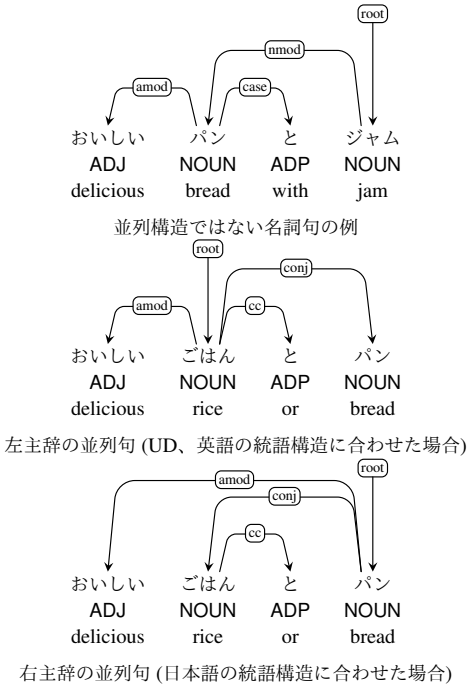

図 7　日本語における名詞句の並列構造

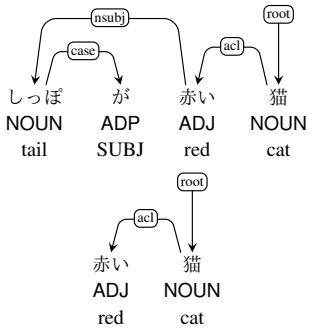

図 8　日本語における節と句の違い

7. まとめと今後の展望

本稿では現代日本語書き言葉均衡コーパス (BCCWJ) から Universal Dependencies(UD) のフレームワークに変換した UD Japanese-BCCWJ を構築した。そして、BCCWJ と UD の違いに触れ、その構築手順や特徴について説明した。UD Japanese-BCCWJ は 2018 年 4 月に公開されている [(10)]。

しかし本稿で議論したように、UD Japanese-BCCWJ あるいは日本語版 UD において検討しなくてはいけない問題点が存在する。例えば単語の単位認定が短単位であるのは UD の統語的な単語単位としてふさわしいとは言い難いため、長単位などの別の単語単位のコーパスも用意する必要があるだろう。

それぞれの日本語版 UD では、基としているコーパスが異なるために、品詞体系などの違いから、ルールがそれぞれ異なってしまっている。例えば、UD Japanese-KTC は句構造ツリーバンクから構築されており、BCCWJ の係り受け構造から変換されたものではない。そこで、今後は日本語 UD において、なるべく同一のルールで構築できるように、UD Japanese-BCCWJ で用いたルールに従って構築できるように調整を行いたいと考えている。これにより日本語 UD 間でのコーパスの差異を減らすことができると考えられる。

謝　辞

(10) http://universaldependencies.org/ にて UD Japanese-BCCWJ として配布されている。また BCCWJ の中納言アカウントを持っている場合、https://bccwj-data.ninjal.ac.jp/mdl にて変換済みのデータをダウンロードすることができる。

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) の成果である。

文　献

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Hěctor Martínez Alonso, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li (2017). “CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.” *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–19.

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki (2018). “Universal Dependencies Version 2 for Japanese.” *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1824–1831. Miyazaki, Japan.

Takaaki Tanaka, and Masaaki Nagata (2013). “Constructing a Practical Constituent Parser from a Japanese Treebank with Function Labels.” *Proceedings of 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL’2013)*, pp. 108–118. Seattle, Washington, USA.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto (2016). “Universal Dependencies for Japanese.” *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1651–1658.

Toshinobu Ogiso, Asuko Kondo, Yoko Mabuchi, and Noriko Hattori (2017). “Construction of the ‘Corpus of Historical Japanese: Meiji-Taisho Series I - Magazines’.” *Proceedings of the 2017 Conference of Digital Humanities (DH2017)*. Montrèal, Canada.

Mai Omura, Yuta Takahashi, and Masayuki Asahara (2017). “Universal Dependency for Modern Japanese.” *Proceedings of the 7th Conference of Japanese Association for Digital Humanities (JADH2017)*, pp. 34–36.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*,

48:2, pp. 345–371.

Masayuki Asahara, and Yuji Matsumoto (2016). “BCCWJ-DepPara: A Syntactic Annotation Treebank on the ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58. Osaka, Japan.

植田禎子・飯田龍・浅原正幸・松本裕治・徳永健伸 (2015).「『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.

Sadao Kurohashi, and Makoto Nagao (2003). *Building a Japanese Parsed Corpus – while Improving the Parsing System.*, Chap. 14 pp. 249–260. *Treebanks: Building and Using Parsed Corpora.*: Springer, Dordrecht.

Shinsuke Mori, Hideki Ogura, and Tetsuro Sasada (2014). “A Japanese Word Dependency Corpus.” *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 753–758. Reykjavik, Iceland.

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』 国書刊行会, pp. 101–123.

Slav Petrov, Dipanjan Das, and Ryan McDonald (2012). “A universal part-of-speech tagset.” *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC2012)*, pp. 2089–2096.

大村舞・浅原正幸 (2017).「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発表論文集, pp. 133–143.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning (2014). “Universal Stanford Dependencies: A cross-linguistic typology.” *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 4585–4592. Reykjavik, Iceland.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto (2004). “Applying conditional random fields to Japanese morphological analysis.” *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Barcelona, Spain.

Milan Straka, and Jana Straková (2017). “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe.” *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99. Vancouver, Canada.

Milan Straka, Jan Hajič, Jana Straková, and Jan Hajič jr. (2015). “Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle.” *Proceedings of Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*.