コーパスの文法と教科書の文法 言語情報付きコーパスに対する学校文法情報付与

今田 水穂(文部科学省)

1. はじめに

係り受け情報つきコーパスに対して、学校文法で使用される主語、述語、修飾語などの「文の成分」に相当する情報を付与することを考える。コーパスは常に研究者が必要とする情報を提供してくれるわけではない。コーパスには多様な情報が付与されているが、研究者の目的もまた多様であり、与えられていない情報は自分で作る必要がある。コーパスは学校文法と同様に文節単位の依存文法を採用しているが、文節や係り受けの属性について必ずしも豊かな情報が付与されているわけではない。本発表では形態論情報などの既存の情報を利用した文法情報の自動付与について説明し、また今後の展望について述べる。

2 コーパスの文法と教科書の文法

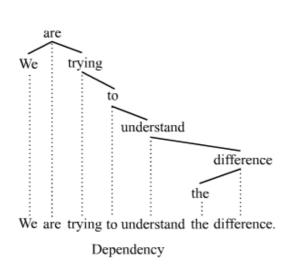
2.1 句構造文法と依存文法

文の構造を記述する方法には句構造文法と依存文法がある。句構造文法は文の構造を句や語の階層構造によって記述するもので、生成文法や HPSG など理論言語学の分野で広く使用されている。依存文法は文の構造を構成要素間の依存関係で記述するもので、学校文法で使用されている係り受け文法は依存文法の一種である。コーパスのアノテーションでは句構造文法も依存文法も使用されるが、日本語のコーパスの場合は学校文法に類似した依存文法が使用される場合が多い。

are

are

We



are trying to understand understand difference the difference.

We are trying to understand the difference.

Constituency (BPS)

図 1 依存文法と句構造文法[1]

依存文法の構造はノード(頂点)とエッジ(辺)からなる有向グラフで表現することができる。

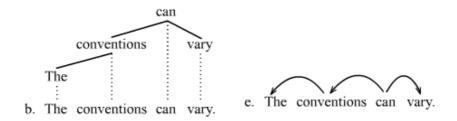


図 2 依存文法の木構造表現(左)と有向グラフ表現(右)[2]

日本語の文の構造を依存文法で表現する場合には、文節単位の依存文法と語単位の 依存文法がある。学校文法は、文節単位の依存文法である。コーパスの言語情報付与 では、文節単位の依存文法、語単位の依存文法、句構造文法のいずれも使用される。



図 3 文節単位の依存文法 (BCCWJ-DepPara — 浅原 2013, p.5)

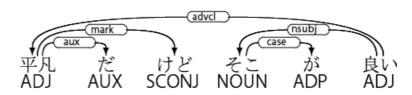


図 4 語単位の依存文法 (日本語 Universal Dependency — 金山他 2015, p.506)

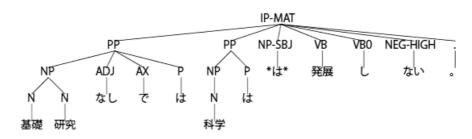


図 5 句構造文法 (KEYAKI Treebank — Butler et al. 2012, p.2)

2.2 教科書の文法

日本の国語教育で使用されている文法を、俗に学校文法と呼ぶ。教科書によって多少の差異はあるが、橋本文法をベースにした文節単位の依存文法である。学校文法では、文の構成要素を主語、述語、修飾語、接続語、独立語という 5 つの「文の成分」に分類する。また、文の成分を作る関係として、並立の関係と補助の関係がある。

表 1 学校文法における係り受け

係り受け関係	文の成分	例
主語・述語の関係	主語、述語	<u>猫が</u> いる
修飾の関係	修飾語	庭に いる (連用修飾語)
		たくさんの 猫 (連体修飾語)
接続の関係	接続語	<u>猫だから</u> かしこい
		猫だ。 <u>だから</u> かしこい
独立の関係	独立語	<u>あっ</u> 猫だ。
並立の関係		<u>犬と 猫</u>
		<u>食べて</u> <u>寝る</u>
補助の関係		<u>遊んで</u> <u>いる</u>

文の成分のうち、主語、連用修飾語、接続語、独立語は述語に係る成分である。



述語も、述語に係る場合がある(連用節)。述語の連用用法は、接続語とされる。



述語ではなく名詞に係る成分は、節か語句かによらず連体修飾語とされる。

庭に いる たくさんの 猫

連体修飾語 連体修飾語

また、「文の成分」を作る関係として、並立の関係と補助の関係がある。



まとめると、学校文法は述語に係る従属連用成分(主語、連用修飾語、接続語、独立語)の分類が中心であり、それに従属連体成分(連体修飾語)や等位成分(並立の関係)などを加えた構成となっている。

2.3 コーパスの文法

ここでは「現代日本語書き言葉均衡コーパス」の係り受け・並列構造アノテーション(BCCWJ-DepPara)について言及する。BCCWJ-DepParaでは、係り受け関係は文節単位の依存文法で記述され、並列・同格関係は係り受けとは別のレベルの情報として記述される。係り受けの種類は通常の係り受け、係り受けなし、文末、文節の連結(文節境界の修正)の4種類であり、並列・同格関係は並列、同格、具体例・総称同格の3種類である。

表 2 BCCWJにおける係り受け・並列関係

タグ	説明	例
D	通常の係り受け	<u>庭に</u> <u>猫が</u> いる
F	係り受けなし	<u>あっ</u> だから いるのか (感動詞、接続詞)
		<u>「</u> 猫が いる 」 (記号など)
Z	文末	猫が <u>いる</u>
В	文節の連結	猫が <u>遊んで</u> <u>いる</u>
Parallel	並列	<u>犬</u> と <u>猫</u>
Apposition	同格	<u>主人公</u> <u>のび太</u> が
Generic	具体例·総称同格	<u>ドラえもんなど</u> <u>猫</u> が

通常の係り受けは、従位接続一般を表し、学校文法の主語、連用修飾語、連体修飾

語、接続語(従属節)などを包括する。係り受けなしは、主に接続詞、感動詞、記号類である。接続詞は、文および節の単位が未確定であるために、現時点では係り受けなしとされる。文末は文字通り文末の要素に付与されるが、文中に引用文が埋め込まれている場合などは文中の要素に付与される場合もあり得る。係り先なしと文末は、形式上はダミーノードへの係り受けとして表現される。



文節の連結は、単一文節とすべき文節が2文節に分割されている場合に付与される。 学校文法の補助の関係は、BCCWJでは単一文節として扱われるので、このラベルが 付与される。



並列は、学校文法の並立の関係に相当する。ただし、BCCWJでは並列は係り受けとは別のレベルの情報として扱われており、係り受けとしては通常の係り受け(D)が割り当てられる。また、BCCWJの並立関係は名詞の並立のみを扱い、述語の並立に対しては付与されない。



同格は、学校文法では扱われない。BCCWJでは、同格と具体例・総称同格の2種類の同格関係が扱われる。並列関係と同様、係り受けとは別のレベルの情報であり、係り受けとしては通常の関係(D)が割り当てられる。

3. アノテーション

3.1 アノテーションの設計

タグセットは学校文法に準拠し、主語、修飾、接続、独立、並立、補助を用いる。 また、修飾については連用修飾と連体修飾の区別を追加する。述語は、係り元ではな く係り先の属性なのでタグセットには含めない(係り元としては接続や連体修飾とな る)。また、学校文法がカバーしない属性を補完するため、既存コーパスの係り受け関 係スキーマから係り受けなし、文末、その他を追加する。

表 3 学校文法アノテーションのタグセット

タグ	例	
主語	<u>猫が</u> いる。	
連用修飾	<u>庭に</u> いる。	(連用修飾句)
連体修飾	<u>庭の</u> 猫	(連体修飾句)
	かしこい 猫	(連体修飾節)
接続	猫だから かしこい。	(連用修飾節)
	<u>だから</u> かしこい。	(接続詞)
独立	<u>あっ</u> 猫だ。	(感動詞)
並立	<u>犬と猫/犬も</u> 猫も/ <u>犬から</u> 猫まで	(名詞の並立)
	<u>食べて</u> 寝る。	(述語の並立)
補助	遊んでいる。	
係り先なし	<u>第1章</u> 猫	
係り先なし 文末		

3.2 人手によるアノテーション

アノテーションの方法としては、人手でアノテーションする方法と、既存の言語資源を利用して自動的にアノテーションする方法がある。人手によるアノテーションでは、作業環境を用意する必要がある 1 。今回のアノテーションでは Excel ワークシートを利用する。

¹ 作業内容によってはアノテーション作業専用のツールを使う必要がある。例えば、ChaKi.NET は文節境界、係り受け、並列関係などの編集ができる。Tagrin はテキスト中の任意の範囲にセグメントとリンクを付与することができ、述語項構造アノテーションなどに使用されている。



図 6 作業用の Excel ワークシート

複数の作業者が同一ファイルのアノテーションを行う。作業結果の評価は、作業者間の評定の一致度を示す κ 係数により行う。 κ 係数は、2 人の評定が偶然一致する率 p_e と、実際に一致した率 p_o を用いて以下の式で計算する。

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

κ係数の意味は、直感的には次の図で理解することができる。

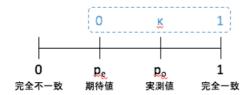


図7 к係数

また、複数の作業者による作業結果は、多数決によって単一ファイルに統合し、成果物として利用する。

3.3 規則によるアノテーション

アノテーションの自動化には、規則によるものと、統計的手法によるものがある。 規則によるアノテーションでは、既存の言語資源によって得られる情報から、規則に 基づいて新たな情報を付与する。利用可能な既存の言語資源としては、形態論情報、 係り受け・並列関係(BCCWJ-DepPara)、述語項構造(BCCWJ-PAS)などがある。以下 に例を示す。

X 1 301 O Linguistani O C 1 KAKI / / V 1 V								
形態論	出現形	犬	と	猫	<u>が</u>	いる	0	
	品詞	名詞	格助詞	名詞	格助詞	動詞	補助記号	
	活用形	*	*	*	*	終止形	*	
係り受け	文節 ID	0 1		1		2		
	係り先 ID			2		·1 Z		
	係り受け	D		D				
並列	並列		<u>Parallel</u>		Parallel			
述語項構造	述語項構造		,			述語		
学校文法		並列		主語		文末		

表 4 既存の言語資源を利用した学校文法アノテーション

まず形態論情報を参照する。文節の接続は、原則として文節の最右要素の接続形式によって決定される。最右要素が名詞の場合、接続は確定されない。副詞、連体詞、接続詞、感動詞の場合は、それぞれ連用修飾語、連体修飾語、接続語、独立語になる。活用語(動詞、形容詞、形容動詞、助動詞)の場合は活用形に依存する。連用形、連体形、終止形はそれぞれ接続語、連体修飾語、文末になる。助詞は、格助詞、副助詞は連用修飾語、接続助詞は接続語、終助詞は文末の場合が多い。ただし格助詞は「が」なら主語、「の」「との」「という」なら連体修飾語になり、副助詞も「は」などは主語になる場合がある。

次に係り受け情報を参照する。係り受けなし(F)の文節には、接続詞、感動詞、記号類があるが、接続詞、感動詞はそれぞれ接続語、独立語とするので、それ以外のものを係り受けなしとする。文末(Z)は、文末とする。文節の連結(B)は、補助の関係に相当するものとそれ以外のものを含む。形態論情報で接続(述語の連用用法)と判定されたものを補助の関係に書き換え、それ以外のものはその他にする。ここまでで接続の特定されていない文節については、係り受け関係にある別の文節の属性を参照して接続を特定する。係り先が名詞であれば連体修飾語であり、述語であれば連用修飾語である。

並列(Parallel)は、複数ある並列要素のうち、最右以外のものに並列タグを付与する。 同格(Apposition)、具体例・総称同格(Generic)についてはこのアノテーションではサポートしない。連体修飾語と見なせるものは連体修飾語とし、それ以外のものはその 他とする。述語項構造は「は」など副助詞を伴う文節について、主語と連用修飾語の 区別をするために使用する。

これらの一連の処理を実行するために、kokugo_tagger という名前の Ruby スクリプトを作成した。このスクリプトは CaboCha 形式のファイルを入力、拡張 CaboCha 形式のファイルを出力する。



図 8 kokugo_tagger の実行画面

解析結果の評価は、予め用意した正解データと対照し、正解率(accuracy)、再現率 (recall)、適合率(precision)、F値(F-score)などを求めることによって行う。ここでは、人手で作成したデータがまだ存在しないため、割愛する。

3.4 統計的手法によるアノテーション

統計的手法によるアノテーションとして、ここでは YamCha²を使用する方法を示す。YamCha は SVM(Support Vector Machine)を使用した汎用テキストタガーで、品詞付与、固有表現認識、文節境界付与などの目的で使用される。学習と評価のために、予め一定の量の正解データを作成しておく必要がある。

まず、図 9 のようなファイルを作成して YamCha に学習させ、破線部分の情報を変数として実線部分のラベルを予測するモデルを作成する。B は文節頭、I は文節頭以外を意味する。

² http://chasen.org/~taku/software/yamcha/

詰め	動詞	連用形	В	修飾(連体)
将棋	名詞	*		修飾(連体)
o o	助詞	*		修飾(連体)
本	名詞	*	В	補語
を	助詞	*	1	補語
買2	動詞	連用形	В	接続
τ	助詞	*	I	接続
き	動詞	連用形	В	文末
まし	助動詞	連用形	1	文末
た	助動詞	終止形	1	文末
<u>ه</u>	補助記号	*	1	文末

図 9 YamCha の学習用データ

このモデルを使用し、タグが付与されていないファイルに対して自動的にタグを付与する。以下は、評価のために既にタグが付与されている正解データに対して自動付与を行ったものである。5 列目が正解データ、6 列目が自動付与結果である。文節末形態素(図中の破線)に付与されたタグを、その文節の属性と見なす。

	空白	*	В	主語	補語
それ	代名詞	•	1	主語	修飾(連用)
かく	助詞	*	1	主語	並立
私	代名詞	•	В	主語	修飾(連用)
が	助詞	*	1	主語	主語
マザー	名詞	•	В	補語	補語
を	助詞	•	1	補語	補語
見	動詞	連用形	В	修飾(連体)	修飾(連体)
<i>t=</i>	助動詞	連体形	1	修飾(連体)	修飾(連体)
最初	名詞	*	В	文末	文末
だっ	助動詞	連用形	1	文末	文末
<i>t</i> =	助動詞	終止形	1	文末	文末
s.	補助記号	•	1	文末	文末

図 10 YamCha の解析結果

解析結果の評価は、規則によるアノテーションの場合と同様、正解率(accuracy)、再現率(recall)、適合率(precision)、F 値(f-measure)によって行う。正解率は、正解データと解析結果が一致した率である。図 10 の場合は、5 文節中 4 文節が一致しているので正解率は 0.8 である。

再現率は解析結果が正解データを再現できた率、適合率は解析結果による予測が正解データと一致していた率であり、F 値は再現率と適合率の調和平均(F=2rp/(r+p))である。図 10 の「主語」の場合は、正解データで主語とされる 2 文節のうち、解析結

果も主語なのは 1 文節なので再現率は 0.5、解析結果で主語とされる 1 文節のうち、正解データでも主語なのは 1 文節なので適合率は 1.0 であり、F 値は $2\times0.5\times1.0/(0.5+1.0)$ で 0.66 である。

ここでは、人手で作成したデータがまだ存在しないため、参考として kokugo_tagger で作成したデータを使用し、50 ファイルで学習して別の 50 ファイルで評価した結果を示す。全体の正解率は 0.9365 だった。ラベル別の再現率、適合率、F 値は以下の通りである。 true、positive はそれぞれ、正解データ、解析結果に該当のラベルが付与されていた頻度を表し、true-positive は両者に一致して該当のラベルが付与されていた頻度を表す。使用した kokugo_tagger のバージョンの都合により、ラベルの種類は 3.1 節表 3 で示したタグセットとは異なる。

丰	5	VomCho	17 F	ス	解析結果の評価
ᅏ	ก	Yamu;na	W -	\sim	

label	true	positive	true-positive	recall	precision	f-measure
主語	1518	1492	1443	0.9506	0.9672	0.9588
補語	3555	3748	3432	0.9654	0.9157	0.9399
連用修飾	6102	5859	5393	0.8838	0.9205	0.9018
連体修飾	6700	6680	6652	0.9928	0.9958	0.9943
接続	2584	2594	2521	0.9756	0.9719	0.9737
独立	41	37	32	0.7805	0.8649	0.8205
並立	1160	1201	847	0.7302	0.7052	0.7175
同格	39	13	0	0.0	0.0	NaN
係りなし	410	371	318	0.7756	0.8571	0.8143
文末	3205	3206	3184	0.9934	0.9931	0.9933
文節連結	168	179	157	0.9345	0.8771	0.9049
その他	209	311	80	0.3828	0.2572	0.3077
合計	25691	25691	24059	0.9365	0.9365	0.9365

4. 課題と展望

4.1 作成したデータの利活用

このアノテーションで作成したデータは学校文法に準拠しているので、国語教育分野での利用に適している。図 11 は「児童・生徒作文コーパス」(小学校・中学校の児童・生徒の作文を収集したコーパス) を CaboCha と kokugo_tagger で解析し、文の成分の学年別の推移を調べたものだが、学年が上がるにつれて連体修飾成分の使用頻度が増えていくことが分かる(宮城・今田 2015)。現在、このコーパスの一部(10万短単位程度)について人手による形態論・係り受け情報付与の作業が進められているので、今後このデータについて再度学校文法アノテーションをしたい。



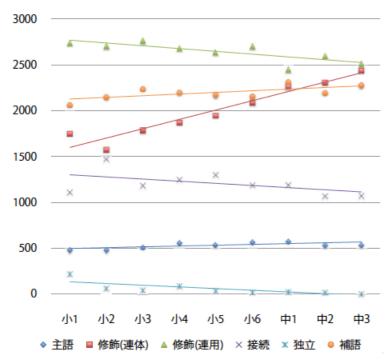


図 11 「児童・生徒作文コーパス」における文の成分の構成比(宮城・今田 2015)

また、主語、連用修飾語、連体修飾語などの要素は、文法研究で一般に用いられる概念であり、汎用的に利用することができる。例えば、コーパスから主語と述語の対や、連体修飾語と非修飾語の対をリストアップしたいときに、これらの情報を付与したコーパスや自動付与するシステムがあれば、独自に主語の判別や連体と連用の判別などをする必要が無くなり、研究者間で同一の言語資料を共有したり、追試したりすることが容易になる。

4.2 アノテーションの拡張と応用

アノテーションの拡張や応用の可能性として、次のようなものが考えられる。

- 従位接続と等位接続
- 連用と連体
- 述語と非述語
- 述語の下位分類(動詞述語、形容詞述語、名詞述語)
- 節の下位分類
- 連用修飾語の下位分類(補語、修飾語、状況語など)
- 連体修飾語の下位分類(属格の用法分類など)

従位接続と等位接続、連用と連体、述語と非述語は、タグセットの体系性に関わる。

学校文法に準拠したタグセットは利便的ではあるが、理論的に十分体系的とは言えない。まず、接続には従位接続と等位接続があるが、従位接続が主語、修飾語、接続語、独立語などに細分化されているのに対して、等位接続は特に区別なく並立の関係としてまとめられている。次に、従位接続には連用と連体があるが、連用が主語、修飾語、接続語、独立語などに細分化されているのに対して、連体は連体修飾語にまとめられている。また文の成分の分類において節と句の区別は重視されておらず、連用修飾節は接続語に、連体修飾節は連体修飾語に包摂される。研究の目的によっては従位・等位、連用・連体、述語・非述語(節・句)などの属性が付与されている方が、有用性が高い。

述語的成分については、2 つの拡張が考えられる。1 つは述語の下位分類であり、もう1つは節の下位分類である。述語の下位分類は、教科書でも言及される場合があるが、動詞述語、形容詞述語、名詞述語の3つがある。名詞に限らず、他の非活用語も助動詞「だ」など繋辞(コピュラ)を伴って述語として機能することがあるため、総称してコピュラ述語と呼ぶ場合もある。ただしコピュラは省略される場合も多く、名詞述語の判定は簡単ではない。そのため、名詞述語を特定する情報が利用可能であることは有用である。節の下位分類は、節境界認定の研究(佐藤・丸山 2015)が進んでいるので、いずれ利用可能になることを俟ちたい。

非述語的成分については、連用的成分、連体的成分について、それぞれ下位分類する拡張が考えられる。日本語記述文法研究会(2010)は文の成分を一般的な学校文法における分類より細かく分類しているが、そのうち連用修飾語に相当するものとしては補語、修飾語、モダリティ修飾語、状況語が挙げられており、連体修飾語は規定語という名称が与えられている。名詞の意味を研究するにあたって、補語が特定されていること、属格「の」の用法が下位分類されていることなどは重要であると考えられ、今後、検討したい。

5. まとめ

学校文法およびコーパスの係り受け情報が文節単位の依存文法であることについて 説明し、コーパスに学校文法の情報を人手、規則、統計的手法でアノテーションする 方法について説明し、作成したデータの利活用および拡張・応用の可能性について述 べた。今年度中に人手データを作成し、自動付与システムの開発、評価や、成果物の 利活用を進めたい。

参考文献

浅原正幸(2013)「『現代日本語書き言葉均衡コーパス』係り受け・並列構造アノテーション作業メモ Version 0.6.0」Technical report.

金山博・宮尾祐介・田中貴秋・森信介・浅原正幸・植松すみれ(2015)「日本語 Universal Dependencies の試案」『言語処理学会第 21 回年次大会発表論文集』505-508. 黒橋禎夫・居倉由衣子・坂口昌子(2000)「形態素・構文タグ付きコーパス作成の作業 基準(version 1.8)」Technical report.

- 佐藤理史・丸山岳彦(2015)「節境界認定に関する諸問題」第8回コーパス日本語学ワークショップ,国立国語研究所.
- 日本語記述文法研究会[編](2010)『現代日本語文法 1』くろしお出版.
- 宮城信・今田水穂(2015)「児童・生徒作文コーパスの設計」第7回コーパス日本語学ワークショップ,国立国語研究所.
- Butler, Alastair, Tomoko Hotta, Ruriko Otomo, Kei Yoshimoto, Zhen Zhou, Hong Zhu (2012) "Keyaki Treebank: Phrase Structure with Functional Information for Japanese", テキストアノテーションワークショップ, 国立情報学研究所.

図表引用

- [1] https://en.wikipedia.org/wiki/File:Wearetryingtounderstandthedifference_(2).j pg, Author: Tjo3ya (CC BY-SA 3.0)
- [2] https://commons.wikimedia.org/wiki/File:Conventions.jpg, Author: Tjo3ya (C C BY-SA 3.0)

これは、筑波大学応用言語学研究会の発表原稿作成規定です。予稿集の体裁統一のため、作成規定に従って原稿を作成していただきますよう、ご協力をお願いします。原稿は、研究会係でページ番号の挿入などの編集をするため、Microsoft Word や、テキスト形式で保存されたデータでご入稿ください。3また、文字化けの可能性がある場合は、PDFもあわせてお送りください。

(1) ページ数

講演:20ページ以内、研究発表:8ページ以内。

- (2) タイトル、所属・氏名、本文の基本レイアウト
 - a. 余白: MS Word の「標準」を使用 (上: 35.01mm、左右下: 30mm)。
 - b. 1ページあたりの文字数と行数:38字/38行。(?)
 - c. タイトル: MS ゴシック 14 ポイント (副題は 13)、中央揃え。
 - d. 所属・氏名: MS 明朝: 10.5 ポイント、右揃え。
 - e. 本文: MS 明朝 11 ポイント、両端揃え。英数字は Century。
 - f. 脚注:ページ下。10.5 ポイント。
 - g. 参考文献:本文と同様。
 - h. 図表:白黒印刷可能なものを使用。

原稿締切日:2015年9月21日(月)

原稿提出の際、以下の質問について、あわせてお知らせください。 原稿送り先:郡司結花(yuka.w.m.gunji@gmail.com)

よろしくお願いいたします。

³ この原稿作成規定は、昨年度からの新しい試みのため、不十分な点があるかと思います。 お気づきの点がございましたら、ご遠慮なくお知らせください。

- ① 発表に Power Point を使用 (する・ しない)。
- ② 発表に Power Point を使用する場合、そのための追加資料を配布 (する・しない)。
- ③ 発表に Power Point を使用する場合、パソコンは (自分(もしくは発表者同士) ・ 事務室) のものを使用する。
- ④ 発表原稿 PDF の応用 HP への掲載を希望 (する ・ しない)。 (掲載は、研究会終了後となります。)
- ⑤ その他、何か伝えたいこと、またお気づきの点などがありましたら、お知らせく ださい。