「象は鼻が長い」構文が使われるのはいつか：文内要因の検討

「象は鼻が長い」構文が使われるのはいつか：文内要因の検討

吉田 真大 (千葉大学大学院人文公共学府)*
傳 康晴 (千葉大学大学院人文科学研究院)

When you use the construction "Zou ha hana ga nagai.": Within-sentence factor

Mahiro Yoshida (Graduate School of Humanities and Studies on Public Affairs, Chiba University)
Yasuharu Den (Graduate School of Humanities, Chiba University)

要旨

本研究では、複数通りの表現が想定されるときに、どのような要因が選択に影響するのか、について検討した。構文レベルに焦点を当て、「象は鼻が長い」(「A ハ B ガ P」構文) と「象の鼻が長い」構文 (「A ノ B ハ P」構文) とを比較して、それらの文内における要因の存在について考察した。UD Japanese-BCCWJ に収録された実例からモデルを構築してあてはめた。その結果、B に相当する部分の長さ、文後方に位置する修飾部分、文の主辞 P の品詞の変数が選択に特に強い影響を与えていることが示唆された。

1. はじめに

同じことを表すために、複数通りの言い方をすることができる。例えば、語のレベルでは同義語という言葉があり、自明であるといえよう。「良い天気であること」を「晴れ」と表すこともできれば、「晴天」と表すこともできる。他にも、「学校」と「学び舎」、「机」と「デスク」など、二つの語の関係は様々だとしても、外延がほとんど同じである語は豊富にあるだろう。例えば語用論のレベルでも、本来の意味に由来するニュアンスの差はありつつも、複数通りの表現をすることができる。レストランを想定し、「今食べている料理がもの凄く美味しいこと」と表したいと思ったとする。「これおいしいね」「これまた食べに来たいね」「こんなの食べたことない」など、同じ意図を仮定しても様々な言い方がある。

これらは同じ意味や、意図を伝えるためにいくつかのアプローチがあるという例である。では命題ではどうだろうか。例えば「太郎と次郎が、太郎が兄、次郎が弟の兄弟関係にあること」という命題を表現するとき、「太郎には次郎という弟がいる」「次郎は二人兄弟の弟で、兄の名は太郎である」など、表す意味は共通で形式の異なる複数の表現を想定することが可能だろう。これらの表現は、原則的に交換可能であると言って良いように感じられる。しかし、これらはどのような理由をもって使い分けられているのだろうか。

本研究では「象は鼻が長い」構文について取り上げる。「象は鼻が長い」といえば、三上 (1960) で取り上げられて以来、特殊な構造をもつ構文として研究対象とされてきた。しかし、その意味構造を取り上げられたり、助詞「は」の分析に用いられることはあっても、利用され

* 24dm1125(α)student.gs.chiba-u.jp

る要因について検討した研究は、管見の限りない。

本研究では、同じ命題を示す 2 つの構文「象は鼻が長い」構文をその言い換えといえる「象の鼻が長い」構文と比較し、それぞれの構文がどのような要因をもって利用されるのかについて、UD Japanese-BCCWJ に含まれる実例データを用いて検討した。また、「筆者が○○を強調したかったから/主張したかったから」といった質的な分析ではなく、客観的に定義することが可能な説明変数を用いたモデルを構築し、それを分析した。

2. 方法

2.1 データ

本研究では、『現代日本語書き言葉均衡コーパス』(BCCWJ: Balanced Corpus of Contemporary Written Japanese) (Maekawa et al. 2014) のコアデータに対して、Universal Dependencies（以下、UD）(McDonald et al. 2013) の枠組みに則ったアノテーションを付与したコーパスである、UD Japanese-BCCWJ(Asahara et al. 2018) を用いた。

2.2 データ抽出

本節では、本研究の対象である 2 つの構文（「象は鼻が長い」「象の鼻は長い」）を認定し、コーパスからサンプルを抽出する条件について述べる。

説明に先んじて、各構文をより一般的にした呼称を導入する。以下では「象は鼻が長い」構文を「A ハ B ガ P」構文、これに対して「象の鼻が長い」構文を「A ノ B ハ P」構文と呼ぶこととする。

2.2.1 「A ハ B ガ P」構文

まず「A ハ B ガ P」構文について、具体例を（1）と図 1 に示しながら記述する。なお、図中の矢印は文節ごとの依存関係を示す。矢印上に書かれた文字はその種別（UD における dep_type）を示す。依存関係ラベルセットについて、詳しくは浅原ほか (2019) を参照されたい。

(1) 中小零細企業取材は数多く手がかかる割に、大手に比べるとインパクトが小さい。

【出典】 sent_id: PB56_00007-51

当該構文と認定する条件を以下のように設定した。まず、「A ハ B ガ P」構文の A に相当する助詞「は」を伴う文節（例における「取材は」）と、B に相当する助詞「が」を伴う文節（例における「インパクトが」）とが、この順で現れることである。そして、それに加えて、どちらも文の主辞を含む文節（例における「小さい」）に他の格要素なしで係っていること、を条件として設定した。また、これらの依存関係についても、助詞「は」を伴う文節は dislocated（転位された要素）、助詞「が」を伴う文節は nsubj（名詞句主語）であることを求めた。

そして浅原ほか (2019) の記述を参考にして、その他の要素について許容する事項について

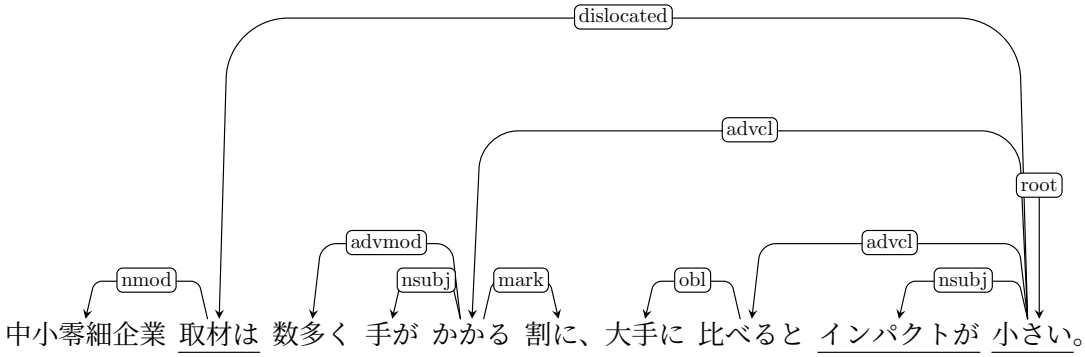


図 1　（1）における文節レベルの依存関係

設定した。助詞「は」を伴う文節と助詞「が」を伴う文節の 2 つ以外に、いくつかの要素について文中に現れることや主辞に係ることを許容した。まず、節レベルの依存関係としては、advcl（副詞的修飾節）、advmod（副詞修飾語）である。名詞句レベルの依存関係としては、形容詞的要素である nmod（名詞修飾語）、acl（形容詞的修飾節）、amod（形容詞修飾語）、det（限定詞）である。合計 6 種類の要素については、主辞に係っているサンプルも含めた。

最後に、抽出した「A ハ B ガ P」構文のサンプルのそれぞれについて、「A ノ B ハ P」構文を用いて表現することができるかどうか、目視によるチェックが実施された。交換可能でないと判断されたサンプルは対象から削除された。交換可能合計 302 件のサンプルを「A ハ B ガ P」構文として対象とした。

2.2.2　「A ノ B ハ P」構文

次に、「A ノ B ハ P」構文について述べる。具体例は（2）と図 2 によって示す。

(2)　その結果、工業材料による建築物が主流となり建築士の木造への理解は低くなっていった。

【出典】 sent_id: OW6X_00123-60

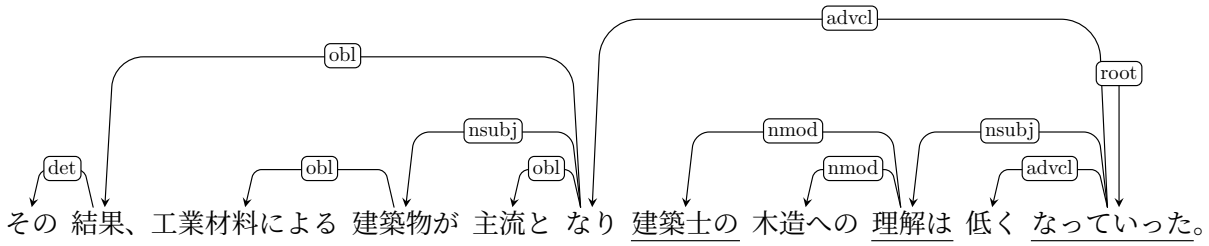


図 2　（2）における文節レベルの依存関係

当該構文と認定する条件として、以下のように設定した。まず、助詞「の」を伴う文節（例における「建築士の」）が助詞「は」を伴う文節（例における「理解は」）に係っていることで

ある。次に、その「は」を伴う文節が文の主辞を含む文節（例における「なっていった」）に他の格要素なしで係っていること、とした。当該構文において、A は助詞「〜の」を伴う文節に係る部分、B は助詞「〜は」を伴う文節に係る部分、P は文の主辞となる語をさすこととする。なお、助詞「〜の」を伴う文節を複数含む場合には、一番先頭に位置するもの（主辞の word_id が最も小さいもの）を A とした。これを統一したのは、交換可能性を保つためであった。

「A ハ B ガ P」構文と同様に、依存関係についても条件を設定した。助詞「の」を伴う文節は nmod（名詞修飾語）、助詞「は」を伴う文節は nsubj（名詞句主語）であることを条件とした。

その他の要素については、前項と同様に、依存関係が advcl、advmod、nmod、acl、amod、det である要素が文中に現れている、もしくは文の主辞に係っているサンプルも含めた。

最後に、抽出した「A ノ B ハ P」構文のサンプルのそれぞれについて、「A ハ B ガ P」構文を用いて表現することができるか、目視によるチェックが実施された。交換可能でないと判断されたサンプルは対象から削除された。合計 448 件のサンプルを「A ノ B ハ P」構文として対象とした。

2.3 モデル

本研究では、「各サンプルを構成する要素が、どのような要因によって各構文として構成・表現されるのか」を明らかにするため、文中の各要素が与えられたときに 2 つの構文のどちらを選択するかというモデルについて検討した。

本研究では、以下で述べる 11 の変数からなる変数群を説明変数、選択された構文（「A ハ B ガ P」構文は 0、「A ノ B ハ P」構文は 1 として計算した）を応答変数とした一般化線形モデル（Generalized Linear Model, 以下 GLM）を用いてモデルを構築した。モデルを用い、説明変数の中でも構文の選択に強い影響がある変数について検討した。なお、モデルの当てはめには R の glm 関数を用いた。また、誤差構造には二項分布を用いた。交互作用は本モデルにおいては除外された。

モデル構築にあたり、前項でコーパスから抽出して得られたサンプルを、2 つの構文の比率を統制した層化抽出によって分割した。750 件のサンプルのうち、601 件をモデルの学習用データとして、149 件を評価用データとして用いた。

2.3.1 要因・説明変数

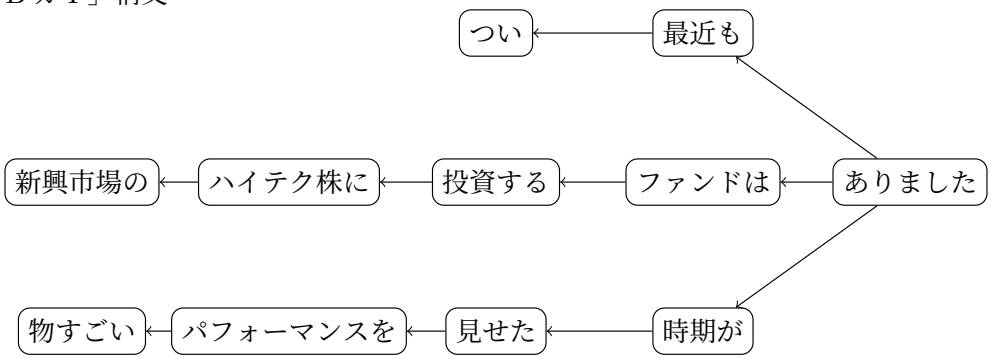
本研究では、各サンプルが有する特徴による要因（以下文内要因）について検討した。本研究で取り上げる文内要因は「長さ」「修飾成分」「品詞」「特徴語」の 4 つである。モデルではそれらの要因を表現した合計 11 の変数を用いた（表 1 参照）。

まず、1 つ目の要因として「長さ」に着目した。構文選択時の要因として、A や B の何らかの長さが関係している可能性が挙げられる。そこで、本研究ではそれぞれの文字数を A の長さ、B の長さとし、変数として用いた。なお、A、B の範囲については、依存関係を用いて表現した木構造から部分木を抽出することによって定義した。具体例は図 3 を参照されたい。原則的には、A、B の主辞を根とする部分木の全体を A、B とした。ただし、「A ノ B ハ P」構

文の場合について、「〜の〜の」と、助詞「の」を含む文節が連続する場合には、最も左側に位置する文節の主辞を A の主辞として、A を定義した。図 3 の「A ノ B ハ P」構文では、”クール・ビューティー”が A の主辞である。

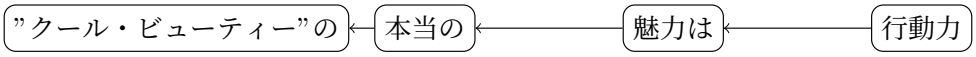
モデルにおいては、各文における A、B の文字数をそれぞれ len_treeA、len_treeB として変数に設定した。例示した図 3 の「A ハ B ガ P」構文では、len_treeA が 20、len_treeB が 18 である。

「A ハ B ガ P」構文



【出典】sent_id: PM23_00013-41

「A ノ B ハ P」構文



【出典】sent_id: PM11_00013-93

図 3　部分木の範囲例

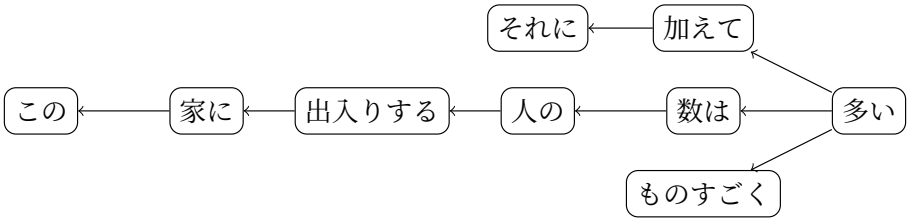
2 つ目の要因として、「修飾成分」に着目した。修飾成分がいくつ、どのくらいの長さで主辞にかかっているか、が構文選択に影響しているという可能性があると考えられた。本研究では、主辞に係っている要素のうち、先述の A、B 以外の部分を修飾成分とした。またその位置によって、A と B よりも左側なら pre、と A と B よりも右側であれば post、と二つに分類した。例えば、図 4 における「それに加えて」は pre、「ものすごく」は post である。

モデルにおいては、pre、post が何個ずつ主辞にかかっているかを、n_pre、n_post、その文字数を len_pre、len_post として表現した。図 4 の例では、n_pre、n_post は共に 1、len_pre は 7、len_post は 5 である。

3 つ目の要因として「品詞」に着目した。A、B の主辞は抽出の手続き上、基本的に名詞に限られる。しかし、文の主辞である P には制限がなく、その品詞が構文選択に関係している可能性があると考えられた。

モデルでは、pos2_P として、文の主辞の品詞が、名詞類・動詞・形容詞のいずれかというカテゴリー変数を設定した。なお、ここで用いた名詞類とは、副詞や補助記号など、動詞・形容詞のいずれでもないものを含めたものである。

4 つ目の要因として「特徴語」に着目した。本研究では、特定の語が各要素の主辞として現れたときに構文選択に影響を与えている可能性が想定された。特徴語を決定するため、2 つの



【出典】sent_id: PB39_00023-188（一部改変）

図 4　修飾成分の例

構文の A、B、P の各語の頻度（ただし書字出現形ではなく語彙素での頻度）を計算した。構文間で頻度の差が 10 以上あるものを特徴語として設定した。特徴語は以下の通りである。A では特徴語としたものはなかった。B では、「事」の 1 語を特徴語とした。P では、「有る」「多い」「する」の 3 語を特徴語とした。

モデルにおいては、B や P が各特徴語と等しいかどうか、という二値の変数として、feature_事、feature_有る、feature_多い、feature_する、の 4 つの変数を設定した。

表 1　説明変数群

要因	説明変数	概要
長さ	len_treeA	文字数（部分木 A）
	len_treeB	文字数（部分木 B）
修飾成分	n_pre	修飾成分の数（pre）
	n_post	修飾成分の数（post）
	len_pre	文字数（pre）
	len_post	文字数（post）
品詞	pos2_P	P の品詞（名詞類/動詞/形容詞のいずれか）
特徴語	feature_事	B の主辞が「事」かどうか
	feature_有る	P の主辞が「有る」かどうか
	feature_多い	P の主辞が「多い」かどうか
	feature_する	P の主辞が「する」かどうか

3.　結果

GLM の結果は、表 2, 3 に示した。表 2 は 6 つの量的変数（len_treeA, len_treeB, n_pre, n_post, len_pre, len_post）と、4 つの二値変数（feature_事, feature_有る, feature_多い, feature_する）についてである。

なお、3 水準以上のカテゴリー変数である pos2_P については、まず R の drop1 関数を用いて GLM の結果に基づいて、尤度比検定を行った。表 3 はその結果である。表 3 のとおり、有

意差が認められた。そのため、さらに R の glht 関数を用いて Tukey の補正による多重比較を実施し、水準間の比較を行った。多重比較の結果は、表 4 に示した。

11 の説明変数のうち、6 つの変数において（len_treeB, n_post, len_pre, len_post, pos2_P, feature_有る）有意差が認められた。

また、評価のために、当該モデルに 149 件の評価用サンプルを適用し、構文選択を予測させた。その結果、正答率は 81.8% であった。

表 2　GLM の結果（量的変数, 二値変数）

	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	2.493	0.310	8.033	9.52e-16	***
len_treeA	0.004	0.018	0.223	0.824	
len_treeB	-0.279	0.035	-7.881	3.24e-15	***
n_pre	-0.168	0.310	-0.542	0.588	
n_post	1.460	0.371	3.936	8.29e-05	***
len_pre	0.025	0.013	1.967	0.049	*
len_post	0.071	0.031	2.295	0.022	*
feature_事	-0.703	1.346	-0.522	0.601	
feature_有る	-1.223	0.563	-2.173	0.030	*
feature_多い	-0.446	1.267	-0.352	0.725	
feature_する	-0.362	1.150	-0.315	0.753	

注: Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

表 3　pos2_P に係る尤度比検定の結果

	DF	Deviance	AIC	LRT	Pr(>Chi)	
<none>		490.12	516.12			
pos2_P	2	526.30	548.30	36.175	1.395e-08	***

注: Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

表 4　pos2_P に係る多重比較の結果

	Estimate	Std. Error	z value	Pr(> |z|)	
動詞 - 名詞類	-1.373	0.279	-4.924	<1e-04	***
形容詞 - 名詞類	-1.755	0.360	-4.869	<1e-04	***
形容詞 - 動詞	-0.382	0.338	-1.130	0.492	

注: Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

4. 考察

本項では、3 節において有意差の認められた 6 つの説明変数のうち、主な 3 つについて考察する。

4.1 len_treeB

len_treeB の Estimate は-0.279 であった。これは、部分木 B の長さが長ければ長いほど「A ハ B ガ P」である可能性が高まる、と解釈可能である。これを説明する一つの可能性が挙げられる。助詞「は」による主題化をする際、その対象があまりに長くなることが避けられるのではないかという解釈である。len_treeA の分布について、二つの構文で大きな差はなかった。A、B、P、修飾部分（Pre、Post）が構文選択前に所与のものであると仮定した時、B の部分が長かった場合に主題化が抑制されてガ格として収まるのではないだろうか。逆に B が短く済む場合は、A、B まで含めて主題化することに制限がかかっていない可能性を考えることができる。

4.2 n_post

n_post の Estimate は 1.460 であった。post の修飾成分が増えると、「A ノ B ハ P」である可能性が高まる、と解釈可能である。これは、主題を提示する役割を持つ助詞「は」と文の主辞 P との距離が離れすぎることを回避している、という解釈が可能ではないだろうか。これは、主題の「は」が文の主辞にかかるときに、その依存関係を見失うことを防ぐ効果があるように考えられる。「A ハ B ガ P」構文の場合、既に部分木 B の分、助詞「は」と P の間に距離ができている。しかし、「A ノ B ハ P」構文の場合、助詞「は」を伴うのは部分木 B である。もし修飾成分 post がなければ助詞「は」と P の距離はないため、修飾成分 post を挿入する余裕があるといえる。修飾部分 post が多くなる場合には「A ノ B ハ P」構文を選択する、という説明を考えることができる。

4.3 pos2_P

表 4 に示された結果から、P の主辞が動詞のとき、名詞類のときと比べて有意に「A ハ B ガ P」構文である可能性が有意に高いことを読み取ることができる。同様に、P の主辞が形容詞のときも、名詞類のときと比べて有意に「A ハ B ガ P」構文である可能性が高いことを読み取ることできる。また形容詞の場合と動詞の場合を比較した場合には、構文選択に有意差は認められなかった。

この結果は、P が動詞や形容詞のときと、名詞のときとでは、コピュラ文かどうか、という点で質的に異なるという点が大きく影響している可能性がある。より詳しく言えば、「A ハ B ガ P」構文において、P が動詞や形容詞のとき、「A は」は、「B が P」が成立するためのある種の条件や限定のような意味をもっていると解釈できる。それに対し、「A ノ B ハ P」構文では、このような条件を示すような意味はなく、P の対象の外延を定める役割に留まっているように読むことができる。このような質の差が、構文の選択に対して効果を持っている可能性がある。

5. おわりに

本研究では、同じ命題内容を表しているといえる 2 つの構文「象は鼻が長い」構文（「A ハ B ガ P」構文）と「象の鼻が長い」構文（「A ノ B ハ P」構文）において、構文を選択する際にどのような要因が関係しているのかについて、GLM によってモデルを構築することで検討した。結果として、部分木 B の長さや修飾部分 post の数、文の主辞 P の品詞などの変数が大きく影響している可能性が示唆された。

今後の課題としては、主に 3 つのことが挙げられる。

まず、客観性である。対象が真に交換可能かどうかを第三者の目をもって調査・確認する必要性があると考えられる。実例で選択されていない方の構文に加工して呈示する必要があるため、多少の「不自然さ」は必然的に発生することが想定される。調査の手続きや交換可能の基準について精査しながら進めなくてはならないだろう。

次に、本研究で挙げた変数群が説明変数として過不足がないか、という点である。ここまでに挙げたものの交互作用や、新たな変数の候補の有無についてなど、モデルに含める変数の取捨選択に対して検討を続けていく必要があると考える。

そして、有意差のあった説明変数群が、いかにして選択に影響を与えているのかという面の説明が不足している点である。本研究では、各変数の値のみを計算してモデルを構築し、それを当てはめることで検討してきた。今後は、各変数どうしの関係性や構文ごとの変数の分布を観察するなど、考察の材料を収集することでより説得的・整合的な説明ができるようになるだろうと思われる。また、これらの構文にのみ効果のあるルールがあるとは考えにくい。より一般的な法則を参照・照合することで説明を付与していく必要があるだろう。

最後に展望を述べる。本研究では文内要因を検討したが、文内に含まれない要因（文脈要

因）についての検討をすることで、発展させていくことができると考える。例えば、他の文のトピックを分析し、当該構文に含まれる単語との関係性を考察することなどが考えられる。

文　献

三上章 (1960).『象ハ鼻ガ長イ』 第 1 版版　くろしお出版.

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, pp. 345–371.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee (2013). “Universal Dependency Annotation for Multilingual Parsing.” *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.

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 LREC 2018*, pp. 1824–1831.

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.