木形状分布の分析：自然言語の句構造とランダム木について

石井太河　宮尾祐介
東京大学
{taigarana,yusuke}@is.s.u-tokyo.ac.jp

概要

自然言語の構文木はランダムではない．では，自然言語の構文木は「木」として，どのような性質を持っているのか？本研究では，特に，左右の分岐の深さなどの木の形状に着目する．構文木の形状は，言語により異なることがあり，多様な言語を特徴付ける要素として重要である．一方で，同一言語内でも構文木の形状は一定ではなく，定性的には特徴を捉えきれない．そこで，本研究では，句構造ツリーバンクに関して木形状の分布を計算し，言語による差異・共通点をより詳細に調査する．また，ランダム生成した木と比較し，自然言語特有の現象について考察する．

1　はじめに

自然言語は文法構造を持つ．文法構造は，世界に存在する多様な自然言語の共通性や差異を考える上で重要な要素となっている [1]．例えば，日本語は構文木の枝分かれが左に深い傾向があり左分岐な言語とされる一方で，英語は逆に構文木が右に深い傾向があり右分岐とされる [2]．本研究では，言語を特徴づける要素として構文木の形状とその分布に着目し，「自然言語の構文木は『木』として，どのような性質を持っているのか？」という問いに取り組む．

言語を特徴付ける要素として構文木の形状に着目することは，言語学的・認知科学的な観点から重要である．例えば，トップダウンやボトムアップといった parsing strategy によってはメモリの観点で処理しやすい木構造が異なることが知られており [3]，実際に人間が脳内で構文木形状の異なる言語に対し異なる parsing strategy を使用することも示唆されている [4]．また，応用面では，構文木の形状に関するバイアスを明示的に導入することで教師なし構文解析の性能が向上することも報告されている [5, 6]．

一方で，自然言語の構文木は同一言語内でも異なる形状のものがありえる．しかしながら，WALS [1] といった既存の言語類型論的特徴の多言語データベースは，同一言語内の特徴の分布まで捉えるには至っていない [7]．

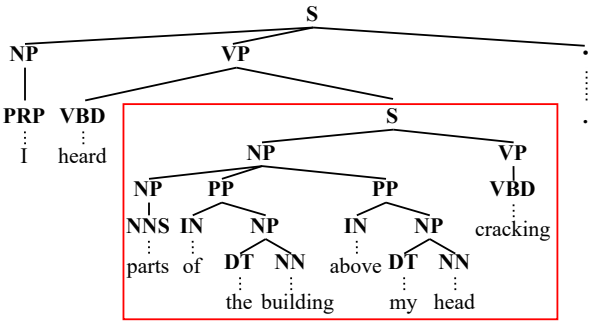

図 1　構文木とその部分木の例

そこで，本研究では，複数の自然言語のツリーバンクに対し様々な木形状の分布を句カテゴリごとに計算し，自然言語間での共通点や差異を分析する．また，田中 [8] にならい，自然言語の統計情報を元にランダム生成された木と比較を行い，自然言語の文法性が木形状に与える影響についても議論する．

2　背景

2.1　構文木の定義

多言語に関して大量の文法構造データから文法的特徴の分布を計算する研究は，依存構造に関しては行われてきた [9]．これに対し，本研究では句構造を分析の対象とする．すなわち，本研究で分析対象とする構文木は，葉が単語からなり，各ノードに句カテゴリのラベルが付いた木である．なお，構文解析などでは 2 分木のみを扱うこともあるが [10, 11]，本研究では一般の木を対象とする．[1)]

2.2　木形状およびその指標

木形状（tree shape，tree topology）は、ラベルなしの木について定められる．本研究では，Chan ら [5] にならい，木全体の形状だけでなく，与えられた木

1)　Crossing bracket については考慮しない．

表 1　本研究で使用する木形状指標．図 1の構文木の部分木（赤枠部分）に対する計算値が示されている．

主な計測対象	指標	計算例
左右の分岐	corrected Colles index	−0.095
	equal weights Colles index	0.233
	Roger's J index	0.333
中央埋め込み	max center embedding	2
木の平坦さ	aspect ratio	1.00
木の高さ	node level	4
木の大きさ	num leaves inside	8
	span ratio	0.727
部分木の相対位置	num leaves outside/left/right	3/2/1
	node depth	2
	outside center embedding	1

に対する部分木の相対的な関係も広義の木形状として扱う．本研究で使用する指標については 表 1にまとめられている．これらの指標は大まかに，左右の分岐，中央埋め込み，木の平坦さ，木の大きさ，部分木の相対位置などを計算している．

左右の分岐　左右の分岐は，語順や依存関係の順序といった文法的特徴の差異が強く表れると考えられている．表 1にある指標は，tree balance の指標 [12, 13, 14] を左右に拡張した 3 つの指標 [15] であり，主に各ノードの左右の部分木の葉数の差を元に計算される．これらは，$[-1, 1]$ に値を取り，-1 に近いほど左分岐，1 に近いほど右分岐となる．Corrected Colles index は，equal weights Collex index よりも根に近いノードの分岐を強く評価する．Roger's J index は葉数差の符号を扱うため，上記の二つよりも分岐を荒く評価する．

中央埋め込み　中央埋め込みは，メモリ使用量といった認知的観点で議論されることが多い [16]．既存研究では left-corner parsing strategy での最大スタックサイズとして計測されることが多いが [16]，左右に関する対称性を欠くといった問題があるため [17]，本研究では「左右の端の子となっていないノードが根からのパス上にいくつあるか」として計算している．[2] Max center embedding は部分木の根から全ての葉までのパスに対してこれを計算し，その最大値を取る．

木の平坦さ・高さ　平坦さ・高さは，カテゴリがどの程度入れ子になっているのかを表すと解釈できる．aspect ratio は中間ノード数を葉数で割った値であり，木が平坦であるほど値が 0 に近くなる．単項ノードがなければ $(0, 1)$ の値をとるが，あれば 1 以上の値も取りうる．Node level はノードから子孫の葉までの最大のパス長として計算される．

2）[18] と異なり，自然言語の文法的な制約は考慮しない．

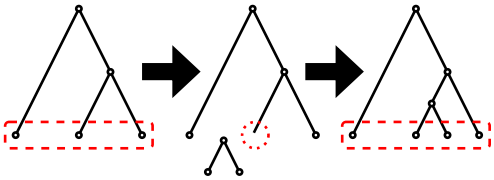

図 2　Yule model の 1 ステップの例．葉を選択し cherry に置換することを繰り返し，木を生成する．

木の大きさ　句の大きさを表す．Num leaves inside は部分木の葉数だが，span ratio は構文木全体に対して部分木が占める割合を計算する．

部分木の相対位置　あるカテゴリが構文木中の特定の位置に出現するかどうかを調べることができる．Num leaves outside/left/right は，部分木の両側・左側・右側にある葉数を計算する．node depth は部分木までの深さであるが，outside center embedding は構文木の根から部分木の根までのパスがどれほど中央埋め込みされているかを計算するため，常に node depth 以下の値を取る．

2.3　ランダム木の生成

本研究では，自然言語の構文木と比較する対象のランダム木として Yule model [19] で生成された木を扱う．Yule model は実装の簡単さなどからランダム木の生成モデルとして基本的なものとなっている [19]．図 2にあるように，Yule model は，一つの葉から開始し，全ての葉のリストから一様ランダムに選んだ一つを cherry に置換する，というプロセスを葉数が一定になるまで繰り返すものである．ここで，cherry とは，根から 2 つの葉が生えた木のことを指す．本研究では，葉数が 2 以外のものも cherry と呼ぶ．

Yule model は，階層的な生成である点において確率文脈自由文法（PCFG）と類似する．しかしながら，PCFG が全ての葉数のラベル付きの木の分布を定めるのに対し，Yule model はある葉数のラベル無しの木の分布を定めるという点で異なる．加えて，PCFG は句カテゴリを持つが，Yule model は持たない．

本研究では，自然言語ツリーバンクの統計量を元に Yule model をパラメタ化し，使用する．

表 2 各ツリーバンクのカテゴリごとのデータ数．#Trees は，部分木ではなく木全体の個数を示す．括弧内のものはデータ数が小さいか，そもそもアノテーションされていないため分析対象から除外する．

	S	NP	VP	ADVP	PP	ADJP	#Trees
PTB	117K	386K	180K	24K	116K	17K	49K
CTB	355K	751K	641K	208K	62K	40K	132K
KTB	84K	152K	(-)	12K	127K	(5)	37K
KorTB	37K	41K	38K	(601)	(-)	(1K)	(5K)
FTB	37K	151K	20K	(1K)	82K	24K	21K

3 実験設定

3.1 自然言語ツリーバンク

本研究では，Penn Treebank（PTB；英語）[20] の WSJ セクション，Chinese Treebank（CTB；中国語）[21]，Keyaki Treebank（KTB；日本語）[22]，Korean Treebank（KorTB；韓国語）[23]，French Treebank（FTB；フランス語）[24] の 5 つの言語のツリーバンクを分析に用いる．[3] 前処理としては，punctuation や数字はそのままにし，null element のみからなる部分木を削除した後，delexicalization を施し preterminal を葉とする．今回は階層的な句構造の木形状を分析することが目的であるので，preterminal を葉とすれば十分である．

カテゴリごとの分析を行うにあたっては，文全体（S），名詞句（NP），動詞句（VP），副詞句（ADVP），前・後置詞句（PP），形容詞句（ADJP）を対象とする．[4] なお，表 2にあるように，データ数が 10000 未満のカテゴリは分析対象から除外する．

3.2 ランダム木

本研究では，Yule model をパラメタ化し，いくつかの統計量が自然言語ツリーバンクと同じになるようなランダム木を生成する．パラメタとしては，「木全体の葉数」，「ノードの次数」，「各葉数での葉の選択位置」をツリーバンクからカウントし，経験分布として用いる．ここで，「各葉数での葉の選択位置」とは，葉のリスト（図 2赤点線枠）から一つ選択し cherry に置換する際の位置のことである．

生成にあたり，まずは木全体の葉数 L を経験分布からサンプルする．次に，各葉数での葉の選択位置 i と次数 d をサンプルし，i 番目の葉を次数 d の cherry と置換する，というのを葉数が L を超えるまで繰り返す．[5] 各葉数での葉の選択位置のカウントは，Yule model と逆のプロセスを辿ることで行う．一つの木で複数のプロセスがあり得る場合はランダムに一つ選択する．

実験では，各ツリーバンクに対し，それぞれの統計量を用いて 100000 個のランダム木を生成する．

3） 予測モデルの学習は行わないため，データセットは分割せず全て分析に用いる．

4） 各ツリーバンク特有のカテゴリの対応付けは 付録 A に記す．

表 3 自然言語ツリーバンクの木形状分布の共通部分の割合

	S	NP	VP	ADVP	PP	ADJP
corrected Colles index	0.10	0.61	0.18	0.86	0.18	0.69
equal weights Colles index	0.08	0.61	0.18	0.86	0.21	0.70
Roger's J index	0.10	0.62	0.19	0.87	0.25	0.70
max center embedding	0.52	0.64	0.70	0.85	0.64	0.58
aspect ratio	0.11	0.15	0.28	0.79	0.37	0.34
node level	0.58	0.71	0.65	0.85	0.74	0.66
num leaves inside	0.56	0.72	0.63	0.92	0.79	0.72
span ratio	0.52	0.58	0.63	0.72	0.72	0.46
num leaves outside	0.52	0.60	0.67	0.70	0.58	0.66
num leaves outside left	0.68	0.61	0.74	0.83	0.56	0.79
num leaves outside right	0.44	0.67	0.59	0.81	0.63	0.77
node depth	0.48	0.66	0.58	0.63	0.66	0.68
outside center embedding	0.56	0.51	0.51	0.65	0.49	0.62

表 4 各ツリーバンクと Yule model の木形状分布の共通部分の割合．KorTB はデータ数が少ないため参考値となる．

	PTB	CTB	KTB	KorTB	FTB
corrected Colles index	0.85	0.79	0.75	(0.88)	0.79
equal weights Colles index	0.57	0.78	0.74	(0.53)	0.50
Roger's J index	0.66	0.83	0.61	(0.56)	0.52
max center embedding	0.88	0.96	0.91	(0.89)	0.78
aspect ratio	0.92	0.75	0.68	(0.93)	0.74
node level	0.86	0.90	0.82	(0.76)	0.79

4 実験結果・考察

木形状の分布を分析するにあたり，本研究では，histogram intersection（HI）を用いる．HI とは，複数のヒストグラムの共通部分の合計カウントである．ヒストグラムが正規化されている場合は [0, 1] に値を取り，分布が何割程度共通しているかを示す．表 3と表 4では，共通部分が 1/4 以下の場合は青色，3/4 以上の場合は赤色で示す．なお，ヒストグラムのビン数は全ての設定で 100 とする．

4.1 自然言語間の比較

表 3は，各カテゴリに対しデータ数が十分ある自然言語ツリーバンク全てに対し木形状指標の分布の HI を計算した結果である．まず第一に，カテゴリにより HI 値の高低が異なることが確認でき，各カテ

5） 葉数 n での葉の選択位置の経験分布が無い場合は一様分布を用いる．

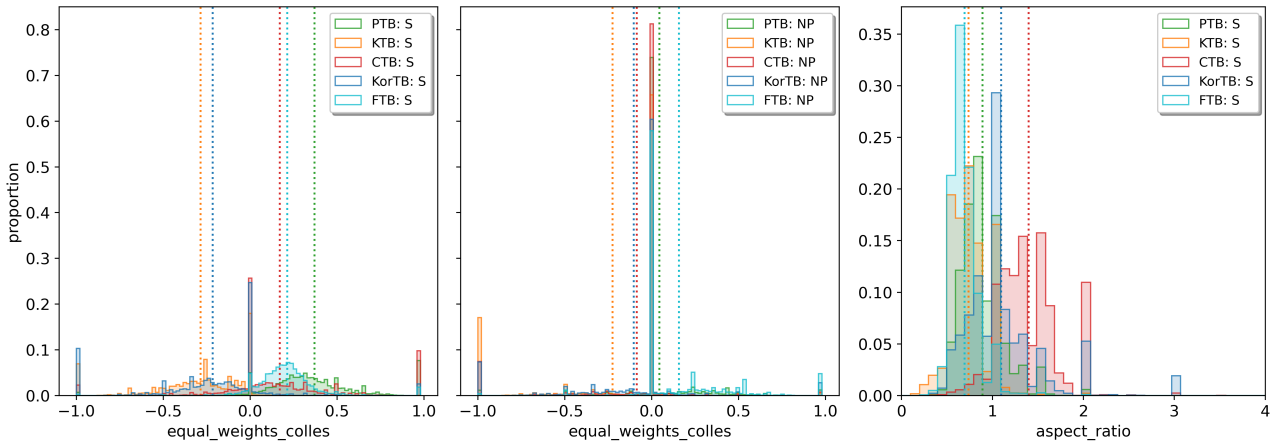

図 3　自然言語ツリーバンクの木形状分布：（左・中）S・NP に対する equal weights Colles index，（右）S に対する aspect ratio．点線は，各ツリーバンクに対する平均値を示す．

ゴリが異なる木形状特性を持つことが示唆される．次に，ADVP はどの指標に対しても HI が高いが，これは ADVP がツリーバンクによらずほとんど短いスパンしかないためだと考えられる．

一方で，左右の分岐に関する 3 つの指標はどれも，S，VP，PP でかなり低く，言語間に大きな差があると言える．実際に図 3（左）を見ると，KTB，KorTB（日本語，韓国語）は平均的に左分岐，CTB，FTB，PTB（中国語，フランス語，英語）は平均的に右分岐となっており，定性的な言語の分岐 [2] と一致することが分かる．また，どの言語でも分布に広がりがあり，同一言語内でも構文木の形状にばらつきがあることが確認できる．

これに対し，NP ではどの分岐指標でも共通部分が 6 割ほどあり，S，VP，PP に比べ言語差が小さい．図 3（中）からは，NP は言語によらず左右の分岐の差が小さく，バランスしている木の数が多いことが見られる．これは，Zhang ら [4] による中国語・英語の観察結果と一致し，多言語への拡張になる．

また，aspect ratio は，S，NP，VP で HI が低く，分布の重なりが小さい（図 3右）．詳しい原因は明らかでないが，aspect ratio は葉あたりのノード数を意味するので，ツリーバンクのアノテーションスタイルの違いなどが分布の違いの要因として推測できる．

4.2　自然言語構文木とランダム木の比較

表 4は，Yule model でランダム生成した木と，元となるツリーバンクの間の HI である．Yule model はラベルなしの木を生成するため，カテゴリごとの分析は行わず，木全体の形状に対する結果のみを記している．驚くことに，多くツリーバンク・指標で HI が 7.5-9 割程度と高く，Yule model のような単純なランダム木生成モデルでも自然言語の構文木に似た形状を生成できることが分かる．

一方で，左右の分岐に関しては，CTB 以外で corrected Colles index がその他の二つの指標よりも HI が高い傾向がある．Corrected Colles index は他の二つと異なり，根に近い部分の分岐を優先的に評価する性質があることから，Yule model で生成した木は，葉に近いノードほど自然言語の構文木とは左右の分岐バランスが異なってしまうことが分かる．i 番目の葉を選択するという Yule model の特性上，葉の数が多いほど，葉の位置が持つ構造情報が希薄になってしまうのではないかと推測できる．これは，言語の階層的生成において，カテゴリといった文法的要素が構造情報を下層に伝える上で重要であることを，木形状の観点から示唆する．

5　結論と今後の展望

本研究では，自然言語の構文木が木としてどのような性質を持つのか？という問いに木形状の観点から取り組んだ．自然言語ツリーバンクに関して，句カテゴリごとに木形状分布を計算した結果，特に左右の分岐や平坦さにおいて言語間で分布に差異があることを定量的に確認した．今後は，単一指標だけでなく，複数指標を組み合わせた際の分布も分析することで，より詳細に言語の性質を分析できることが期待される．一方で，単なる平均ではなく分布を分析対象とする以上，より多くのデータが必要になるという問題がある．Blasi ら [9] のように，パーサーを用いて大規模に生成したデータを分析するといった対応も必要になると考えられる．

謝辞

本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2108 の支援を受けたものです．

参考文献

[1] Matthew S. Dryer and Martin Haspelmath, editors. **WALS Online (v2020.3)**. Zenodo, 2013.

[2] Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu. An empirical comparison of unsupervised constituency parsing methods. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3278–3283, Online, July 2020. Association for Computational Linguistics.

[3] Steven P Abney and Mark Johnson. Memory requirements and local ambiguities of parsing strategies. **J. Psycholinguist. Res.**, Vol. 20, No. 3, pp. 233–250, May 1991.

[4] Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong. Is the brain mechanism for hierarchical structure building universal across languages? an fMRI study of Chinese and English. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 7852–7861, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[5] Samuel W K Chan, Lawrence Y L Cheung, and Mickey W C Chong. Tree topological features for unlexicalized parsing. In **Coling 2010: Posters**, pp. 117–125, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[6] Nickil Maveli and Shay Cohen. Co-training an Unsupervised Constituency Parser with Weak Supervision. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 1274–1291, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling language variation and universals: A survey on typological linguistics for natural language processing. **Comput. Linguist.**, Vol. 45, No. 3, pp. 559–601, September 2019.

[8] 田中久美子. 言語とフラクタル: 使用の集積の中にある偶然と必然. 東京大学出版会, 東京, 2021.

[9] Damian Blasi, Ryan Cotterell, Lawrence Wolf-Sonkin, Sabine Stoll, Balthasar Bickel, and Marco Baroni. On the distribution of deep clausal embeddings: A large cross-linguistic study. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3938–3943, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. The Infinite PCFG Using Hierarchical Dirichlet Processes. In **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)**, pp. 688–697, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[11] Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. Unsupervised Recurrent Neural Network Grammars. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 1105–1117, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Stephen B. Heard. Patterns in Tree Balance Among Cladistic, Phenetic, and Randomly Generated Phylogenetic Trees. **Evolution**, Vol. 46, No. 6, pp. 1818–1826, 1992.

[13] Arne O. Mooers and Stephen B. Heard. Inferring Evolutionary Process from Phylogenetic Tree Shape. **The Quarterly Review of Biology**, Vol. 72, No. 1, pp. 31–54, March 1997.

[14] James S. Rogers. Central Moments and Probability Distributions of Three Measures of Phylogenetic Tree Imbalance. **Systematic Biology**, Vol. 45, No. 1, pp. 99–110, March 1996.

[15] Taiga Ishii and Yusuke Miyao. Tree-shape uncertainty for analyzing the inherent branching bias of unsupervised parsing models. In Jing Jiang, David Reitter, and Shumin Deng, editors, **Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)**, pp. 532–547, Singapore, December 2023. Association for Computational Linguistics.

[16] Marten van Schijndel, Brian Murphy, and William Schuler. Evidence of syntactic working memory usage in MEG data. In **Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics**, pp. 79–88, Denver, Colorado, June 2015. Association for Computational Linguistics.

[17] 能地宏. Left-corner Methods for Syntactic Modeling with Universal Structural Constraints（言語の普遍性を取り入れた統語モデリングのための左隅型解析法）. PhD thesis, 総合研究大学院大学, 2016.

[18] Ethan Wilcox, Roger Levy, and Richard Futrell. Hierarchical representation in neural language models: Suppression and recovery of expectations. In **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 181–190, Florence, Italy, August 2019. Association for Computational Linguistics.

[19] Mareike Fischer, Lina Herbst, Sophie Kersting, Luise Kühn, and Kristina Wicke. Tree balance indices: A comprehensive survey, September 2021.

[20] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. **Computational Linguistics**, Vol. 19, No. 2, pp. 313–330, 1993.

[21] Nianwen Xue, Xiuhong Zhang, Zixin Jiang, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. Chinese treebank 9.0 ldc2016t13, 2016.

[22] Alastair Butler, Zhu Hong, Tomoko Hotta, Ruriko Otomo, Kei Yoshimoto, and Zhen Zhou. Keyaki treebank: phrase structure with functional information for japanese. In **Proceedings of Text Annotation Workshop**, p. 41, 2012.

[23] Na-Rae Han, Shijong Ryu, Sook-Hee Chae, Seung yun Yang, Seunghun Lee, and Martha Palmer. Korean treebank annotations version 2.0 ldc2006t09, 2006.

[24] Anne Abeillé, Lionel Clément, and François Toussenel. Building a treebank for french. In Anne Abeillé, editor, **Treebanks: Building and Using Parsed Corpora**, pp. 165–187. Springer Netherlands, Dordrecht, 2003.

A ツリーバンクごとのカテゴリの対応づけ

表 5は，各ツリーバンク特有の句カテゴリと本研究で扱う句カテゴリの対応付けを示す．

表 5 ツリーバンクごとのカテゴリの対応付け．ハイフンは対応するカテゴリが無いことを表す．

	S	NP	VP	ADVP	PP	ADJP
PTB	S	NP	VP	ADVP	PP	ADJP
CTB	IP	NP	VP	ADVP	PP	ADJP
KTB	IP	NP	-	ADVP	PP	ADJP
KorTB	S	NP	VP	ADVP	-	ADJP
FTB	SENT, Sint Srel, Ssub	NP	VPinf VPpart	AdP	PP	AP