児童作文における係り受け距離と階層距離

著者	今田 水穂
雑誌名	言語資源活用ワークショップ発表論文集
巻	6
ページ	338-347
発行年	2021
URL	http://doi.org/10.15084/00003507

児童作文における係り受け距離と階層距離

今田 水穂 (筑波大学) *

Dependency distance and hierarchical distance in children's writing

Mizuho Imada (University of Tsukuba)

要旨

児童作文の文節係り受け構造について、係り受け距離と階層距離 (係り受けの深さ) の分布を調べた。係り受け距離和と階層距離和の頻度分布はいずれも対数正規分布に従っており、それを文節数 -1 で除した係り受け平均と階層距離平均も同様の分布だった。係り受け距離平均と階層距離平均は文節数に従って大きくなるので、学年を変量効果として $\mu = (a_f + a_r)\log(n/2)$ で線形混合モデル分析を行った。固定効果は後者の方が大きく、全体としては長い係り受けよりも深い係り受けを使って文を長くすることが分かった。また、変量効果を見ると小学校低学年から中学年にかけては長い係り受けを比較的多く使用し、高学年以降は比較的使わなくなっていくこと、ほぼ全学年を通じて学年が上がるほど深い係り受けをより多く使用するようになることが分かった。

1. はじめに

文の統語的複雑さは、文の長さや文中の節の数などによって評価される。しかし、同じ長さの文であっても構造による複雑性の違いがある。そこで本稿は、文節係り受け構造の複雑さを数値化する方法を考え、この指標が児童の作文能力の発達を分析する上で有用な指標になり得るかを検討する。

日本語における文節係り受け構造については、係り受け距離の分布が Zipf の法則に従う (丸山・荻野 1992, 金 1996) ことなどが報告されているが、本研究では文節単位ではなく文単位で構造の複雑性を数値化するため、文節数、係り受け距離平均 (MDD)、階層距離平均 (MHD) を調べる。文節数については、古橋 (2012) が日本語の青空文庫と京都大学テキストコーパス (毎日新聞) のテキストにおける文長の分布を調査しており、文字単位では対数正規分布によく当てはまるが、文節単位では文庫では対数正規分布とガンマ分布の当てはまりに差がなく、新聞ではガンマ分布の方がよく当てはまることを報告している。MDD と MHD については、Jing and Liu (2015) が英語とチェコ語について調査しており、これらが非対称な形状の分布を持つこと、文長と正の相関を持つこと、英語は MDD、チェコ語は MHD を増加させることで文を長くする傾向があることなどを報告している。また、Komori et al. (2019) は日本語学習者の作文における文長、MDD、MHD の分布を調査し、学習者の習熟度によって MHD に

* imada.mizuho.gf@u.tsukuba.ac.jp

は有意な差が見られるが、MDD の変化はそれほど明確ではないことを報告している。

本稿は日本語を母語とする小学生、中学生の作文における係り受け構造の複雑さの発達を調べる。MDD と MHD は文長 (文節数) と正の相関を持つため、これらを切り分けて分析する必要がある。しかし MDD と MHD は文節数に従って線形に増加するわけではないので、適切な回帰モデルを検討する必要がある。そのため、まず文節数、係り受け距離和 (SDD)、階層距離和 (SHD) の分布を調べ、その特徴を明らかにする (平均ではなく和を用いるのは、単にデータの分布を見るためには除算して平均にする必要がないためである)。次にその分布の特徴に基づいて MDD 及び MHD と文節数の回帰モデルを構築し、学年を変量効果とする線形混合モデル分析を行う (回帰モデルで平均を使うのは、その方がモデルが単純になるためである)。

2. 定義

ある文節の係り先までの距離を係り受け距離 (dependency distance, DD)、文末に到達するまでの係り受けの数を階層距離 (hierarchical distance, HD) とする。文中の全ての文節の DD の総和を係り受け距離和 (sum of dependency distandes, SDD)、hd の総和を階層距離和 (sum of hierarchical distances, SHD) とする。

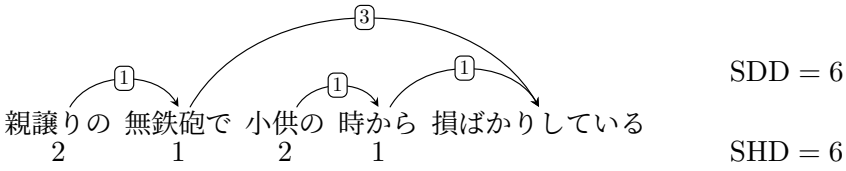


図 1　係り受け距離と階層距離

文節数 n の文において、係り受けの数は $n-1$ なので、SDD と SHD を $n-1$ で除したものをそれぞれ係り受け距離平均 (mean of dependency distances, MDD)、階層距離平均 (mean of hierarchical distances, MHD) とする。

3. データ

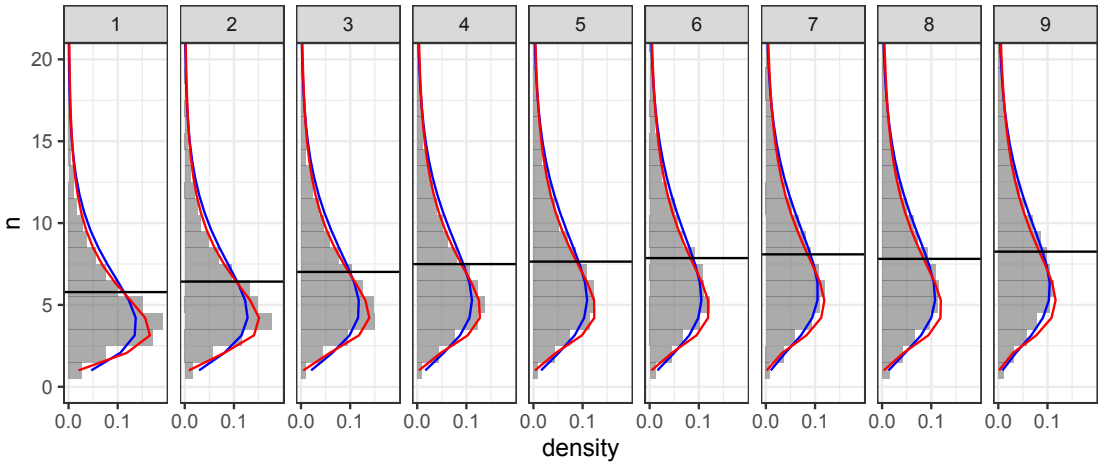
「児童・生徒作文コーパス」(宮城・今田 2018) を使用する。このコーパスは 2014 年から 2016 年にかけて「夢」「頑張ったこと」という 2 つの課題で小学 1 年〜中学 3 年の児童・生徒の作文を収集したもので、CaboCha/UniDic による形態素・構文解析が施されている。コーパスの規模は以下の通りである (grade は学年で 1〜9 が小 1〜中 3 を表す。segments は文節数)。小学校は 1 学年 2 クラス、中学校は 1 学年 4 クラスを調査対象としているため、作文数は中学校が小学校の 2 枚程度となっている。

表 1　児童・生徒作文コーパスの規模

grade	documents	sentences	segments
1	405	2896	16749
2	406	3999	25691
3	402	5125	35961
4	433	6190	46385
5	448	6524	49884
6	455	6332	49774
7	930	15030	121578
8	932	16010	125091
9	914	15397	127101

4.　文節数の分布

学年別の 1 文あたり文節数の分布をヒストグラムで確認する。文節数が 100 を超える長い文もあるので、文節数 20 までの範囲を示す。図中の水平線は平均値、曲線は赤が対数正規分布、青がガンマ分布の確率密度曲線である。



平均値を見ると、全体的には学年に従って文節数は増大するが、小学校低学年から中学年にかけての変化と比べると、高学年以降の変化はそれほど大きくない。新井ほか (2017) は小学校と中学校の理科教科書について、小学校から中学校に上がる段階で 1 文あたりの係り受けの数に明白な増加が見られることを指摘している (文の係り受けの数は文節数 −1 なので、これは実質的に文節数を数えているのと等価と考えられる)。本研究の調査からは、小学校と中学校の教科書の間に見られる文の複雑さのギャップに反して、児童の作文においては、小学校高学年と中学校の間でそのような明白な差異は確認できないことが分かる。

分布については、R 言語の fitdistr 関数によるフィッティングで得た対数尤度を示す。対数尤度が大きいほど当てはまりが良いとされる。小学校では対数正規分布の方が当てはまりが良

く、中学校ではガンマ分布の方が当てはまりが良い。

表 2　文節数におけるガンマ分布と対数正規分布の対数尤度

grade	nobs	gamma	lognormal
1	2896	-7302.59	-7143.91
2	3999	-10379.27	-10173.96
3	5125	-13680.14	-13482.98
4	6190	-16899.55	-16774.30
5	6524	-17945.36	-17847.14
6	6332	-17657.44	-17631.28
7	15030	-41722.60	-41861.24
8	16010	-44203.95	-44493.58
9	15397	-42889.96	-43221.80

古橋 (2012) は日本語の文節数の分布について、青空文庫では対数正規分布とガンマ分布の当てはまりに差がないが、京都大学テキストコーパス (毎日新聞) ではガンマ分布の方が当てはまりが良いことを報告している。文庫や小学生のテキストでは対数正規性が高く、新聞や中学生のテキストでは対数正規性が低いことから、文体の違いによって対数正規性に差が出る可能性が考えられるが、より多くの種類のテキストを分析しなければ結論できない。

5. 係り受け距離和と階層距離和の分布

文節数 8 の文を例として係り受け距離和 (SDD) と階層距離和 (SHD) の分布を考える。日本語のように係り受けが前方から後方へ一方向にかかる言語では、文節数 n の文の可能な係り受け構造の数はカタラン数 c_{n-1} で計算することができ、$n = 8$ のときは $c_7 = 429$ である。

$$c_n = \frac{(2n)!}{(n+1)!n!}$$

全ての構造が同じ確率で生起すると仮定した場合、SDD と SHD は全く同様の頻度分布を持つ。この分布は対数正規分布によく当てはまる (対数尤度は、対数正規分布が -1180.965、ガンマ分布が -1181.483)。[(1)]

[(1)] SDD ではなく DD についてであるが、Liu (2007) は中国語の依存構造について、実際のデータ、およびランダムに構造を作成したデータでは DD はゼータ分布に従うが、係り受けの交差を含む非文法的な構造も認めたランダムな構造データではハイパーポワソン分布に従うという興味深い指摘をしている。

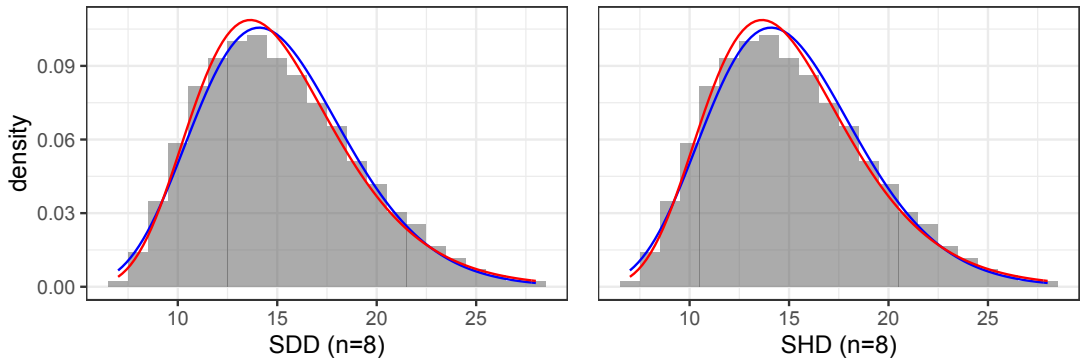


図 2　文節数 8 の文における SDD と SHD の分布 (可能な構造数)

実データでも SDD と SHD は対数正規分布することが予想される。そこで実際のデータの分布を確認する。

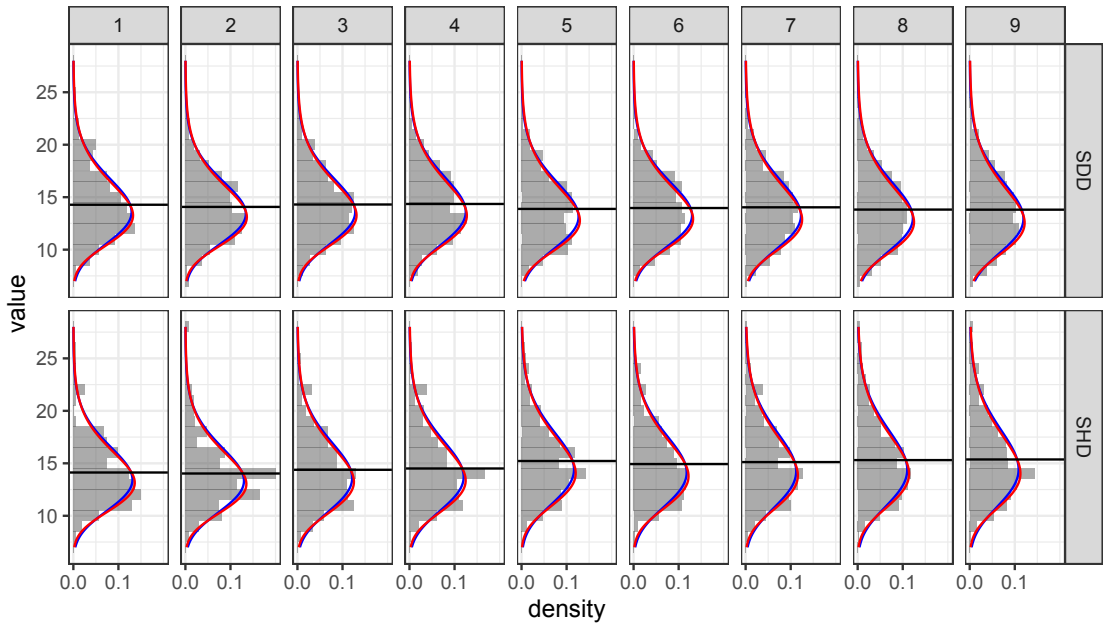


図 3　文節数 8 の文における SDD と SHD の分布 (作文データ)

平均値は小学校低学年では SDD の方が僅かに高いが、それ以降は SHD の方が高い。いずれの値も全学年を通じて大きく変化しないように見えるが、厳密には SDD は増加した後減少、SHD はほぼ一貫して増加するようである (6 節参照)。分布は、対数正規分布にもガンマ分布にも見える。両分布でフィッティングした対数尤度を以下に示す。

表 3　SDD と SHD におけるガンマ分布と対数正規分布の対数尤度

		logLik.(SDD)		logLik.(SHD)	
grade	nobs	gamma	logno	gamma	logno
1	161	-410.64	-409.88	-408.17	-405.18
2	273	-686.11	-686.83	-694.69	-689.18
3	386	-987.74	-986.76	-1008.81	-1005.09
4	503	-1300.34	-1301.70	-1319.45	-1312.44
5	568	-1460.85	-1460.74	-1508.14	-1502.51
6	579	-1474.99	-1476.71	-1530.51	-1523.42
7	1366	-3555.19	-3559.23	-3678.74	-3666.77
8	1487	-3873.89	-3879.66	-4033.86	-4016.12
9	1519	-3998.65	-4005.06	-4141.69	-4123.61

SDD は、小学 2 年と 4 年でガンマ分布の方が当てはまりが良いが、それ以外の学年では対数正規分布の方が当てはまりがよい。SHD は、全ての学年において対数尤度の方が当てはまりが良い。全体としては、SDD、SHD ともに対数正規分布の方が当てはまりが良いと考えられる。

6. 線形回帰

文節数 n が大きくなるほど SDD と SHD も大きくなるので、SDD と SHD を目的変数、n を説明変数として回帰分析することを考える。ただし、モデルを単純にするために SDD、SHD ではなく MDD、MHD を使う。文節数 n のときの SDD、SHD の最大値は $\frac{n(n-1)}{2}$、最小値は $n-1$ である。文節数 n の文における係り受けの数は $n-1$ なので、MDD $= \frac{\mathrm{SDD}}{n-1}$、MHD $= \frac{\mathrm{SHD}}{n-1}$ である。従って MDD、MHD の最大値は $\frac{n}{2}$、最小値は 1 である。これらは $\left(\frac{n}{2}\right)^1$、$\left(\frac{n}{2}\right)^0$ なので、中央値も $\left(\frac{n}{2}\right)^a$ に従うと予想される。MDD、MHD は SDD、SHD を $n-1$ で除しただけなので、頻度分布は SDD、SHD と同じく対数正規分布である。対数正規分布の中央値 e^μ は、対数変換して正規分布にしたときの平均 μ に対応する。従って MDD、MHD は次の単純な式で回帰分析できる。

$$\log \mathrm{MDD} = a \log \frac{n}{2}$$
$$\log \mathrm{MHD} = a \log \frac{n}{2}$$

データの分布を確認する。図中の赤線は fitdistr 関数で推定した μ である。

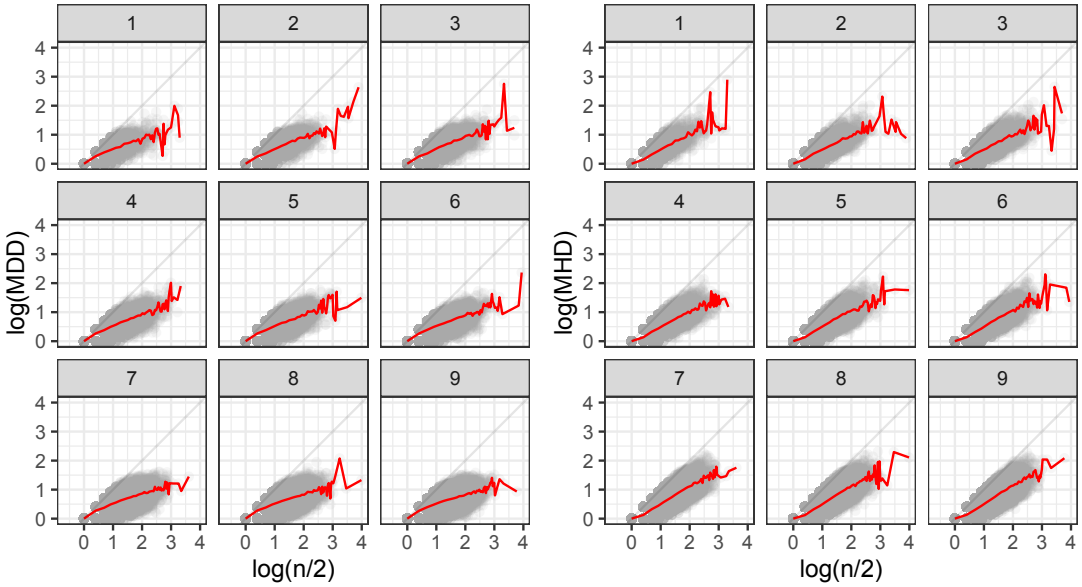


図 4　作文コーパスにおける MDD と MHD の分布

μ は原点を通る直線に近い分布を示しており、概ね想定したモデルに従っている。全ての学年を一度に分析するために、以下のモデルで線形混合モデル分析を行う。a_f は固定効果、a_g は学年による変量効果である。

$$\log \mathrm{mdd} = (a_f + a_g) \log \frac{n}{2}$$
$$\log \mathrm{mhd} = (a_f + a_g) \log \frac{n}{2}$$

分析は R 言語の lmer 関数で行った。結果を以下に示す。

	log MDD	log MHD
log(n/2)	0.47***	0.52***
	(0.01)	(0.01)
AIC	−2164.92	−2673.42
BIC	−2137.18	−2645.68
Log Likelihood	1085.46	1339.71
Num. obs.	76586	76586
Num. groups: school_year	9	9
Var: school_year log(n/2)	0.00	0.00
Var: Residual	0.06	0.06

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

表 4　MDD と MHD の線形混合モデル分析

固定効果は log MDD が $a_f = 0.47$、log MDD が $a_f = 0.52$ であり、MHD の方が傾きが大きい。これは全学年を通じた傾向として、児童は係り受け距離の長い文節よりも、階層の深い文節を増やすことによって文を長くする傾向があることを示している。Jing and Liu (2015) は英語とチェコ語について、文長と MDD 及び MHD の間に相関があることを指摘しており、さらに両言語とも MHD の方が文長との相関が強いことを報告している。本稿の分析結果は Jing and Liu (2015) の報告と合致するものであり、文構造の複雑化において線形的な拡張よりも階層的な拡張が好まれる傾向は、広範な言語で観察されるものと予想される。(2)

次に、学年ごとの変量効果を確認する。

表 5　学年による変量効果

grade	log MDD	log MHD
1	0.0055	−0.0218
2	0.0121	−0.0312
3	0.0137	−0.0212
4	0.0171	−0.0159
5	0.0043	0.0033
6	−0.0003	0.0059
7	−0.0096	0.0236
8	−0.0175	0.0233
9	−0.0253	0.0341

log MDD の a_g は小学 1 年から 4 年にかけて増加した後、小学 5 年で小学 1 年以下の水準まで急激に減少し、その後は中学 3 年まで減少を続ける。log MHD のの a_g は小学 1 年から 2 年にかけて減少するが、その後は一貫して増加を続ける。全体としては (固定変量から分かるように) 全学年を通じて MHD の増加が MDD の増加よりも文節数の増加に寄与しているが、相対的には中学年までは線形的な文の拡張によって MDD を増加させる傾向が強く、高学年以降はその傾向が弱くなると考えられる。これは文節数が小学校中学年まで増加し、その後はそれほど変化しないことと関連しているかも知れない。すなわち中学年までは短い文の使用頻度が高く、短い文では階層を増やすのではなく、動詞に係る名詞や副詞を増やすことで文を長くしていることが考えられる。

Komori et al. (2019) は中国語母語の日本語学習者の作文について、大学 2 年と 3 年では文長と MHD に差が見られるが、MDD には明確な差が見られないことを指摘している。文長と MDD、MHD の間には相関があり、Komori et al. (2019) のデータは回帰分析を行なっていな

(2) Jing and Liu (2015) はまた、英語はチェコ語よりも MDD が増大しやすく、チェコ語は英語より MHD が増大しやすいことを示すいくつかの証拠を挙げているが、そのモデルは文長と MDD、MHD の間に線形関係を仮定したモデルと思われ、本稿のモデルとは異なるので直接的な比較はできない。一方で、英語とチェコ語の語順の自由さの違いに言及している点は注目に値する。Jing and Liu (2015) はチェコ語は比較的語順が自由な言語であるにも関わらず英語より MDD が小さいと述べているが、語順の自由さはむしろ長い構成素を文の外側に配置するなどの方法で MDD を小さくするために有利に働くだろう。

いため、文長の増加が MDD および MHD の変化にどの程度の影響を及びしているか不明だが、MHD の方が MDD よりも増加しやすいという点は、本稿における小学 5 年以降のデータと一致している。

7. まとめ

児童作文における文節係り受け構造の複雑さについて検討した。まず文節数、係り受け距離和 (SDD)、階層距離和 (SHD) の分布を確認したところ、文節数は学年に従って増加するものの、高学年以降はそれほど顕著に変化しないことが分かった。興味深いことに学齢によって対数正規性に差があり、小学校では対数正規分布、中学校ではガンマ分布によく当てはまった。同じ長さの文における SDD と SHD の分布は概ね対数正規分布に従うようであり、ほとんどの学年で SHD の方が大きく、また学年による変化は SDD、SHD ともそれほど大きくない。

次に係り受け距離平均 (MDD) および階層距離平均 (MHD) と文節数 n について線形回帰を試みた。$\log y = (a_f + a_g) \log \frac{n}{2}$ で線形混合モデル分析したところ、固定変量 a_f は MHD の方が MDD よりも大きかった。これはデータ全体の傾向として、同じ長さの文であれば MHD の方が MDD より大きくなることを意味する。可能な全ての構造が等確率で生起するのであれば、この値は等しくなるはずなので、実際のデータでは距離の長い係り受けよりも階層の深い係り受けを増やすことによって文を長くすることが好まれると解釈することができる。学年による変量効果 a_g からは、同じ長さの文における MHD は概ね学年に従ってより大きな値を取るようになるのに対して、MDD は中学年までは増加するが高学年以降は減少することが分かった。この結果は、ある一定の発達段階に達した後は、距離の長い係り受けを避ける傾向はより強くなることを示唆する。

本稿は作文データにおける文の統語的複雑さを観察、分析したものだが、文の複雑さと処理の困難さの関係は慎重に検討しなければならない。データに見られる文の複雑さの発達は、単に複雑な文を処理する能力の向上を示しているのではなく、複雑な文を回避する能力の発達も反映している可能性がある。また、複雑な構造ほど直ちに処理が困難になるというわけでもなく、複雑な構造において処理速度が向上する Anti-locality 現象が日本語においても観察されている (浅原ほか 2019)。引き続き、多様な観点から分析を行う必要がある。

謝　辞

本研究は JSPS 科研費 19K23068 の助成を受けたものです。

文　献

丸山宏・荻野紫穂 (1992).「日本語における文節間係り受け関係の統計的性質」　情報処理学会第 45 回全国大会講演論文集 (人工知能及び認知科学), pp. 173–174.

金明哲 (1996).「文節の係り受け距離の統計分析」　社会情報 = Social Information, 5:2, pp. 1–11.

古橋翔 (2012).「文の長さ分布に見られる対数正規性」　第 1 回コーパス日本語学ワークショップ予稿集, pp. 93–98.

Yingqi Jing, and Haitao Liu (2015). “Mean Hierarchical Distance Augmenting Mean Dependency Distance.” *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 161–170. Uppsala, Sweden: Uppsala University, Uppsala, Sweden.

Saeko Komori, Masatoshi Sugiura, and Wenping Li (2019). “Examining MDD and MHD as Syntactic Complexity Measures with Intermediate Japanese Learner Corpus Data.” *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pp. 130–135. Paris, France: Association for Computational Linguistics.

宮城信・今田水穂 (2018).「『児童・生徒作文コーパス』を用いた漢字使用能力の発達過程の分析」　計量国語学, 31:5, pp. 352–369.

新井庭子・分寺杏介・石原侑樹・松崎拓也・影浦峡 (2017).「テキストの読みを困難にする特徴の計量分析:小・中理科教科書を対象として」　計量国語学, 31:2, pp. 144–159.

Haitao Liu (2007). “Probability Distribution of Dependency Distance.” *Glottometrics*, 15, pp. 1–12.

浅原正幸・小野創・宮本 エジソン正 (2019). “BCCWJ-EyeTrack.” 言語研究, 156, pp. 67–96.